Beyond the Undermining Effect: Extrinsic Rewards Preserve Neural Intrinsic Reward

Ayabe, Hiroaki

The Graduate University for Advanced Studies, SOKENDAI

School of Life Science

Department of Physiological Sciences

CONTENTS

Appendix Supplementary Materials

Abstract

The Self-Determination Theory posits that extrinsic reward withdrawal often undermines intrinsic motivation. This undermining effect varies according to the reward category (e.g., monetary vs. praise). However, the mechanism underlying this variability remains unclear, partly because previous studies have conceptualized intrinsic motivation as a single construct without distinguishing *liking* (intrinsic value) from *wanting* (motivational drive). Failure to make this distinction may obscure the effects of rewards on each component. We hypothesized that extrinsic reward withdrawal would differentially affect intrinsic value and motivational drive. To test this hypothesis, we conducted functional magnetic resonance imaging (fMRI) on 54 participants who were randomly assigned to one of three conditions: control (no rewards), monetary reward, or social reward. Participants in the monetary and social reward groups played a stopwatch game twice, first with and then without extrinsic rewards. The control group received no extrinsic rewards during the task. We assessed intrinsic value through fMRI-measured activation of the reward system and motivational drive through voluntary engagement during the task intermission. Intrinsic value, measured by activation in predefined reward-related regions, declined in the control group but was preserved in both extrinsic reward groups after withdrawal. Thus, both monetary and social rewards were associated with preserved intrinsic value. However, intrinsic behavioral motivation declined in the monetary group but was maintained in the social group. These findings suggest that extrinsic reward withdrawal differentially influences intrinsic value and motivational drive across reward categories, thereby clarifying the variability of the undermining effect.

Keywords: intrinsic motivation, Self-Determination Theory, Incentive Sensitization Theory, preserving effect, undermining effect, social reward, monetary reward.

1. Introduction

1.1. Background

Intrinsic motivation, defined as engaging in activities autonomously out of interest or curiosity, forms the foundation for sustained learning and creativity across educational and workplace contexts [1]. Self-Determination Theory (SDT) posits that fulfilling the basic psychological needs of autonomy, competence, and relatedness is essential for maintaining intrinsic motivation [2–3]. This framework highlights the role of intrinsic motivation in supporting learning, creativity, and job satisfaction [4].

The undermining effect refers to a decline in intrinsic motivation that follows the withdrawal of an extrinsic reward [5–7]. Deci's seminal study [8] demonstrated that a decline in free-choice behavior—a proxy for intrinsic motivation—after monetary reward withdrawal could not be explained by Operant Conditioning Theory [9]. According to SDT, monetary rewards, when perceived as evaluative, can undermine autonomy and thus reduce intrinsic motivation. By contrast, social rewards, particularly when perceived as informational, support competence and relatedness, fostering internalization of the extrinsic reward that sustains intrinsic motivation [10–11].

However, for decades, scholars have debated whether extrinsic rewards satisfy or thwart these needs [5,12–14], with particular controversy surrounding the relative effects of different reward categories, including monetary incentives and social approval [15–16]. Although the importance of intrinsic motivation is widely recognized, the psychological mechanisms through which extrinsic rewards sustain or enhance it remain poorly

understood. Consequently, incentive systems in educational and organizational contexts, although designed to promote intrinsic motivation, have often produced unintended outcomes, including widened achievement gaps, reduced creativity, and increased turnover [17–19]. This paradox underscores the unresolved question of how extrinsic rewards interact with intrinsic motivation, and indicates that guidance for designing effective institutional interventions remains limited.

One important reason why the mechanisms underlying the undermining effect remain unresolved is persistent methodological limitations. The effects of extrinsic rewards vary with the measurement indices used (e.g., self-reports, free-choice behavior, and performance quality or quantity), undermining the consistency of the conclusions drawn [14,20]. Self-reports alone are insufficient; converging evidence indicates that combining free-choice behavior with neural measures offers a more valid assessment of intrinsic motivation [20–21]. Recent SDT reviews highlight the methodological limitations of prior intrinsic motivation research, particularly the reliance on single indicators such as free-choice behavior and the inadequacy of treating intrinsic motivation as a unitary construct [3,22–23]. These reviews also emphasize the importance of neuroscientific approaches that go beyond surface behaviors and capture the underlying reward processes. For example, recent functional magnetic resonance imaging (fMRI) studies revealed that reward system activation during task performance represents the intrinsic value of the action through action-outcome contingency [24–25].

1.2. Distinguishing Liking and Wanting in the Undermining Effect

Neuroimaging studies on the undermining effect suggest that neural substrates vary across reward categories and that neural reward values can dissociate temporally and functionally from subsequent actions. Murayama et al. [26] first found that withdrawing monetary rewards significantly decreases reward-circuitry activation—a "neural" undermining effect. However, Albrecht et al. [27] did not replicate this effect for monetary rewards and, in the case of social rewards, observed sustained or even enhanced activation after withdrawal. These contrasting findings reveal discrepancies between neural activation and behavioral predictions, challenging simplistic models that assume a direct link between extrinsic reward events, neural values, and subsequent behavior. These perspectives are consistent with recent integrative psychological-neuroscientific reviews on the role of rewards in motivation, indicating that extrinsic rewards are neither inherently detrimental nor beneficial for intrinsic motivation but instead exert complex, context-dependent effects.

A theoretical obstacle to understanding inconsistent neural and behavioral findings, and why reward categories differ in their undermining effects, is that traditional research has not distinguished between two core components of intrinsic motivation: ***liking*** (hedonic experience of reward value) and ***wanting*** (motivational drive to pursue rewards) [3,23]. Developments in Incentive-Sensitization Theory (IST) also provide converging evidence for this distinction, conceptualizing *liking* and *wanting* as neurologically dissociable processes [28–30]. Studies within the IST framework further show that extrinsic rewards can affect *liking* and *wanting* in distinct ways, challenging the long-standing assumption—metaphorically termed a "powerful meme" by Robinson and Berridge [31]—that rewards influence both processes in the same way and at the same time.

Evidence also shows that these dissociations occur not only in drug addiction but also in non-pharmacological rewards, such as gambling, eating behaviors, and monetary and social rewards. These rewards may engage separable neural mechanisms for *liking* and *wanting* [30,32].

An important insight from IST research is that associations *learned* through conditioning and reward-behavior contingencies—stimulus cues that link behaviors with rewards—can differentially modulate *liking* and *wanting* [28,30–31]. For example, the Pavlovian-instrumental transfer (PIT) effect shows that a stimulus paired with a reward can increase the motivational drive (*wanting*) to perform the associated action without substantially affecting *liking* [33–34]. Similarly, *learning* driven by the cognitive evaluation of rewards or reevaluation of social value can lead to fluctuations in *wanting*, while leaving *liking* largely unchanged [25,32]. Together, these findings challenge the assumption that extrinsic rewards uniformly influence *liking* and *wanting* responses.

In the present study, we hypothesized that *different categories of extrinsic rewards influence liking and wanting differently*, given their varying undermining effects. To test this theoretical hypothesis, we first conceptualized the distinction between liking and wanting according to the IST, which explains the pathophysiology of addiction in which the discrepancy between *wanting* and *liking* occurs. Reward is a psychological construct comprising three components: *liking* (hedonic experience or pleasure), *wanting* (motivational drive to obtain pleasure), and *learning* or prediction of future values [30–31]. *Liking* refers to a positive hedonic impact, which is neurally represented in opioid-mediated hedonic hotspot networks in the orbitofrontal cortex (OFC), insular cortex, nucleus accumbens (NAcc), ventral pallidum, and pontine

parabrachial nucleus [28–30]. Thus, *liking* can be quantified as neural responses within these networks during a task, in which the action-outcome contingency is fulfilled [24]. *Wanting* refers to the core process of incentive salience, neurally represented by mesolimbic dopaminergic networks [30–31]. As incentive salience is triggered by Pavlovian cues leading to seeking and consuming their associated reward [30–31], free-choice tasks with appropriate cues, such as the presence of the game, can quantify the degree of *wanting* by counting the number of voluntary task executions [8,26,33].

Second, technically, this study evaluated *liking* during task engagement and wanting during a subsequent free-choice period. This design is based on the concept that the *liking* (hedonic experience) emerges during performance that can be measured neurally, whereas wanting (motivational drive to pursue) precedes performance, and thus can be measured behaviorally as a trial number. We adopted the two-phase reward presentation–withdrawal structure proposed by Murayama et al. [26]. in which the stopwatch (SW) task requires participants to press a button to stop the stopwatch precisely at 5.0 s. The SW task is known to elicit high-level intrinsic motivation through action-outcome contingency [24]. Therefore, contrasting the SW task with the watch-and-stop (WS) task—which lacks an action–outcome contingency because the stopwatch automatically stops at approximately 5.0 s—isolates the neural responses of intrinsic reward, that is, **"***liking***.***"* We repeated this fMRI measurement twice: in the first phase, the SW task success was externally rewarded with money or praise, while in the second phase, no external reward was applied. We modeled task-related activation covering the entire construct of the SW task, including the cue, hold, run, and feedback, and contrasted it with those of the WS task. *Wanting* was measured by SW trial number during a free-choice period after the first and second phases of the

fMRI. This classical behavioral measure has been widely utilized since the seminal study by Deci et al. [8]. By comparing the second phase with the first, we can depict the after-effect of the external rewards on "*liking*" and "*wanting*", and their modality-specificity.

1.3. Neural Evaluation of Liking as an Intrinsic Reward

Traditional measures of *liking* suffer from methodological limitations, such as self-report biases and the unstable validity of momentary hedonic indicators. Because *liking* reflects an internal hedonic value, a valid assessment requires real-time neural activity within the reward circuitry. The intrinsic-reward stopwatch paradigm developed by Miura et al. [24] addressed this limitation by capturing *liking*-related activity across the full action–outcome contingency [28,35]. Building on this approach, the present study incorporated monetary and social extrinsic rewards to enable a more comprehensive real-time assessment of intrinsic reward.

1.4. Objectives

This study aimed to clarify how the presentation and withdrawal of extrinsic rewards influence *liking* during task performance and *wanting* during subsequent free-choice. To this end, we used a two-phase reward presentation–withdrawal design combined with an intrinsic-reward paradigm, enabling simultaneous measurement of neural *liking* and behavioral *wanting*. *Liking* was indexed by reward-circuitry activation, and *wanting* was indexed by voluntary task selections during the free-choice

period. Based on the theoretical hypothesis established in Section 1.2, we tested two specific hypotheses:

H1 (*liking*): Both monetary and social rewards preserve neural *liking* compared with the no-reward condition.

H2 (*wanting*): Following reward withdrawal, *wanting* decreases in the monetary reward condition (undermining effect), but not in the social reward condition.

Support for both hypotheses would indicate that *liking* is preserved across reward categories, whereas *wanting* diverges depending on reward category.

2. Materials and Methods

2.1. Ethics Approval

This study was approved by the Bioethics Committee of the National Institute of Natural Sciences, Japan (Approval No. EC01-035), in accordance with the Declaration of Helsinki [36]. Written and verbal informed consent was obtained from all participants, who were fully debriefed upon completion (Supplementary Material S1).

2.2. Participants

Fifty-eight healthy, right-handed adults aged 18–44 years (mean age = 26.6 ± 7.2 years; 32 women) were recruited by the National Institute for Physiological Sciences, Japan. Due to a program malfunction, data from four participants were lost, leaving 54 valid datasets for the final analyses. Participants were randomly assigned

to one of three groups: C (control), M (monetary reward), or S (social reward), with n = 18 per group. None of the participants had a history of neurological disorders or MRI incompatibility and they all completed two experimental tasks. A priori power analysis using G*Power [37] indicated that a sample size of N = 48 was required to detect an effect size of f = .30 at α = .05 with 80% power (see Supplementary Material S2).

2.3. Experimental Tasks

We used two tasks adapted from Murayama et al. [26]. The SW task requires participants to press a button to stop the stopwatch precisely at 5.0 s and is designed to elicit high levels of intrinsic motivation. In contrast, the WS task requires participants to press a button after the stopwatch automatically stops at approximately 5.0 s. As it lacks a success–failure outcome and provides no feedback, this monotonous task serves as a control condition for the SW task. The success criterion for the SW task was set at 4.94–5.06 s, based on a preliminary study, to ensure moderate difficulty and promote optimal intrinsic reward (see Supplementary Material S3). The average success rate for the SW task was 65.1%, confirming its moderate level of difficulty. The task materials were prepared and presented with Presentation® software (Neurobehavioral Systems, Albany, CA, USA) (RRID: SCR_002521).

2.4. Experimental Procedure

This study adopted a two-phase structure: reward presentation (Experiment phase 1) and withdrawal (Experiment phase 2). Each phase consisted of two parts: an fMRI scan session and a free-choice period (see Table 1), with the former designed to assess intrinsic reward (neural *Liking*) and the latter to evaluate

motivational drive (behavioral *Wanting*). In this study, general terms are written in lowercase (e.g., *liking* and *wanting*), while measured variables are capitalized (e.g., *Liking* and *Wanting*) to clearly distinguish the terminology.

In Experiment phase 1, participants were assigned to three groups (C, control; M, monetary reward; and S, social reward) and performed the SW and WS tasks. In Experiment phase 2, all participants completed the same procedures without extrinsic rewards, allowing assessment of changes in motivation following reward withdrawal. Before each phase, practice tasks were administered to standardize proficiency and minimize the influence of practice effects or individual differences in skills on measurement outcomes. Experiment phase 2 used procedures identical to Experiment phase 1 but was introduced with a different cover story to minimize expectancy and learning effects. To ensure that the relationship between the Experiment phases was not disclosed, independent experimenter teams conducted each Experiment phase with distinct cover stories. Specifically, Experiment phase 1 was described as "baseline data collection for improving fMRI analysis accuracy," whereas phase 2 was described as "assessment of attentional intensity via recognition of action outcomes." This information control minimized expectancy, practice, and learning effects from Experiment phase 1 influencing Experiment phase 2 and prevented inadvertent intervention, ensuring procedural independence and measurement reliability between Experiment phases. This manipulation was approved by the ethics committee, and all participants underwent informed consent reaffirmation (debriefing) immediately after completing Experiment phase 2, during which the true purpose and procedures of the study were explained (see Supplementary Material S1).

In Experiment phase 1, participants completed 30 trials for each of the SW and WS tasks in randomized order (see Figure 1). Each trial comprised a cue presentation (1.5 s), HOLD phase (1.5 s), RUN phase (5.0 s), and feedback presentation (3.0 s), with intertrial intervals pseudorandomly varied between 1.0 and 5.0 s to prevent predictability. Group M received 200 yen (USD 1.43) as a monetary reward for each successful SW trial, while Group S received an applause sound as a nonverbal social reward for success. In contrast, Group C received no reward stimuli. To balance overall rewards, Groups C and S received a fixed participation fee of 3,000 yen (USD 21.43), while Group M received a performance-based reward of the same amount, assuming a 50% success rate.

After scanning, participants entered a three-minute free-choice period, termed "waiting time," in a private room. During this time, participants could freely choose to perform either the SW or WS task, or read weekly, scientific, or regional magazines. Participants' behavior was recorded unobtrusively, and *Wanting* was defined as the number of voluntary *SW*-task initiations during the 3-min free-choice period. During this period, participants were free not to interact with the PC at all (e.g., reading magazines or doing nothing). If they chose to use the PC, they could freely select either the SW or WS task on each trial. Consequently, SW-task initiations formed a continuous behavioral index ranging from 0 (no PC engagement or only WS selections) to 14 (continuous voluntary SW selections).

In Experiment phase 2, all participants completed the scan session and free-choice period under the same extrinsic reward–free conditions as Group C in Experiment phase 1. For participants in Groups M and S, the withdrawal of previously provided performance-based rewards (monetary or applause sound) served as the reward

withdrawal manipulation. Furthermore, to eliminate motivational bias based on task performance, all participants received a uniform participation fee of 4,000 yen (USD 28.57). This design enabled the measurement of within-group changes in intrinsic motivation before and after reward withdrawal in Groups M and S, as well as consistent intergroup comparisons with the reward-free Group C to assess the effects of extrinsic rewards.

Table 1.

Experiment Phase Structure and Reward Conditions

Phase	Experiment phase 1		Experiment phase 2	
Group	Scan session 1	Free-choice period 1	Scan session 2	Free-choice period 2
C	No extrinsic reward	Number of plays in the SW task	No extrinsic reward	Number of plays in the SW task
M	Monetary reward per success		No extrinsic reward	
S	Applause sound on success		No extrinsic reward	

Note. SW = stopwatch. In Experiment phase 1, the three groups received reward manipulations according to the reward conditions (none, 200 yen monetary reward, or applause). In Experiment phase 2, all the tasks were performed without *extrinsic* rewards. Neural activity was recorded during scanning, and SW task selection during the free-choice period indexed behavioral *Wanting*.

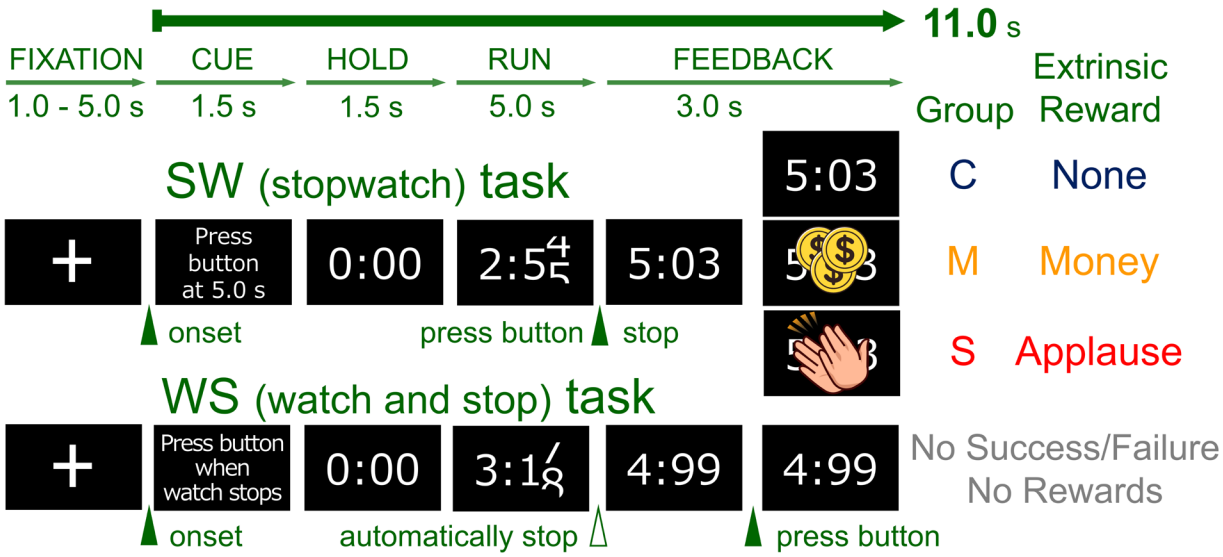

Figure 1. Trial Timeline for the Stopwatch (SW) and Watch-and-Stop (WS) Tasks.

Each trial lasted 11.0 s. In the SW task, participants aimed to stop the watch exactly at 5.0 s; in the WS task, they responded after the watch stopped automatically (approximately 5.0 s). The WS task provided no success/failure feedback or rewards. In Experiment phase 1, participants in Group C performed with no extrinsic reward upon success, Group M with a monetary reward, and Group S with a social reward (applause-like sound). In Experiment phase 2, all groups performed the SW task without extrinsic rewards. Each phase included 60 trials (30 SW and 30 WS) presented in randomized order, with pseudorandomized intertrial intervals (1.0–5.0 s) to reduce predictability.

2.5. fMRI Data Acquisition

During the fMRI, visual stimuli were projected onto a screen in the MRI scanner room through a liquid-crystal display projector (CP-SX12000J; Hitachi Ltd., Tokyo, Japan). Participants in the scanner looked at the screen via a mirror in front of their faces. Participants' responses were recorded using an MRI-contingent button device (HHSC1×4-D; Current Designs, Inc., Philadelphia, PA, USA) held in the right hand. fMRI data were acquired using a 3-Tesla Siemens Verio scanner equipped with a 32-channel head coil. A T2*-weighted echo-planar

imaging sequence was employed (TR = 1.0 s; TE = 31 ms; flip angle = 55°; matrix = 104 × 104; field of view = 208 × 208 mm; slice thickness = 2 mm; 72 slices; echo spacing (ES) = 0.75 ms; bandwidth (BW) = 1,850 Hz/pixel). A multiband sequence acquired eight slices simultaneously per TR [38]. Each run, repeated twice, lasted approximately 15 min and yielded 825 volumes.

2.6. fMRI Data Preprocessing and Statistical Analysis

The data were preprocessed and analyzed using SPM12 (Wellcome Centre for Human Neuroimaging, UCL) in MATLAB 2023a (MathWorks Inc., Natick, MA, USA). The first four volumes of each functional run were discarded to achieve magnetic equilibrium. The remaining images were realigned for motion correction. The mean functional image (computed after realignment) was then directly normalized to the Montreal Neurological Institute (MNI) EPI template [39–40] using a nonlinear basis function. The normalized images were spatially smoothed with an 8-mm full-width-at-half-maximum (FWHM) Gaussian kernel. A high-pass filter with a cutoff of 128 s removed low-frequency drifts [41]. To confirm that the 8-mm smoothing did not induce partial-volume mixing in small nuclei such as the ventral tegmental area (VTA) and pallidum, we repeated the region-of-interest (ROI) extraction using a 4-mm FWHM kernel. The β estimates from the 4-mm and 8-mm kernels were highly correlated (r = 0.80–0.99 across ROIs), and all main results were preserved, indicating that smoothing-related mixing did not materially affect the conclusions (see Supplementary Material S4).

Liking was indexed using the SW–WS contrasts. The WS task was identical to the SW task in visual input, motor output, and timing but intentionally omitted the game-like features that elicit intrinsic motivation.

Subtracting WS from SW removed non-motivational sensory and motor components, isolating neural activity specific to the intrinsic reward of the SW task, as outlined by Murayama et al. [26]. Because the task type (SW or WS) was randomly determined in advance (not chosen by participants), motivational drive (*Wanting*) was largely confined to the pre-cue phase and minimized during subsequent task execution. During the task execution period, participants' attention was necessarily focused on the stopwatch timing, which further reduced anticipatory *Wanting* for the subsequent trial. Under this design, BOLD responses in reward-related regions associated with opioid-mediated hedonic processing were used as a neural proxy for the emotional/feeling component of *Liking*. This neural *Liking* was measured while minimizing potential contamination from task-related *Wanting*, which occurs before the *Liking* measurement epoch. Behavioral *Wanting* was separately measured during the free-choice period (i.e., number of SW task plays).

At the first level, a general linear model (GLM) was specified [42], including six motion parameters as covariates. Task epochs, from cue onset to feedback offset (11.0 s; see Figure 1), were modeled as single boxcar functions, consistent with approaches treating intrinsic reward as a coherent action–outcome episode [24]. This approach was adopted because, given the properties of the hemodynamic response function, separating the short consecutive epochs within each trial would lead to unavoidable multicollinearity and statistically unstable estimates [43–44]. Conceptually, the SW–WS contrast captures the action–outcome contingency that defines the SW task, which constitutes a coherent psychological sequence from cue to feedback. Contrasts comparing the SW and WS trials within each scan session (1SW > 1WS and 2SW > 2WS) were defined as indices of intrinsic reward. Boxcar regressors were convolved with the canonical hemodynamic response function, and

parameter estimates were obtained using SPM's prewhitening procedure. Because the traditional AR (1) + white-noise model can fail to adequately whiten fast-TR data, temporal autocorrelations were modeled using the FAST approach [45–46], estimated from pooled active voxels, and subsequently used to whiten the data.

For ROI analyses of the reward-related responses, we used the individual-level contrast images (con*.nii) for each participant and both sessions. The basal ganglia regions were defined using the high-resolution anatomical atlas of Pauli et al. [47], which provides separate NIfTI (Neuroimaging Informatics Technology Initiative) masks for the bilateral VTA, NAcc, putamen (PUT), external globus pallidus (GPe), internal globus pallidus (GPi), and caudate nucleus (CN) (12 regions in total). For each ROI, all voxel-wise contrast estimates were averaged, and then the values from the left and right hemispheres were further averaged to obtain a single representative signal. Contrast estimates were extracted using MarsBaR (version 0.45) [48], an established ROI toolbox for SPM. This approach captures subtle activation patterns that whole-brain analyses may miss [47,49].

2.7. Statistical Analysis

Neural and behavioral data were analyzed using R software (version 4.4.1). Neural *Liking* was defined as the SW–WS contrast estimate at the first-level GLM. Contrast values were extracted for six a priori reward-related ROIs using MarsBaR and averaged to derive an ROI-mean *Liking* score. An exploratory analysis of neural measures was conducted for each reward condition and each ROI using repeated-measures ANOVA with the Experiment phase as the independent variable. Changes across Experiment phases were then computed as

Δ*Liking* = Experiment phase 2 − Experiment phase 1. Group differences in Δ*Liking* were evaluated using linear models with planned difference-in-differences (DiD) contrasts (ΔM−ΔC, ΔS−ΔC). As a sensitivity analysis, Bayesian linear mixed models (LMM) with participant-level random intercepts were also fitted, and 95% highest posterior density (HPD) intervals—equivalent to 95% credible intervals (CrI) when posterior distributions are approximately symmetric—were computed to quantify estimation uncertainty. Bayesian estimation was implemented to appropriately model parameter uncertainty and to obtain stable estimates in small-sample contexts [50]. Model fit was evaluated using information criteria (AIC, BIC) obtained from the frequentist LMMs, and an intermediate model without interactions was selected. Detailed analysis settings and the theoretical rationale are provided in Supplementary Material S5.

Behavioral *Wanting* was analyzed as the number of voluntary SW plays (0–14). Based on the observed distribution of this count data, three candidate models were considered: Poisson, zero-inflated Poisson (ZIP), and zero-inflated binomial (ZIB). The consideration of ZIP and ZIB was motivated by the presence of excess zeros relative to a standard Poisson distribution. Model fit was evaluated using leave-one-out cross-validation (LOO). The comparison indicated that both the ZIP and ZIB models provided superior predictive performance to the Poisson model, with ZIP showing the highest expected log predictive density (elpd_loo); however, the difference between ZIP and ZIB was small ($|z| < 2$), indicating comparable goodness of fit. Given the small difference, the ZIB model was selected as the primary analytic model because it more accurately reflects the underlying behavioral structure of the task. Specifically, structural zeros arise when participants choose not to interact with the PC at all, whereas binomial choice processes characterize PC users who repeatedly choose

between SW and WS on each trial. The ZIB model explicitly accommodates both mechanisms, making it theoretically most coherent with the data-generating process. Therefore, the ZIB model was retained as the main analytic framework (see Supplementary Material S6).

The ZIB mixed-effects model included random intercepts for participants and fixed effects for reward condition (C/M/S), Experiment phase (1/2), and their interaction. Phase differences (Δ = Experiment phase 2 − Experiment phase 1) were computed from the marginal posterior means. Group-level effects were assessed using planned DiD contrasts (ΔM−ΔC, ΔS−ΔC). Bayesian estimation was implemented via MCMC sampling (4,000 iterations, 2,000 burn-in, 4 chains), and posterior means and 95% HPD intervals—equivalent to CrI under symmetric posteriors—were reported. All model specifications and priors followed standard recommendations for mixed-effects modeling in small-sample behavioral data.

2.8. Performance Accuracy as a Potential Confound

To assess whether task performance confounded our motivational indices, we examined the influence of success rate on *Liking* and *Wanting*, and found that performance accuracy did not account for the observed neural or behavioral effects. Success rate was defined as the proportion of successful SW trials during each fMRI session (per participant, Experiment phase, and group). At the neural level, we compared mixed-effects models for *Liking* using ROI-mean SW–WS contrast estimates (β) with and without standardized success rate (*zSuc*) and its interaction with Sess (Experiment phase): $Liking_{mean} \sim$ Cond × Sess [+ *zSuc* + Sess:*zSuc*] + (1|PtsID); Analogous models were fitted for Δ*Liking* and ROI-wise Δ*Liking*. At the behavioral and subjective levels, we

computed Spearman correlations between success rate and neural *Liking* (mean SW–WS β across VTA, NAcc, GPi, GPe, PUT, and CN), subjective *Liking* (post-task questionnaire ratings), and *Wanting* (number of SW plays during the free-choice period). Full results of these supplementary analyses are reported in Supplementary Materials S7 (model comparisons) and S8 (correlation matrices).

2.9. Questionnaire Survey

After completing the experiment, participants rated their enjoyment of the SW task using two self-report items on a 5-point Likert scale (1 = not at all; 5 = very much). The mean of these items was computed as the subjective *Liking* score [2]. Additionally, four constructs—autonomy, competence, relatedness, and expectancy—were assessed using two items based on established scales [51]. For detailed item wording and a full list of all ten items, see Supplementary Material S9. These subjective ratings were collected as supplementary information to support interpretation but were not considered the primary outcome measures of the study. Because the questionnaire was administered after the completion of all tasks, subjective *liking* ratings reflected retrospective evaluations rather than online hedonic experience during task engagement. Prior research shows that retrospective affective judgments can diverge from real-time emotional responses due to memory reconstruction, belief-based inferences, and peak–end biases [52–53]. In contrast, neural activation in hedonic/reward-related regions provides an index of immediate hedonic impact ("core *liking*"), which does not necessarily correspond to consciously accessible subjective reports [32]. Consistent with earlier work demonstrating dissociations among subjective, behavioral, and neural indicators of intrinsic motivation [26],

subjective ratings in the present study were used as a secondary complementary measure, whereas neural activation served as the primary measure of *liking*.

3. Results

3.1. fMRI Results

Figure 2 illustrates changes in per-trial fMRI contrast estimates (SW–WS) across six ROIs between Experiment phases 1 and 2, representing *Liking*—the neural index corresponding to the behavioral measure of *Wanting* (SW task selections). The ROIs were delineated using the high-resolution atlas by Pauli et al. [47]. In all reward conditions and ROIs—except for the PUT in Group S at Experiment phase 1 (95% confidence interval (CI) [-0.01, 0.04])—the lower bound of the 95% CI exceeded zero, indicating positive *Liking* during task performance (see Supplementary Material S10 for ROI-specific CIs).

Repeated-measures ANOVAs on Experiment phase effects for each group revealed a significant main effect of Experiment phase in Group C, with *Liking* decreasing from Experiment phase 1 to Experiment phase 2 ($F(1, 17) = 6.02$, $p < .05$, $\eta_p^2 = .26$). No significant phase effects were found in Group M ($F(1, 17) = .57$, $p = .46$, $\eta_p^2 = .03$) or Group S ($F(1, 17) = .55$, $p = .47$, $\eta_p^2 = .03$). These results indicate a decrease in *Liking* only in Group C. All groups exhibited a significant main effect of ROI: Group C ($F(2.05, 34.88) = 7.61$, $p < .01$, $\eta_p^2 = .31$), Group M ($F(2.14, 36.44) = 8.78$, $p < .001$, $\eta_p^2 = .34$), and Group S ($F(2.13, 36.13) = 14.54$, $p < .001$, $\eta_p^2 = .46$). However, confirmatory subjective *Liking* ratings obtained after the experiment showed the opposite pattern: *Liking* persisted in Group C but declined in Groups M and S. This dissociation from the neural *Liking*

results highlights the limitations of self-report measures (see Supplementary Material S11 for details). Significant Experiment phase × ROI interactions emerged in Group M ($F(2.23, 37.94) = 3.23$, $p < .05$, $\eta_p^2 = .16$) and Group S ($F(1.50, 25.52) = 8.53$, $p < .01$, $\eta_p^2 = .33$), suggesting that the pattern of neural responses across ROIs changed differentially following reward withdrawal.

Since Experiment phase 2 employed an identical no-reward condition across all groups (see Table 1), the difference score of *Liking* (Experiment phase 2 – Experiment phase 1) corresponds to the extrinsic reward effect for each condition. Comparing these scores between Groups C and M or S provides a valid evaluation of whether monetary and social rewards influence neural *Liking* compared to the non-extrinsic reward condition. This design allowed us to control for repetition effects and the presence or absence of extrinsic rewards on intrinsic rewards. The following section reports the analyses of these *difference scores* to enable more stringent comparisons across the reward conditions.

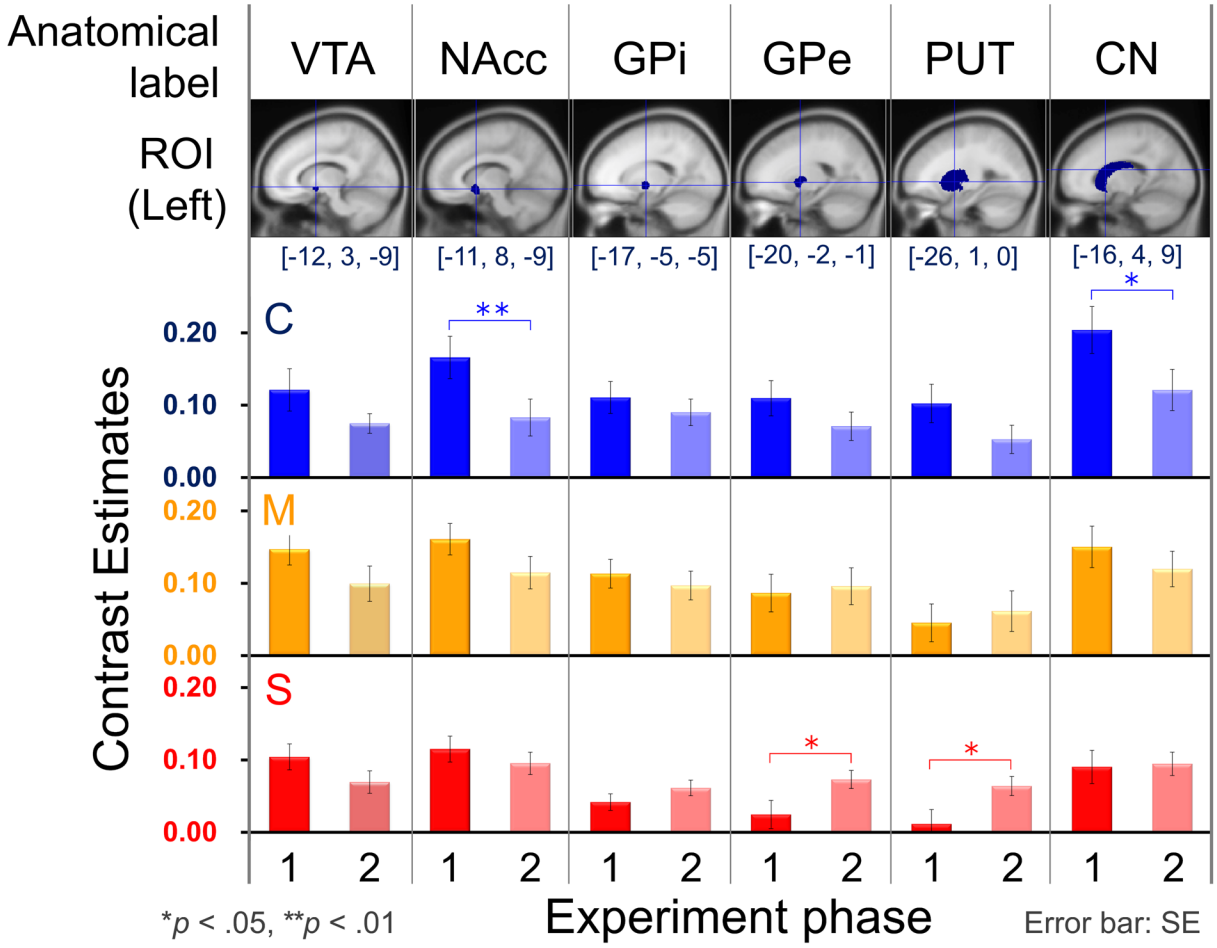

Figure 2. *Liking* by ROI and Experiment Phase.

ROIs: VTA = ventral tegmental area; NAcc = nucleus accumbens; GPi/GPe = internal/external segments of the globus pallidus; PUT = putamen; CN = caudate nucleus. Groups: C = control group (n = 18), M = monetary reward group (n = 18), S = social reward group (n = 18). Experiment phases: 1 = Experiment phase 1 (reward presentation), 2 = Experiment phase 2 (reward withdrawal). Contrast estimates (SW > WS) are presented for each region of interest (ROI) in Experiment phase 1 and Experiment phase 2 across the three reward conditions. Sagittal slice images for each ROI are illustrated using left-hemisphere data from Pauli et al. [47], with Montreal Neurological Institute (MNI) coordinates indicating the center of mass. *p < .05, **p < .01 (unadjusted p values for within-condition Experiment phase effects [Experiment phase 2 vs. Experiment phase 1] tested separately for each ROI; see Supplementary Material S12).

3.2. Difference Score Analysis

To assess changes in neural *Liking* while controlling for the no-reward condition, difference scores (contrast estimates in Experiment phase 2 minus those in Experiment phase 1) were analyzed. The normality of these scores was confirmed by box plots (Supplementary Material S13). A two-way ANOVA with reward condition (Group C/M/S) and ROI as factors revealed a significant main effect of reward condition ($F(2,306) = 9.78$, $p < .001$, $\eta_p^2 = .06$), indicating that changes in *Liking* differed depending on reward presence. The main effect of ROI was significant but small in magnitude ($F(5,306) = 2.93$, $p < .05$, $\eta_p^2 = .04$), and the reward condition × ROI interaction was non-significant ($F(10,306) = 0.53$, $p = .87$, $\eta_p^2 = .02$), suggesting a consistent pattern across ROIs. Post hoc Holm-corrected tests revealed that Group C had significantly lower difference scores than Group M ($p < .05$) and Group S ($p < .001$). The comparison between Group C and the combined means of Groups M and S was also significant ($p < .001$), whereas Groups M and S did not differ ($p = .11$).

To quantify group differences more directly, we conducted planned DiD contrasts (Δ = Experiment phase 2 – Experiment phase 1). ROI-averaged linear mixed-model contrasts showed significantly higher Δ values for both reward groups relative to the control group (ΔM−ΔC = 0.04, SE = 0.02, 95% CI [0.01, 0.07], $p < .01$; ΔS−ΔC = 0.07, SE = 0.02, 95% CI [0.04, 0.09], $p < .001$). These contrasts indicate that the monetary and social reward conditions preserved neural *Liking* more robustly across Experiment phases than the no-reward condition and thereby supporting H1.

To further validate these findings and quantify uncertainty, an LMM was applied, with reward condition (C/M/S) and Experiment phase (1/2) as fixed effects and participants as random intercepts (see Supplementary

Material S14 for ROI- and condition-specific estimates). Figure 3 shows a forest plot of the posterior means and 95% CrIs for each ROI. In Group C, all posterior means were negative. For the VTA, NAcc, and CN, the upper bounds of the 95% CrIs were entirely below zero, as shown in the plot as points and bars fully below the zero line, indicating a credible decline in *Liking*. In contrast, for Groups M and S, all 95% CrIs crossed zero, with most means near or above zero, suggesting no credible change in *Liking* after reward withdrawal. This Bayesian evidence, consistent with the ANOVA results, supports H1 and indicates that both monetary and social rewards preserve neural *Liking* even after extrinsic rewards are removed. This pattern suggests a potential effect of extrinsic reward, which we refer to as the "*preserving effect*."

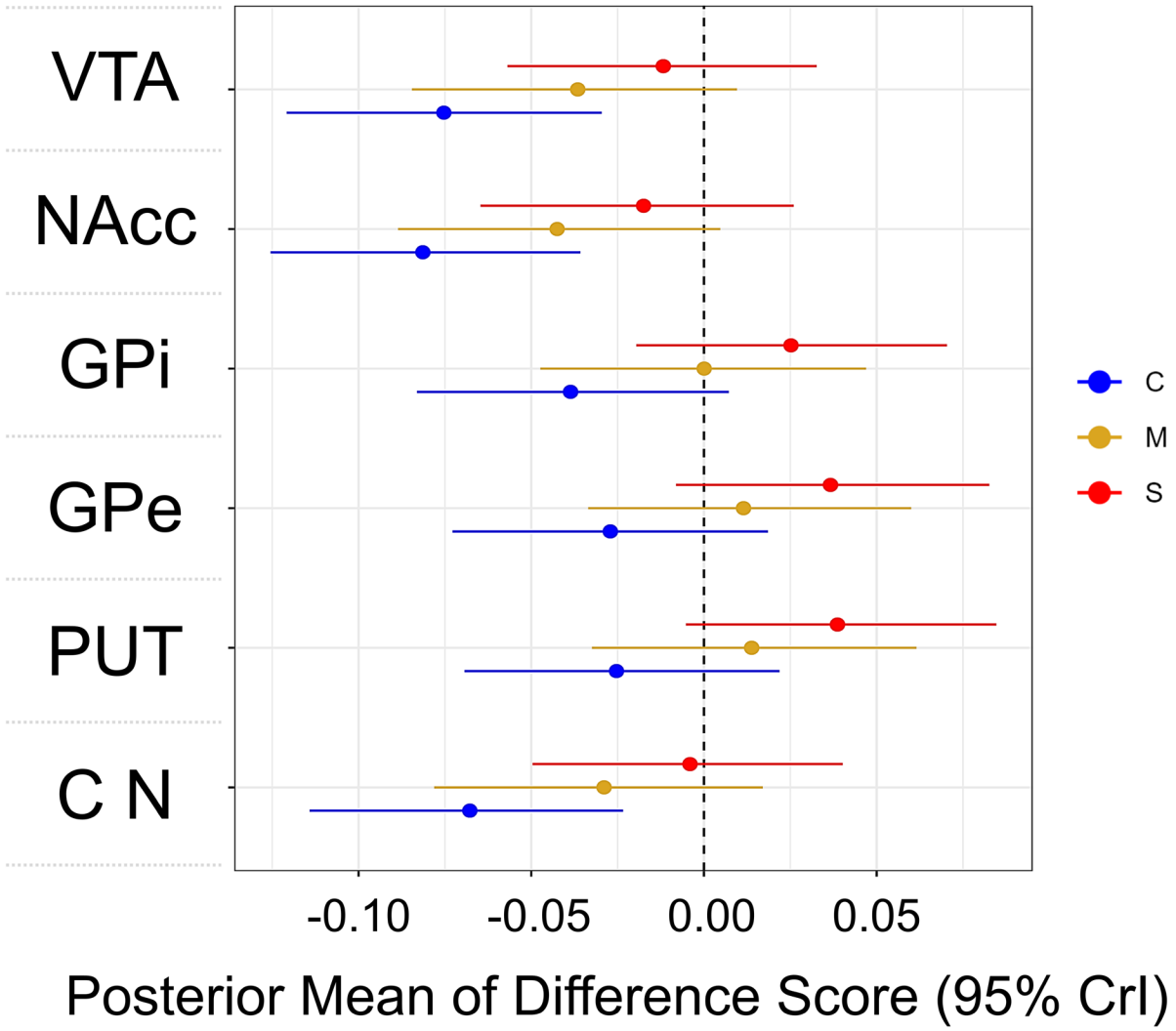

Figure 3. Forest Plot of Difference Scores by ROI and Reward Condition (Bayesian posterior estimates).

Each point represents the posterior mean of the difference score (Experiment phase 2 – Experiment phase 1), and each horizontal line shows the 95% CrI. Positive values indicate increased *liking* in Experiment phase 2 relative to Experiment phase 1, whereas negative values indicate decreased *liking*. Groups: C = control group (n = 18), M = monetary reward group (n = 18), S = social reward group (n = 18). A Bayesian linear mixed model was fitted with fixed effects for reward condition (C/M/S) and Experiment phase (1/2) and random intercepts for participants, estimated in *brms* (Stan) over 4,000 iterations (2,000 burn-in, 4 chains; $\hat{R}$ = 1.00).

3.3. Behavioral Results

Figure 4 (left panel) illustrates the distribution of free-choice SW plays across Experiment phases. Because these counts were bounded (0–14) and strongly zero-inflated, *Wanting* was analyzed using a ZIB mixed-effects model. At the group-comparative level, planned DiD contrasts showed no reliable differences relative to Group C ($\Delta M - \Delta C = -0.61$, 95% CrI [−2.78, 1.56]; $\Delta S - \Delta C = -0.28$, 95% CrI [−2.45, 1.89]). Thus, reward withdrawal did not produce a statistically robust group-level change when evaluated against the no-reward condition. The posterior estimates of within-group phase changes (Figure 4, right panel) revealed a more differentiated pattern. The Δ*Wanting* values were: Group C = −1.17 (95% [−2.96, 0.19]), Group M = −1.80 (95% CrI [−4.01, −0.45]), and Group S = −1.42 (95% CrI [−3.44, 0.17]). Only Group M showed a 95% CrI entirely below zero, indicating a credible decline in Wanting. In contrast, both Groups C and S exhibited CrIs spanning zero, reflecting statistically indeterminate changes. Bayes factors corroborated this pattern: substantial evidence for a phase effect in Group M ($BF_{10} \approx 8.55 > 3$; directional $BF_{10} \approx 399 > 100$), but not in Groups C or S ($BF_{10} < 1$) [50,54–55].

Considered together, the DiD contrasts indicate that the decline in Group M was not clearly larger than that in Group C at the conservative group-comparative level. Nonetheless, the Bayesian within-group estimates—directly visualized in Figure 4—show that *Wanting* decreased selectively and reliably only in Group M. This pattern provides cautious but convergent support for H2, suggesting that monetary reward withdrawal induces a reliable reduction in behavioral *Wanting*, whereas the no-reward and social-reward conditions do not.

Because WS plays were extremely rare and exhibited pronounced zero inflation, we additionally summarized the proportion of participants who never selected the WS task in each free-choice period. In Free-choice Period 1, the proportions were 94% in Group C (17/18, 95% CI [0.73, 1.00]), 100% in Group M (18/18, 95% CI [0.82, 1.00]), and 78% in Group S (14/18, 95% CI [0.52, 0.94]). In Free-choice Period 2, the corresponding proportions were 72% in Group C (13/18, 95% CI [0.47, 0.90]), 89% in Group M (16/18, 95% CI [0.65, 0.99]), and 89% in Group S (16/18, 95% CI [0.65, 0.99]). Mean WS play counts remained well below one trial per period (Free-choice Period 1: C = 0.06 ± 0.24, M = 0.00 ± 0.00, S = 0.33 ± 0.69; Free-choice Period 2: C = 0.61 ± 1.33, M = 0.22 ± 0.73, S = 0.11 ± 0.32). These results confirm that WS engagement was minimal and that the WS task functioned effectively as a low-motivation motor control.

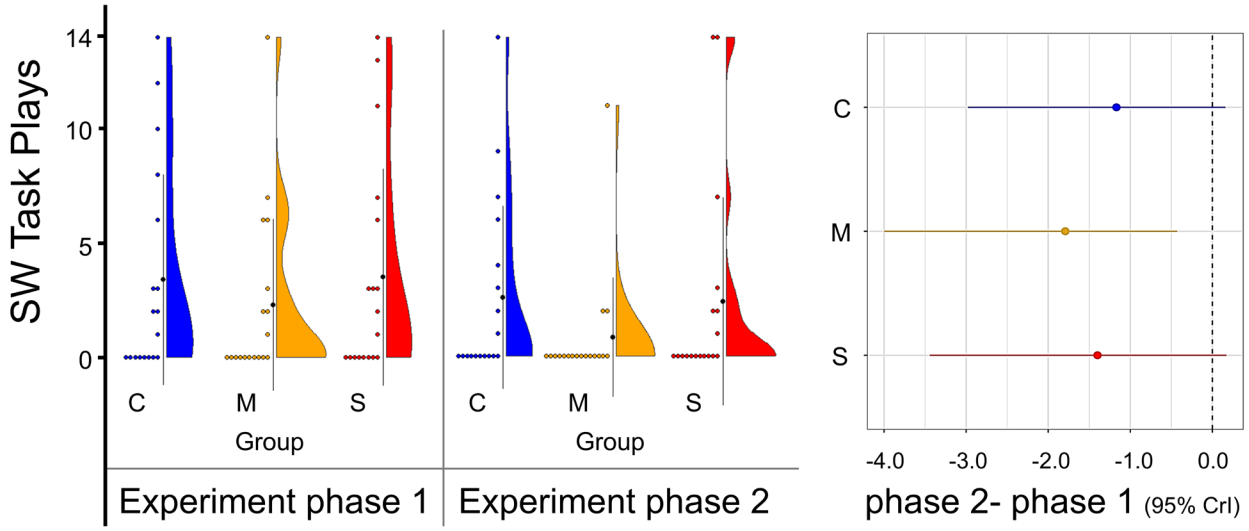

Figure 4. Results of Free-Choice SW Task Plays (*Wanting*).

Groups: C = control group (n = 18), M = monetary reward group (n = 18), S = social reward group (n = 18).

Left panel: Dot plots and violin plots show the distribution of SW task plays for each group during Experiment phase 1 and Experiment phase 2. Dot plot showing group means (black dots), standard deviations (error bars) and individual

data (colored points). Violin plot visualizing the distribution of SW task plays. The possible range of SW task plays was 0–14. **Right panel**: Bayesian posterior estimates of group-level changes in SW task plays (*Experiment phase 2 − Experiment phase 1*). Each point represents the posterior mean, and each horizontal line indicates the 95% credible interval estimated from a zero-inflated binomial mixed-effects model (fixed effects: group and experiment phase; random intercepts for participants). MCMC sampling was performed over 4,000 iterations (2,000 burn-in, 4 chains; $\hat{R}$ = 1.00). The vertical dashed line denotes no change. Positive values indicate increased free-choice SW plays (greater behavioral *Wanting*) from Experiment phase 1 to phase 2, whereas negative values indicate decreased *Wanting*.

3.4. Brain–Behavior Relationships

Spearman's rank correlation analyses were conducted on the Experiment phase-wise difference scores for *Liking* and *Wanting* to assess the associations. As shown in Table 2, correlations were small ($|r| < .31$) and non-significant ($p > .10$) across all reward conditions, suggesting independence between neural *Liking* changes and behavioral *Wanting* changes. Confirmatory analyses of the control WS task yielded no significant correlations (Supplementary Material S15-16).

Table 2

Spearman's Rank Correlations Between Experiment Phase-Wise Difference Scores of Liking and Wanting for Each Reward Condition

Group	VTA	NAcc	GPi	GPe	PUT	CN	Ave
C	.11	−.31	−.17	−.20	−.18	−.31	−.21
M	−.04	.02	.16	.12	.07	.02	.22
S	−.19	−.21	.07	−.01	.00	−.16	−.07
All	−.02	−.18	.02	−.06	−.06	−.16	−.10

***Note*.** This table presents the Spearman's rank correlation coefficients between behavioral changes (difference in SW task plays, *Wanting*) and neural changes (difference in ROI contrast estimates, *Liking*) from Experiment phase 1 to Experiment phase 2. ROIs: VTA, ventral tegmental area; NAcc, nucleus accumbens; GPi/GPe, internal/external segments of the globus pallidus; PUT, putamen; CN, caudate nucleus; Group C, control group (n = 18); Group M, monetary reward group (n = 18); Group S, social reward group (n = 18); All, all participants (N = 54). *Ave* = mean *Liking* change across the six ROIs. No correlations reached significance in any condition ($|r| < .31$, $p > .10$).

To evaluate whether task performance confounded the motivational indices, we conducted supplementary model-based and correlational analyses incorporating success rate. Hierarchical model comparisons showed that adding standardized success rate (*zSuc*) or its change (*ΔSuc*) did not improve the fit of any neural *Liking* model, including ROI-mean *Liking* (|Δelpd| = 2.0), *ΔLiking* (|Δelpd| = 1.1), and ROI-wise *ΔLiking* (|Δelpd| ≈ 0; Supplementary Material S7). Consistent with these model-based results, Spearman correlations between success rate and neural *Liking* were generally small and often nonsignificant across ROIs, reward groups, and Experiment phases (−.39 ≤ r ≤ .08, no

systematic pattern; Supplementary Material S8). Correlations with *Wanting* were more heterogeneous ($-.31 \leq r \leq .61$): a strong positive association was observed only in Group C during Experiment phase 1 ($r = .61$, $p < .01$), whereas other groups and phases showed small or nonsignificant relationships that did not mirror the main group-by-phase effects. In contrast, success rate exhibited moderate-to-strong positive correlations with retrospective subjective *Liking* in some conditions (e.g., $r = .44$–$.69$), indicating that performance accuracy modestly biased post hoc evaluations but did not track the neural or behavioral indices that constituted our primary outcomes. Taken together, these converging findings suggest that task performance does not account for the key neural and behavioral motivational effects reported above.

4. Discussion

4.1. Summary of the Main Findings

We found that withdrawal of extrinsic rewards (monetary and social) did not affect the task-related activation of the reward circuit compared with Group C, in which repetition suppression occurred. The voluntary performance of the SW task during the free-choice period suggested the undermining effect only in Group M. These findings provide support for **H1 (*liking*)** and offer cautious but convergent support for **H2 (*wanting*)**, consistent with our theoretical hypothesis that *different categories of extrinsic rewards influence liking and wanting differently*, challenging the long-standing assumption that *liking* and *wanting* co-vary in parallel and indicating category-dependent diversity in the manifestation of undermining effects.

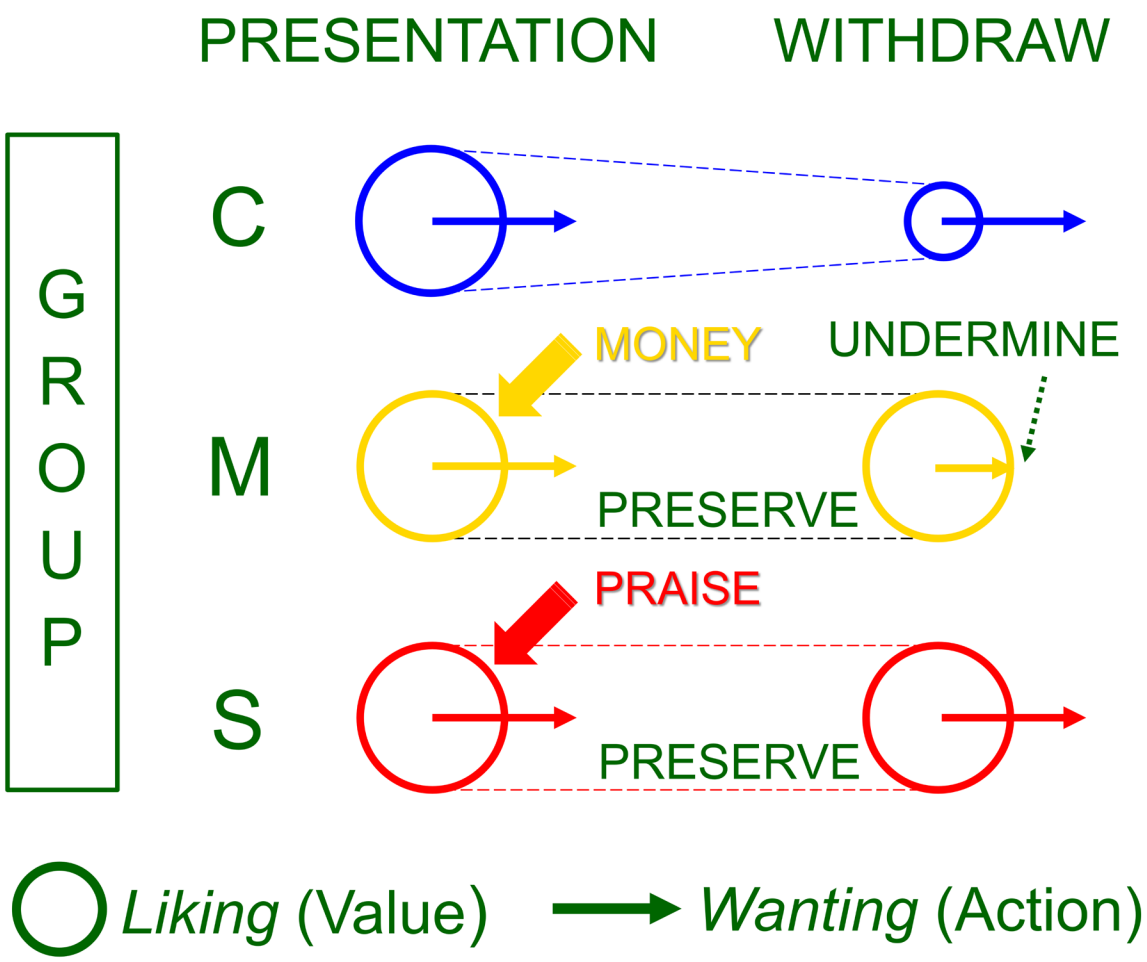

Figure 5. Simplified Model of *Liking* and *Wanting* Changes.

Changes in *Liking* (intrinsic reward; circles) and *Wanting* (motivational drive; arrows) from Experiment phase 1 (PRESENTATION) to Experiment phase 2 (WITHDRAW) are illustrated for C (control group; $n = 18$), M (monetary reward group; $n = 18$), and S (social reward group; $n = 18$). In Group C, *Liking* decreased after reward withdrawal while *Wanting* was sustained, demonstrating a dissociation between the two. In contrast, both Group M and S exhibited sustained *Liking* (preserving effect), but *Wanting* decreased in Group M (undermining effect) and was sustained in Group S. These results indicate different patterns of dissociation between *Liking* and *Wanting* depending on the reward category.

4.2. Overall Interpretation of the Results

Neural *Liking*. *Liking* in Group C decreased from Experiment phase 1 to Experiment phase 2, while in Group M and Group S, it remained relatively stable compared to Group C after extrinsic reward withdrawal. The decrease in Group C is consistent with repetition suppression from repeated exposure to stable action–outcome contingencies [56–60], in accordance with fMRI evidence showing that BOLD responses in regions such as the NAcc and VTA decline when reward-related stimuli are repeatedly presented [58–60]. In contrast, no significant decrease in task-related activation was observed in Groups M and S, both of which showed decreases that were significantly less than that of Group C. Given that Group C did not receive any extrinsic reward, the reward circuitry activation during the SW task, compared with the control WS task, represents the intrinsic reward value (*liking*) of the SW task, and the repetitive involvement of the task induced the repetition suppression. Therefore, the smaller decline of *Liking* observed in Groups M and S appeared to be related to the presence of extrinsic reward in the first phase.

The contrast effect [16,61–62], which explains the undermining effect by the amount of the total reward value (both intrinsic and extrinsic), does not likely explain the results, because despite presumed repetition suppression and a decrease in total reward value, the activation did not change. A more feasible interpretation is that symbolic cues (e.g., successful outcomes enhanced by extrinsic reward) can sustain or recall hedonic value after reward withdrawal [35,63]. This interpretation also aligns with neural models suggesting that symbolic cues contribute to *liking* independently of monetary incentives [64] and with evaluative conditioning theories predicting the maintenance of *liking* over time [65]. While Group C likely experienced saturation and

decay of task value through repeated exposure, participants in Groups M and S may have undergone value updating and stabilization during reward presentation. Symbolic cues in the withdrawal phase may have facilitated the reactivation of the hedonic experience. Given that reward responses in the first phase represent the sum of internal and external rewards, and the second phase lacks the external reward, this finding can be regarded as evidence of the internalization of the extrinsic reward [35].

Harackiewicz [63] argued that monetary rewards, when contingent on performance, could serve as symbolic cues, signifying recognition of effort and achievement. She further proposed that related cues presented after reward withdrawal (e.g., successful outcomes) may evoke hedonic experiences, helping to sustain intrinsic motivation. Although this view has not yet been extensively tested in psychological research, it aligns with the present study and other neural evidence previously reported. For instance, an fMRI study that orthogonally manipulated monetary rewards and symbolic cues (performance or accuracy feedback) revealed selective activation in the anterior and posterior cingulate cortices and the dorsal striatum, while the ventral striatum responded to both types of information [64]. Moreover, within the framework of evaluative conditioning, hedonic affect accompanying rewards can transfer through subsequent task-related cues (symbolic cues) to the conditioned stimulus. This transfer occurs via activation of memory traces (“echo”) in the reward circuitry, preserving affective value even after the original expectation has dissipated [65–66]. This emotion-based reactivation pathway contrasts with the cognition-based pathway described in Cognitive Evaluation Theory, a sub-theory of SDT, where the informational component of verbal feedback enhances perceived competence through cognitive processing [5,13]. Together, these perspectives highlight the need to combine

behavioral measures with neural indices, such as fMRI, to examine how psychological and neural states evolve from reward presentation to withdrawal.

Behavioral *Wanting*. In Group C, *Wanting* remained stable, indicating that repeated task engagement without extrinsic rewards neither enhanced nor diminished motivation. A possible SDT-based interpretation is that the absence of evaluative context preserved *Wanting*. Figure 5 shows that *Wanting* credibly decreased from Experiment phase 1 to Experiment phase 2 in Group M (supported by Bayesian analyses). This is consistent with the classic undermining effect, whereby performance-contingent monetary rewards can act as controlling incentives that reduce autonomy and subsequent motivational drive [5,26]. As Group S showed no undermining effect, the contrast effect cannot explain the monetary reward-specific undermining effect. Instead, this finding is consistent with SDT, suggesting that monetary rewards undermine intrinsic motivation by diminishing perceived autonomy when rewards are experienced as evaluative [2,5,8,16,26]. However, as we did not observe reliable group-level differences, we refrain from further interpretation.

4.3. Theoretical Organization and Value–Action Dynamics

Prior research on intrinsic motivation has often conflated *liking* (intrinsic *value*) and *wanting* (behavioral *action*) as a unitary construct, leading to inconsistent findings [2,14,26]. Figure 5 highlights the theoretical implication of this study: *Liking* (*value*) and *Wanting* (*action*) do not necessarily change in parallel, challenging the long-standing assumption that intrinsic motivation forms a unified construct [2,3]. This aligns with neurobehavioral frameworks indicating that *Liking* and *Wanting* rely on partly dissociable neural systems [29,30].

In Group C, the clearest dissociation was observed: *Wanting* was preserved while *Liking* declined. Behavioral interpretations typically see the control condition as indicating preserved intrinsic motivation, yet the neural data show that stable *action* does not guarantee stable *value* [16,67]. The two extrinsic-reward groups showed differentiated patterns. In Group M, *Liking* was preserved whereas *Wanting* showed a modest decline. Although the decline observed in this study was not pronounced, this pattern is broadly consistent with well-established undermining effects of monetary rewards [5,7,13,14,20]. In Group S, both *Liking* and *Wanting* were preserved, a pattern that accords with evidence that relatedness-supportive feedback facilitates internalization [4,11] and that SDT views socially grounded feedback as promoting autonomous motivation through internalization processes [2,3]. This internalization-based account is also compatible with recent proposals that extrinsic and intrinsic rewards may operate additively rather than competitively, offering a complementary perspective on how preserved value can support subsequent action [68–69]. To organize this pattern, we introduce Value–Action Dynamics (VA-Dynamics) as a descriptive framework. VA-Dynamics does not posit causality but clarifies why preserved value does not necessarily translate into preserved action, offering a conceptual foundation for future studies examining how value and action become coupled or decoupled across reward contexts.

4.4. Methodological Considerations

The present study adopted neural responses of reward circuitry during task engagement as a measure of *Liking*, and *Wanting* was measured during a subsequent free-choice period. In contrast to *wanting*, which can be

quantified through observable behaviors (e.g., free-choice tasks), assessing *liking* has posed substantial challenges. The predominant self-report method is highly susceptible to recall bias, social desirability bias, and participants' tendency to conflate *liking* with *wanting* [3,67,70]. Objective indicators of *liking*, such as transient facial expressions or micro-behaviors, are also problematic due to their fleeting nature and questionable validity [28,35,67].

On the other hand, the core regions of the reward circuitry provide a reliable neural basis for assessing *liking*. Coordinate-based Activation Likelihood Estimation (ALE) meta-analyses have shown that a reward network—comprising the VTA, NAcc, PUT, CN, and both segments of the globus pallidus (GPe/GPi)—is consistently activated during *liking* responses [71–72]. Based on the common-currency hypothesis, this network responds reliably to primary rewards (e.g., food, music) and secondary rewards (e.g., monetary, social incentives) [72–75]. Among these regions, the medial shell of the NAcc—a core component of the reward circuitry—acts as a "hedonic hotspot." Animal studies show that stimulating this area enhances *liking* responses [32,76]. In humans, high-resolution positron emission tomography (PET) imaging shows a correlation between momentary pleasure ratings (*liking*) while listening to music and reward-circuitry activity in the NAcc [71–72,75–77]. Additionally, the PUT and CN contribute to transforming reward value into behavioral choice, while the GPe/GPi track reward prediction errors in both magnitude and direction [78–79]. BOLD responses in these regions correlate closely with real-time hedonic experiences, thereby providing a neural proxy for the quantification of *liking*.

Although recent research has identified neural markers of *liking*, earlier neuroimaging studies of *liking* during task performance were likely constrained by limited temporal analysis windows and experimental designs [26–27]. Most investigations of the neural basis of the undermining effect have focused on the feedback phase, when extrinsic rewards are delivered. However, reward processing unfolds across multiple phases—anticipation, execution, outcome evaluation, and feedback—each supported by distinct neural mechanisms [80–81]. The narrow temporal focus is problematic because *liking* can emerge at any point in a task, not only at reward delivery [35]. The present study extends the paradigm of Murayama et al. [26] by modeling the entire task period from the cue to the feedback, and by incorporating manipulations of extrinsic rewards (monetary and social), enabling a more valid real-time assessment of intrinsic rewards and their fluctuations, ultimately advancing motivation research methodology.

4.5. Internal and External Validity

This study ensured internal and external validity through methodological choices tailored to specific challenges in measuring *liking*. Specifically, subjective ratings have inherent limitations as sole indicators. In this study, subjective evaluations did not significantly correlate with either behavioral or neural measures (see Supplementary Material S17 for details). Therefore, based on previous research, neural indices of *liking* were adopted as the primary outcomes. Reward-related neural activity in the ventral striatum and related regions has been shown to reflect hedonic valuation [32,71–72]. In addition, internal validity was strengthened by randomly assigning participants to minimize selection bias [82] and integrating behavioral and neural measures to ensure

consistency across indices [49,50,56]. ROIs were defined using the high-resolution atlas by Pauli et al. [47] to target areas involved in reward processing. Mixed-effects modeling combined with Bayesian estimation was applied to quantify the uncertainty in localized fMRI signals, enhancing analytical precision [50,55]. Finally, external validity was bolstered by adopting applause as a nonverbal praise signal across cultures and societies [83–85] and by using an adult sample balanced for age and gender, supporting generalizability to diverse educational and workplace settings. These safeguards provide a robust foundation for applying laboratory findings in real-world contexts.

4.6. Limitations and Future Directions

This study has three main limitations. First, in the control condition, a discrepancy emerged: *Liking* decreased while *Wanting* remained stable. Because the control condition serves as a reference point for comparison with the extrinsic reward conditions, this discrepancy warrants careful consideration. Although it was not the primary aim of the study and was treated as an exploratory finding, the pattern was both unexpected and intriguing, warranting further investigation. Second, the mechanisms by which social rewards update and stabilize value, as well as sustain action, create a pattern that contrasts with the one observed for monetary rewards. Future research should directly compare social and monetary rewards to clarify these differences [52–53,71–72,73–74,86–87]. In this study, nonverbal applause was used; therefore, it remains unclear whether the preserving effect of *liking* generalizes to other feedback forms (e.g., verbal vs. nonverbal cues) and schedules (e.g., continuous vs. partial reinforcement). Systematic variation of these factors will clarify the conditions under

which social rewards uniquely support value stabilization and the coordination of preserved value and action. Third, inferences about regional specificity and causal direction within the reward circuitry are limited by the correlational nature of fMRI and the ROI-based analytic approach. Although the preserving effect was observed, the causal contributions of individual regions remain unverified. Integrating model-based fMRI analyses of value updating, stabilization, and reactivation with effective connectivity analyses (e.g., psychophysiological interaction [PPI] or dynamic causal modeling [DCM]) would provide stronger tests of the VA-Dynamics framework. Addressing these limitations will help clarify the VA-Dynamics and advance the refinement of a framework that conceptually reorganizes the perspectives of SDT and IST.

4.7. Implications and Contributions

This study advances motivational theory by empirically dissociating *liking* and *wanting* within the same task and identifying category-specific patterns in value–action interaction. We propose VA-Dynamics as a framework for interpreting these dissociations. This framework reconceptualizes the undermining effect as a context-dependent modulation of value–action coupling, thereby integrating the explanatory scope of SDT [2–3] and IST [26,30–31]. It also extends the common-currency hypothesis [71–74] by showing that neural value representations can persist beyond reward withdrawal, a phenomenon termed the "*preserving effect,*" which has direct implications for understanding long-term motivational maintenance. Methodologically, combining a two-phase fMRI design with the intrinsic reward paradigm allows for the independent and synchronous measurement of hedonic experience (neural value) and *subsequent* behavioral drive (action) using standardized BOLD

difference indices that enhance reproducibility [26,80]. Practically, the findings provide empirical justification for incorporating culturally robust, nonverbal social feedback, such as applause, in incentive systems to sustain both *liking* and *wanting*. This is particularly relevant in contexts where monetary or verbal incentives are less effective.

4.8. Conclusion

This study shows that the withdrawal of extrinsic rewards produces distinct patterns of *liking* (hedonic experience) and *wanting* (motivational drive) depending on the reward category, establishing an empirical basis for clarifying VA-Dynamics. The identification of the "*preserving effect*," in which neural *liking* persists after extrinsic reward withdrawal, introduces a temporal persistence dimension to existing reward-processing models. These findings establish VA-Dynamics as a theoretical framework for explaining category-dependent motivational outcomes and for guiding the design of interventions that match the reward category to context. This conclusion highlights that behavioral *action* alone does not fully capture motivation and may sometimes reflect only its superficial aspects. By revealing the underlying neural activation and illuminating the intrinsic *value* with fMRI, this study revealed a deeper narrative. From this, the interplay between surface and depth—captured by VA-Dynamics—provides insight into longstanding puzzles and suggests new directions for motivation research.

Acknowledgments

This dissertation was completed under the continuous guidance and support of many individuals. I am deeply indebted to my supervisor, Professor Norihiro Sadato, whose profound expertise in cognitive and affective neuroscience provided an essential foundation for this work. His consistent encouragement of theory-driven and interdisciplinary research, together with his emphasis on theoretical rigor and internal coherence, further shaped the direction of this dissertation. I am also especially grateful to Associate Professor Takahiko Koike, who offered exceptionally devoted technical mentorship throughout the entire research process. His hands-on guidance in experimental design and implementation, fMRI data analysis, and the evaluation and interpretation of results was indispensable to the scientific rigor of this dissertation.

I would like to thank the members of my doctoral dissertation committee for their careful examination and constructive feedback. Professor Hiromasa Takemura, who served as the chief examiner, together with Professor Keiichi Kitajo and Professor Hiroki Tanabe (Graduate School of Information Science, Nagoya University), provided detailed and insightful comments, particularly during the final revision stage, which substantially improved the clarity and completeness of this manuscript. I am also grateful to Specially Appointed Professor Masaki Fukunaga (Project Professor), who served as a committee member and provided important guidance regarding the preparation of the dissertation and the administrative procedures for the degree application.

I gratefully acknowledge the contributions of those who supported the conduct of the experiments and the collection of data. In particular, I thank Dr. Ayumi Yoshioka, Dr. Satoshi Izuno, Dr. Kanae Ogasawara, and

Dr. Shohei Tsuchimoto, who, as co-authors of the published article, primarily contributed as members of the experimental staff and supported the development and implementation of the experiments. I also thank Dr. Takaaki Yoshimoto and Dr. Maho Hashiguchi for their substantial assistance with experimental operations and data collection. In addition, I appreciate the invaluable support of the administrative staff: Ms. Reiko Kimura for comprehensive project administration and coordination of meetings, Ms. Megumi Iwase for assistance with research budget management, and Ms. Kuniko Takenaka for support with participant recruitment.

I also extend my sincere appreciation to the members of Achieve Academy, who participated in the pilot studies and supported the preparatory stages of this research. I am particularly grateful to Ms. Noriko Hanaki, whose continuous administrative assistance and personal support—from research planning through manuscript preparation—were invaluable to the completion of this dissertation. I thank all participants who took part in both the main experiments and the pilot studies for generously devoting their time and effort.

This work was supported by JSPS KAKENHI Grant-in-Aid for JSPS Fellows (Grant Numbers 20J23507 and 24KJ0538) awarded to the author, and by JSPS KAKENHI Grant-in-Aid for Scientific Research (A) (Grant Number 24H00622) awarded to Professor Norihiro Sadato.

Portions of this dissertation are based on the article by Ayabe et al. (2026), “Beyond the undermining effect: Extrinsic rewards preserve neural intrinsic reward” (*Behavioural Brain Research*, 501, 115996) [88]. This dissertation has been independently prepared as a degree thesis. Content and figures derived from the above article are included with appropriate citation, in accordance with Elsevier’s author reuse rights for theses and dissertations.

Glossary

Intrinsic Motivation: The self-driven engagement in activities for interest or enjoyment, without external incentives.

Extrinsic Reward: Incentives provided from outside the individual, such as money or praise, to influence behavior.

Reward Withdrawal: The removal of previously given extrinsic rewards, often used to study motivation changes.

Undermining Effect: A decline in intrinsic motivation that occurs after extrinsic rewards are withdrawn.

Preserving Effect: A newly identified phenomenon where neural evidence of intrinsic reward persists even after extrinsic rewards are withdrawn.

***Liking*:** The hedonic experience or intrinsic value associated with a reward, measured neurally (e.g., via fMRI activity in reward circuits).

***Wanting*:** The motivational drive to pursue or re-engage with an activity, often assessed through free-choice behavior.

Self-Determination Theory (SDT): A motivational framework positing that autonomy, competence, and relatedness are essential for sustaining intrinsic motivation.

Incentive Sensitization Theory (IST): A neuroscientific framework that distinguishes between liking (pleasure) and wanting (motivation), showing they are separable neural processes.

Common-Currency System: The shared neural circuitry (e.g., striatum, ventral tegmental area) that processes both monetary and social rewards.

Intrinsic Reward Paradigm: An fMRI-based task (e.g., stopwatch task) designed to measure neural activity related to intrinsic motivation during task performance.

Free-Choice Period: A behavioral assessment phase in which participants can voluntarily choose whether to continue the target task, indexing motivational drive (*Wanting*).

Value–Action Dynamics (VA-Dynamics): A framework proposed in this study describing how reward categories shape the relationship between neural *liking* (value) and behavioral *wanting* (action).

References

[1] F. Guay, R.J. Vallerand, C. Blanchard, On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (SIMS), *Motiv. Emot.* 24 (2000) 175–213. https://doi.org/10.1023/A:1005614228250

[2] R.M. Ryan, E.L. Deci, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being, *Am. Psychol.* 55 (2000) 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[3] R.M. Ryan, E.L. Deci, Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions, *Contemp. Educ. Psychol.* 61 (2020) 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

[4] A. Van den Broeck, J.L. Howard, Y. Van Vaerenbergh, H. Leroy, M. Gagné, Beyond intrinsic and extrinsic motivation: A meta-analysis on self-determination theory's multidimensional conceptualization of work motivation, *Organ. Psychol. Rev.* 11 (2021) 240–273. https://doi.org/10.1177/20413866211006173

[5] E.L. Deci, R. Koestner, R.M. Ryan, A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation, *Psychol. Bull.* 125 (1999) 627–668. https://doi.org/10.1037/0033-2909.125.6.627

[6] B.S. Frey, R. Jegen, Motivation crowding theory, *J. Econ. Surv.* 15 (2001) 589–611. https://doi.org/10.1111/1467-6419.00150

[7] M.R. Lepper, D. Greene, R.E. Nisbett, Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis, *J. Pers. Soc. Psychol.* 28 (1973) 129–137. https://doi.org/10.1037/h0035519

[8] E.L. Deci, Effects of externally mediated rewards on intrinsic motivation, *J. Pers. Soc. Psychol.* 18 (1971) 105–115. https://doi.org/10.1037/h0030644

[9] B.F. Skinner, *Science and Human Behavior*, Macmillan, New York, 1953.

[10] J. Henderlong, M.R. Lepper, The effects of praise on children's intrinsic motivation: A review and synthesis, *Psychol. Bull.* 128 (2002) 774–795. https://doi.org/10.1037/0033-2909.128.5.774

[11] B. Soenens, M. Vansteenkiste, Understanding the complexity of praise through the lens of self-determination theory, in: E. Brummelman (Ed.), *Psychological Perspectives on Praise*, Routledge, New York, 2020, pp. 27–35.

[12] L. Bardach, K. Murayama, The role of rewards in motivation—beyond dichotomies, *Learn. Instr.* 96 (2025) 102056. https://doi.org/10.1016/j.learninstruc.2024.102056

[13] J. Cameron, K.M. Banko, W.D. Pierce, Pervasive negative effects of rewards on intrinsic motivation: The myth continues, *Behav. Anal.* 24 (2001) 1–44. https://doi.org/10.1007/BF03392017

[14] C.P. Cerasoli, J.M. Nicklin, M.T. Ford, Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis, *Psychol. Bull.* 140 (2014) 980–1008. https://doi.org/10.1037/a0035661

[15] A. Fishbach, K. Woolley, The structure of intrinsic motivation, *Annu. Rev. Organ. Psychol. Organ. Behav.* 9 (2022) 339–363. https://doi.org/10.1146/annurev-orgpsych-012420-091122

[16] S. Hidi, Revisiting the role of rewards in motivation and learning: Implications of neuroscientific research, *Educ. Psychol. Rev.* 28 (2016) 61–93. https://doi.org/10.1007/s10648-015-9307-5

[17] A. Rattan, C. Good, C.S. Dweck, "It's ok—not everyone can be good at math": Instructors with an entity theory comfort (and demotivate) students, *J. Exp. Soc. Psychol.* 48 (2012) 731–737. https://doi.org/10.1016/j.jesp.2011.12.012

[18] R.P. Steel, J.W. Lounsbury, Turnover process models: Review and synthesis of a conceptual literature, *Hum. Resour. Manag. Rev.* 19 (2009) 271–282. https://doi.org/10.1016/j.hrmr.2009.04.002

[19] B.A. Hennessey, T.M. Amabile, Creativity, *Annu. Rev. Psychol.* 61 (2010) 569–598. https://doi.org/10.1146/annurev.psych.093008.100416

[20] B.M. Wiechman, S.T. Gurland, What happens during the free-choice period? Evidence of a polarizing effect of extrinsic rewards on intrinsic motivation, *J. Res. Pers.* 43 (2009) 716–719. https://doi.org/10.1016/j.jrp.2009.03.008

[21] K.E. Marsden, W.J. Ma, E.L. Deci, R.M. Ryan, P.H. Chiu, Diminished neural responses predict enhanced intrinsic motivation and sensitivity to external incentive, *Cogn. Affect. Behav. Neurosci.* 15 (2015) 276–286. https://doi.org/10.3758/s13415-014-0324-5

[22] K.P. Peters, E. Grauerholz-Fisher, T.R. Vollmer, A. Van Arsdale, An evaluation of the overjustification hypothesis: A replication of Deci (1971), *Behav. Anal. Res. Pract.* 22 (2022) 258–264. https://doi.org/10.1037/bar0000245

[23] S.I. Di Domenico, R.M. Ryan, The emerging neuroscience of intrinsic motivation: A new frontier in self-determination research, *Front. Hum. Neurosci.* 11 (2017) 145. https://doi.org/10.3389/fnhum.2017.00145

[24] N. Miura, H.C. Tanabe, A.T. Sasaki, T. Harada, N. Sadato, Neural evidence for the intrinsic value of action as motivation for behavior, *Neuroscience* 352 (2017) 190–203. https://doi.org/10.1016/j.neuroscience.2017.03.064

[25] E.M. Tricomi, M.R. Delgado, J.A. Fiez, Modulation of caudate activity by action contingency, *Neuron* 41 (2004) 281–292. https://doi.org/10.1016/S0896-6273(03)00848-1

[26] K. Murayama, M. Matsumoto, K. Izuma, K. Matsumoto, Neural basis of the undermining effect of monetary reward on intrinsic motivation, *Proc. Natl. Acad. Sci. U. S. A.* 107 (2010) 20911–20916. https://doi.org/10.1073/pnas.1013305107

[27] K. Albrecht, J. Abeler, B. Weber, A. Falk, The brain correlates of the effects of monetary and verbal rewards on intrinsic motivation, *Front. Neurosci.* 8 (2014) 303. https://doi.org/10.3389/fnins.2014.00303

[28] K.C. Berridge, T.E. Robinson, What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience?, *Brain Res. Rev.* 28 (1998) 309–369. https://doi.org/10.1016/S0165-0173(98)00019-8

[29] D. Nguyen, E.E. Naffziger, K.C. Berridge, Positive affect: Nature and brain bases of liking and wanting, *Curr. Opin. Behav. Sci.* 39 (2021) 72–78. https://doi.org/10.1016/j.cobeha.2021.02.013

[30] K.C. Berridge, T.E. Robinson, Liking, wanting, and the incentive-sensitization theory of addiction, *Am. Psychol.* 71 (2016) 670–679. https://doi.org/10.1037/amp0000059

[31] T.E. Robinson, K.C. Berridge, The incentive-sensitization theory of addiction 30 years on, *Annu. Rev. Psychol.* 76 (2025) 29–58. https://doi.org/10.1146/annurev-psych-011624-024031

[32] K.C. Berridge, M.L. Kringelbach, Pleasure systems in the brain, *Neuron* 86 (2015) 646–664. https://doi.org/10.1016/j.neuron.2015.02.018

[33] L.H. Corbit, B.W. Balleine, The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell, *J. Neurosci.* 31 (2011) 11786–11794. https://doi.org/10.1523/JNEUROSCI.2711-11.2011

[34] C. Prévost, M. Liljeholm, J.M. Tyszka, J.P. O'Doherty, Neural correlates of specific and general Pavlovian-to-instrumental transfer within human amygdalar subregions: A high-resolution fMRI study, *J. Neurosci.* 32 (2012) 8383–8390. https://doi.org/10.1523/JNEUROSCI.6237-11.2012

[35] R. Pekrun, The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice, *Educ. Psychol. Rev.* 18 (2006) 315–341.

[36] World Medical Association, World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human participants, *JAMA* 333 (2024) 71–74. https://doi.org/10.1001/jama.2024.21972

[37] F. Faul, E. Erdfelder, A.G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, *Behav. Res. Methods* 39 (2007) 175–191. https://doi.org/10.3758/BF03193146

[38] S. Moeller, E. Yacoub, C.A. Olman, E. Auerbach, J. Strupp, N. Harel, K. Uğurbil, Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI, *Magn. Reson. Med.* 63 (2010) 1144–1153. https://doi.org/10.1002/mrm.22361

[39] A.C. Evans, D.L. Collins, S.R. Mills, E.D. Brown, R.L. Kelly, T.M. Peters, 3D statistical neuroanatomical models from 305 MRI volumes, *Proc IEEE-Nucl Sci Symp Med Imaging Conf.* 3 (1994) 1813–1817.

[40] K.J. Friston, J. Ashburner, C.D. Frith, J.B. Poline, J.D. Heather, R.S.J. Frackowiak, Spatial registration and normalization of images, *Hum. Brain Mapp.* 2 (1995) 165–189.

[41] J. Ashburner, K.J. Friston, Unified segmentation, *NeuroImage* 26 (2005) 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018

[42] K.J. Friston, A.P. Holmes, K.J. Worsley, J.P. Poline, C.D. Frith, R.S.J. Frackowiak, Statistical parametric maps in functional imaging: A general linear approach, *Hum. Brain Mapp.* 2 (1994) 189–210. https://doi.org/10.1002/hbm.460020402

[43] R.N.A. Henson, R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, C.J. Price, W.D. Penny, Analysis of fMRI time series: Linear time-invariant models, event-related fMRI and optimal experimental design, in: R.S.J. Frackowiak et al. (Eds.), *Human Brain Function*, Elsevier, London, 2004, pp. 793–822.

[44] J.A. Mumford, J.-B. Poline, R.A. Poldrack, Orthogonalization of regressors in fMRI models, *PLoS One* 10 (4) (2015) e0126255. https://doi.org/10.1371/journal.pone.0126255

[45] Corbin N, Todd N, Friston KJ, Callaghan MF. Accurate modeling of temporal correlations in rapidly sampled fMRI time series. Hum Brain Mapp. 2018;39:3884–3897. doi:10.1002/hbm.24218.

[46] Olszowy W, Aston J, Rua C, Williams GB. Accurate autocorrelation modeling substantially improves fMRI reliability. Nat Commun. 2019;10:1220. doi:10.1038/s41467-019-09230-w.

[47] W.M. Pauli, A.N. Nili, J.M. Tyszka, A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei, *Sci. Data* 5 (2018) 180063. https://doi.org/10.1038/sdata.2018.63

[48] M. Brett, J.L. Anton, R. Valabregue, J.B. Poline, Region of interest analysis using the MarsBaR toolbox for SPM99 [abstract], *NeuroImage* 16 (2002) S497.

[49] R.A. Poldrack, Region of interest analysis for fMRI, *Soc. Cogn. Affect. Neurosci.* 2 (2007) 67–70. https://doi.org/10.1093/scan/nsm006

[50] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC, New York, 2013.

[51] J.S. Eccles, A. Wigfield, Motivational beliefs, values, and goals, *Annu. Rev. Psychol.* 53 (2002) 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

[52] D. Kahneman, A. Riis, Living, and thinking about it: Two perspectives on life, in: F.A. Huppert, N. Baylis, B. Keverne (Eds.), *The Science of Well-Being*, Oxford University Press, Oxford, 2005, pp. 285–304.

[53] M.D. Robinson, G.L. Clore, Belief and feeling: Evidence for an accessibility model of emotional self-report, *Psychol. Bull.* 128 (2002) 934–960. https://doi.org/10.1037/0033-2909.128.6.934

[54] E.-J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q.F. Gronau, M. Šmíra, S. Epskamp, D. Matzke, J.N. Rouder, R.D. Morey, Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications, *Psychon. Bull. Rev.* 25 (2018) 35–57. https://doi.org/10.3758/s13423-017-1343-3

[55] J.K. Kruschke, *Doing Bayesian Data Analysis*, second ed., Elsevier, Amsterdam, 2015.

[56] H.C. Barron, M.M. Garvert, T.E.J. Behrens, Repetition suppression: A means to index neural representations using BOLD?, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371 (2016) 20150355. https://doi.org/10.1098/rstb.2015.0355

[57] J. Larsson, A.T. Smith, fMRI repetition suppression: Neuronal adaptation or stimulus expectation?, *Cereb. Cortex* 22 (2012) 567–576. https://doi.org/10.1093/cercor/bhr119

[58] N. Bunzeck, E. Düzel, Absolute coding of stimulus novelty in the human substantia nigra/VTA, *Neuron* 51 (2006) 369–379. https://doi.org/10.1016/j.neuron.2006.06.021

[59] B.C. Wittmann, N. Bunzeck, R.J. Dolan, E. Düzel, Anticipation of novelty recruits reward system and hippocampus while promoting recollection, *NeuroImage* 38 (2007) 194–202. https://doi.org/10.1016/j.neuroimage.2007.06.038

[60] P. Ghobadi-Azbari, R. Mahdavifar Khayati, H. Ekhtiari, Habituation or sensitization of brain response to food cues: Temporal dynamic analysis in a functional magnetic resonance imaging study, *Front. Hum. Neurosci.* 17 (2023) 1076711. https://doi.org/10.3389/fnhum.2023.1076711

[61] C.F. Flaherty, *Incentive Relativity*, Cambridge University Press, New York, 1999.

[62] C. Torres, M.R. Papini, Incentive relativity, in: J. Vonk, T.K. Shackelford (Eds.), *Encyclopedia of Animal Cognition and Behavior*, Springer, Cham, 2017, pp. 1–4.

[63] J.M. Harackiewicz, I can't explain, in: R. Arkin (Ed.), *Most Underappreciated*, Oxford University Press, Oxford, 2011, pp. 185–187.

[64] D. Pascucci, C. Hickey, J. Jovicich, M. Turatto, Independent circuits in basal ganglia and cortex for the processing of reward and precision feedback, *NeuroImage* 162 (2017) 56–64. https://doi.org/10.1016/j.neuroimage.2017.08.067

[65] F. Aust, J.M. Haaf, C. Stahl, A memory-based judgment account of expectancy-liking dissociations in evaluative conditioning, *J. Exp. Psychol. Learn. Mem. Cogn.* 45 (2019) 417–439. https://doi.org/10.1037/xlm0000600

[66] J.M. Harackiewicz, G. Manderlink, C. Sansone, Rewarding pinball wizardry: Effects of evaluation and cue value on intrinsic interest, *J. Pers. Soc. Psychol.* 47 (1984) 287–300. https://doi.org/10.1037/0022-3514.47.2.287

[67] N. Schwarz, F. Strack, Reports of subjective well-being: Judgmental processes and their methodological implications, in: D. Kahneman, E. Diener, N. Schwarz (Eds.), *Well-Being: The Foundations of Hedonic Psychology*, Russell Sage Foundation, New York, 1999, pp. 61–84.

[68] F. Rong, M. Kleiman-Weiner, Value internalization: Learning and generalizing from social reward, *arXiv* [cs.LG] (2024). https://doi.org/10.48550/arXiv.2407.14681

[69] P. Anselme, S. Hidi, Acquiring competence from both extrinsic and intrinsic rewards, *Learn. Instr.* 92 (2024) 101939. https://doi.org/10.1016/j.learninstruc.2024.101939

[70] E. Diener, R.A. Emmons, R.J. Larsen, S. Griffin, The satisfaction with life scale, *J. Pers. Assess.* 49 (1985) 71–75. https://doi.org/10.1207/s15327752jpa4901_13

[71] G. Sescousse, X. Caldú, B. Segura, J.-C. Dreher, Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies, *Neurosci. Biobehav. Rev.* 37 (2013) 681–696. https://doi.org/10.1016/j.neubiorev.2013.02.002

[72] O. Bartra, J.T. McGuire, J.W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value, *NeuroImage* 76 (2013) 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

[73] K. Izuma, D.N. Saito, N. Sadato, Processing of social and monetary rewards in the human striatum, *Neuron* 58 (2008) 284–294. https://doi.org/10.1016/j.neuron.2008.03.020

[74] H. Kim, S. Shimojo, J.P. O'Doherty, Overlapping responses for the expectation of juice and money rewards in human ventromedial prefrontal cortex, *Cereb. Cortex* 21 (2011) 769–776. https://doi.org/10.1093/cercor/bhq145

[75] E. Mas-Herrero, L. Maini, G. Sescousse, R.J. Zatorre, Common and distinct neural correlates of music and food-induced pleasure: A coordinate-based meta-analysis of neuroimaging studies, *Neurosci. Biobehav. Rev.* 123 (2021) 61–71. https://doi.org/10.1016/j.neubiorev.2020.12.008

[76] S. Peciña, K.C. Berridge, Opioid site in nucleus accumbens shell mediates eating and hedonic "liking" for food: Map based on microinjection Fos plumes, *Brain Res.* 863 (2000) 71–86. https://doi.org/10.1016/S0006-8993(00)02102-8

[77] V. Putkinen, K. Seppälä, H. Harju, J. Hirvonen, H.K. Karlsson, L. Nummenmaa, Pleasurable music activates cerebral μ-opioid receptors: A combined PET-fMRI study, *Eur. J. Nucl. Med. Mol. Imaging* 52 (2025) 3540–3549. https://doi.org/10.1007/s00259-025-07232-z

[78] B.W. Balleine, J.P. O'Doherty, Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action, *Neuropsychopharmacology* 35 (2010) 48–69. https://doi.org/10.1038/npp.2009.131

[79] M.A. Farries, T.W. Faust, A. Mohebi, J.D. Berke, Selective encoding of reward predictions and prediction errors by globus pallidus subpopulations, *Curr. Biol.* 33 (2023) 4124–4135.e5. https://doi.org/10.1016/j.cub.2023.08.042

[80] S.N. Haber, B. Knutson, The reward circuit: Linking primate anatomy and human imaging, *Neuropsychopharmacology* 35 (2010) 4–26. https://doi.org/10.1038/npp.2009.129

[81] B. Knutson, S.M. Greer, Anticipatory affect: Neural correlates and consequences for choice, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363 (2008) 3771–3786. https://doi.org/10.1098/rstb.2008.0155

[82] W.R. Shadish, T.D. Cook, D.T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston, 2002.

[83] A. Crawley, Clap, clap, clap—unsystematic review essay on clapping and applause, *Integr. Psychol. Behav. Sci.* 57 (2023) 1354–1382. https://doi.org/10.1007/s12124-023-09786-9

[84] R.P. Mann, J. Faria, D.J.T. Sumpter, J. Krause, The dynamics of audience applause, *J. R. Soc. Interface* 10 (2013) 20130466. https://doi.org/10.1098/rsif.2013.0466

[85] T.T.D. Vo, K.V. Tuliao, C.-W. Chen, Work motivation: The roles of individual needs and social conditions, *Behav. Sci.* 12 (2022) 49. https://doi.org/10.3390/bs12020049

[86] J. Heyman, D. Ariely, Effort for payment: A tale of two markets, *Psychol. Sci.* 15 (2004) 787–793. https://doi.org/10.1111/j.0956-7976.2004.00757.x

[87] K.D. Vohs, N.L. Mead, M.R. Goode, The psychological consequences of money, *Science* 314 (2006) 1154–1156. https://doi.org/10.1126/science.1132491

[88] H. Ayabe, T. Koike, A. Yoshioka, S. Izuno, K. Ogasawara, S. Tsuchimoto, N. Sadato, Beyond the undermining effect: Extrinsic rewards preserve neural intrinsic reward, *Behavioural Brain Research* 501 (2026) 115996. https://doi.org/10.1016/j.bbr.2025.115996

Appendix - Supplementary Materials

S1. Ethics Approval and Implementation Procedures

All participants were informed of the study's purpose, procedures, expected burden and benefits, handling of personal data, and their right to withdraw, and provided verbal and written informed consent. Withdrawal rights and data-deletion procedures were detailed in the consent form and reconfirmed during the post-study debriefing. Participant burden was limited to less than two hours. MRI contraindications (e.g., implanted medical devices) were screened via a pre-experiment questionnaire and safety check immediately before scanning. As a risk mitigation strategy, participants wore earplugs during scanning, and continuous communication was maintained via an intercom. In the final debriefing, the first author explained the overall study purpose, emphasizing that the two Experiment phases had been presented with different cover stories as independent experiments, and confirming participants' understanding of this rationale. Participants were informed that, upon withdrawal, all collected data (including prior data) would be permanently deleted; a rapid-response system was in place. All data were anonymized and encrypted and used solely for research purposes.

S2. Sample Size Determination and Exclusion Criteria

An a priori power analysis in G*Power 3.1 for a two-factor repeated-measures ANOVA ($f = 0.30$, $\alpha = 0.05$, power = 0.80) indicated a required sample of $N = 48$ [1]. To accommodate planned attrition, 58 participants were enrolled; four datasets were lost due to a program malfunction, leaving 54 for analysis. Exclusion criteria at enrollment (self-report and safety check) were a history of neurological disorder and MRI contraindications; participants requiring vision correction wore MRI-compatible corrective lenses/goggles, therefore, vision correction was not an exclusion criterion. Participants were randomly allocated to three groups: Group C ($n = 18$; 11 women; $M_{age} = 26.5 \pm 7.1$ years), Group M ($n = 18$; 11 women; $M_{age} = 27.0 \pm 7.5$ years), and Group S ($n = 18$; 10 women; $M_{age} = 26.3 \pm 7.1$ years).

S3. Validity of the SW Task and Success Criteria

In a pilot survey (N = 34; M_{age} = 20.7 ± 11.9 years), participants rated subjective *Liking* of the Stopwatch (SW) task as M = 4.0 ± 0.9 on a 5-point scale. The mean reaction time was M = 5.01 ± 0.08 s; thus, success range was defined as 4.94 – 5.06 s. This difficulty was set to optimize intrinsic reward acquisition, consistent with Atkinson's Achievement Motivation Theory [2], Csikszentmihalyi's Flow Theory [3], and Murayama et al. [4]. The Watch-and-Stop (WS) task served as a control condition. Success counts during the fMRI scan sessions confirmed moderate difficulty (approximately 18–21 successes out of 30 trials per Experiment phase; Table S3).

Table S3.

Mean and SD of SW Task Successes during the fMRI Scan Session in Each Experiment Phase

Experiment phase	1		2		Total	
Group	Mean	SD	Mean	SD	Mean	SD
C	18.94	5.60	19.83	4.89	38.78	9.32
M	21.17	5.26	21.28	4.52	42.44	9.08
S	18.17	5.52	17.78	4.22	35.94	9.42
All	19.43	5.51	19.63	4.70	39.06	9.49

Note. C (control group, n = 18), M (monetary reward group, n = 18), S (social reward group, n = 18), All (all participants, N = 54). The SW task was performed 30 times during the fMRI scan session of the Experiment phase.

S4. Comparison of ROI-level SW–WS β estimates between 4-mm and 8-mm smoothing kernels

This table reports the correspondence between SW–WS contrast β estimates extracted from ROIs after applying either 4-mm or 8-mm FWHM spatial smoothing during preprocessing.

For each Experiment phase (1, 2) and ROI (VTA, NAcc, GPi, GPe, PUT, CN), the table shows:

the Pearson correlation between the 4-mm and 8-mm β values ($r_{(4,8)}$),

the mean β for each kernel (Mean (8), Mean (4)),

the difference in means (ΔMean = Mean (4) – Mean (8)),

the standard deviations for each kernel (SD (8), SD (4)),

the SD difference (ΔSD = SD (4) – SD (8)).

Table S4.

ROI-level SW–WS β Estimates with 4-mm and 8-mm Spatial Smoothing

Experiment phase	ROI	$r_{(4,8)}$	Mean (8)	Mean (4)	ΔMean	SD (8)	SD (4)	ΔSD
1	VTA	0.847	0.124	0.099	−0.025	0.099	0.126	+0.027
	NAcc	0.960	0.154	0.174	+0.020	0.104	0.118	+0.014
	GPi	0.864	0.089	0.077	−0.012	0.083	0.080	−0.003
	GPe	0.967	0.075	0.053	−0.022	0.103	0.111	+0.008
	PUT	0.995	0.053	0.042	−0.011	0.108	0.109	+0.001
	CN	0.997	0.152	0.155	+0.003	0.124	0.132	+0.008
	Ave	**0.977**	**0.108**	**0.100**	**−0.008**	**0.091**	**0.091**	**+0.000**
2	VTA	0.835	0.083	0.059	−0.024	0.077	0.117	+0.040
	NAcc	0.950	0.107	0.119	+0.012	0.091	0.103	+0.012
	GPi	0.801	0.085	0.079	−0.006	0.071	0.089	+0.018
	GPe	0.981	0.082	0.073	−0.009	0.083	0.090	+0.007
	PUT	0.995	0.063	0.057	−0.006	0.087	0.086	−0.001
	CN	0.996	0.118	0.125	+0.007	0.099	0.108	+0.009
	Ave	**0.972**	**0.090**	**0.085**	**−0.005**	**0.072**	**0.072**	**+0.000**

Note. Across all ROIs and both Experiment phases, β estimates derived from 4-mm and 8-mm smoothing were highly correlated (r = 0.80–0.99). Mean and SD differences were minimal, indicating that the choice of smoothing kernel had a negligible influence on ROI-level SW–WS estimates.

S5. Statistical Analysis Settings and Rationale

All statistical analyses were conducted in R (version 4.4.1) using the packages dplyr, tidyr, car, lme4, lmerTest, emmeans, brms, and BayesFactor. Free-choice behavior (*Wanting*) was analyzed following the main-text framework: two-way repeated-measures ANOVA (reward condition (Group C/Group M/Group S) × Experiment phase (1/2)), with Holm–Bonferroni post hoc tests and Bayes factors (BF_{10}; Cauchy $r = .707$). Neural *Liking* contrast estimates from six ROIs (VTA, NAcc, GPi, GPe, PUT, CN) were averaged across hemispheres and analyzed via a mixed-design ANOVA and a Bayesian LMM (4 chains, 4,000 iterations, 2,000 warm-ups; brms) to quantify uncertainty [5]. Model fit was assessed with AIC/BIC from frequentist LMMs. Using difference scores (Experiment phase 2 − Experiment phase 1) as the dependent variable, three random-intercept models were compared (Simple, Intermediate, Full), and the Intermediate model was selected ($\Delta\chi^2_{(5)} = 40.459$, $p < .001$ vs Simple; $\Delta\chi^2_{(10)} = 16.550$, $p = .085$ vs Full) with the lowest AIC (−718.40) and BIC (−680.59).

Simple model: Diff ~ Cond + (1 | PtsID)

Intermediate model: Diff ~ Cond + ROI + (1 | PtsID)

Full model: Diff ~ Cond × ROI + (1 | PtsID)

Cond = reward condition (Group C, Group M, Group S); ROI = brain region; PtsID = participant identifier.

Table S5.

Model Comparison for Difference Scores

Model	npar	AIC	BIC	*log*Lik	Deviance	$\Delta\chi^2$	df	*p*
Simple Model	5	–687.94	–669.03	348.97	–697.94	—	—	—
Intermediate Model	10	–718.40	–680.59	369.20	–738.40	40.459	5	<.001***
Full Model	20	–714.95	–639.33	377.47	–754.95	16.550	10	.085†

Note. npar = number of parameters; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; *log*Lik = *log* likelihood; Deviance = model deviance; $\Delta\chi^2 = \chi^2$ difference from *log* likelihood ratio test; df = degrees of freedom; $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

S6. Model Comparison for Behavioral Wanting (ZIB vs. ZIP vs. Poisson)

To determine the appropriate distributional form for free-choice SW play counts, we compared three generalized linear mixed-effects models—Poisson, zero-inflated Poisson (ZIP), and zero-inflated binomial (ZIB)—using Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO). Table S6 reports the expected log predictive density (elpd_loo) and its standard error (SE) for each model.

Although the ZIP model yielded the numerically highest elpd_loo, differences among models were small relative to their uncertainty. Pairwise differences (Δelpd_loo), computed as the difference in elpd_loo between two models, quantify how much better one model predicts held-out data compared with another. When Δelpd_loo is small relative to its SE (i.e., $|z| = \Delta elpd\_loo / SE < 2$), the difference is considered statistically negligible. In the present comparison, all $|z|$ values were < 2, indicating no decisive predictive advantage of ZIP over ZIB.

Given this absence of a statistically meaningful difference and the conceptual advantages of a zero-inflated binomial structure—reflecting (i) structural zeros arising from participants choosing to rest rather than interact with the computer, and (ii) repeated binary choices (perform SW vs. WS) for those who engaged—the ZIB model was retained as the primary analytic model for *Wanting*.

Table S6.

Model comparison using PSIS-LOO

Model	elpd_loo	SE	Interpretation
ZIB	–185	18.3	—
ZIP	–182	15.6	Highest elpd_loo; not decisively better
Poisson	–191	17.8	Lowest predictive accuracy

Thus, even though ZIP had the highest raw elpd_loo value, Δelpd_loo values indicated that these differences were within sampling uncertainty, supporting the conclusion that ZIB is an adequate and theoretically appropriate model for the *Wanting* data.

S7. Model Comparison for Performance Accuracy

Model Comparison: Assessing the influence of task success rate on neural *Liking* and *Wanting*.

To evaluate whether task performance confounded the neural and behavioral motivational indices, we compared hierarchical models with and without success rate covariates, entered as the standardized success rate **(**zSuc**)** or its change **(**ΔSuc**)**. Across all analyses, including *Liking*, Δ*Liking*, ROI-wise Δ*Liking*, and *Wanting*, adding these covariates did not improve model fit.

1 ‣ Neural *Liking* (ROI-mean SW–WS β)

Model 0 (baseline): $Liking_{mean}$ ~ Cond × Experiment phase (=Sess) + (1|PtsID)

Model 1 (with accuracy): Model 0 + *zSuc* + Sess:*zSuc*

Comparison	npar	elpd_loo	SE	Δelpd	Interpretation
Model 0	8	120.0	7.52	—	—
Model 1	10	118.0	7.48	−2.0	No improvement

➞ Adding success rate did not improve model fit (|Δelpd| < 2).
zSuc and Sess:*zSuc* were non-significant ($p \geq .24$; 95% CrI included zero).

2 ‣ Δ*Liking* (ROI-mean)

Model 0: $\Delta Liking_{mean}$ ~ Cond + (1|PtsID)

Model 1: Model 0 + ΔSuc

Comparison	npar	elpd_loo	SE	Δelpd	Interpretation
Model 0	7	53.2	5.4	—	—
Model 1	8	52.1	5.3	−1.1	No improvement

➞ ΔSuc was negligible (post.mean = 0.002, 95% CrI crosses zero).

3 ▸ ΔLiking (ROI-wise)

Model 0: Δ*Liking* ~ Cond + ROI + (1|PtsID)

Model 1: Model 0 + ΔSuc

Comparison	npar	elpd_loo	SE	Δelpd	Interpretation
Model 0	10	401.0	26.5	—	—
Model 1	11	401.0	26.7	~0	No improvement

➞ ΔSuc effect essentially zero.

4 ▸ *Wanting* (ZIB model)

Model 0: Zero-inflated binomial (ZIB), predictors = Cond × Experiment phase

Model 1: Model 0 + *zSuc*

Comparison	elpd_loo	SE	Δelpd	Interpretation
Model 0	−184	15.7	—	—
Model 1	−185	16.1	−1.0	No improvement

➞ Including *zSuc* worsened fit slightly; accuracy does not explain *Wanting*.

Summary

Across all analyses, success rate — entered as zSuc (standardized success rate) or ΔSuc (change in success rate) — showed no meaningful contribution to neural *Liking*, Δ*Liking*, ROI-wise Δ*Liking*, or *Wanting*. These results indicate that performance accuracy does not confound the motivational findings reported in this study.

Note. As a robustness standard, differences in predictive accuracy were interpreted using the guideline that |Δelpd| < 2 indicates negligible improvement, following established PSIS-LOO conventions [5, 6].

S8. Spearman Correlations among Success Rate, Neural Liking, Subjective Liking, and Wanting across Experiment Phases and Reward Groups

Table S8.

Spearman correlations among Success rate, Neural *Liking*, Subjective *Liking*, and *Wanting*

	1. Success rate	2. Neural *Liking*	3. Subjective *Liking*	4. *Wanting*
1. Success rate	—	.08 −.39 −.27 −.11	.51* .59* .27 .44***	.61** .00 .17 .24†
2. Neural *Liking*	.07 −.39 −.03 −.08	—	−.23 .02 −.42† −.27†	−.31 −.13 −.07 −.23†
3. Subjective *Liking*	.69** −.07 .30 .37**	.03 .16 −.15 .02	—	.52* −.02 .46† .31*
4. *Wanting*	.36 −.31 .18 .07	.01 −.29 .09 −.07	.38 .12 .18 .23†	—

Note. Values are Spearman's *r*. The upper triangle (cells above the main diagonal) shows correlations for Experiment phase 1; the lower triangle (cells below the diagonal) shows correlations for Experiment phase 2. Within each cell, the four rows correspond, from top to bottom, to Group C (control), Group M (monetary reward), Group S (social reward), and all participants (combined). Significance levels: $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

S9. Details of Subjective Evaluation Items

This supplement provides details of 10 questionnaire items across five factors, based on Self-Determination Theory [7] and Expectancy–Value Theory [8]. All items were rated on a five-point scale (1 = Not at all true to 5 = Very true). To minimize response bias, positively worded and reverse-scored items were intermixed. Two subjective *Liking* items were adapted from the Enjoyment/Interest subscale [9] and the Pride/Enjoyment item of the Academic Emotions Questionnaire (AEQ) [10]; evaluating "pleasure during the task" and "joy when successful." Autonomy was assessed with two items measuring self-control and lack of control [7]. Competence items were based on Bandura's self-efficacy theory [11] and Harter's competence framework [12]. Relatedness items drew on Baumeister and Leary's belongingness hypothesis [13] and Ryan and Deci's relatedness construct. Expectancy–Value items followed Eccles and Wigfield's model [8]. Full items are presented in Table S9.

Table S9.

Questionnaire Items and Measurement Concepts

Item	Measurement Concept
I felt stimulated and good.	Subjective *Liking* (Enjoyment)
I felt joy when I succeeded in the task.	Subjective *Liking* (Enjoyment)
I was able to work on the task the way I wanted.	Autonomy
I felt pressure. *	Autonomy
I improved as I proceeded.	Competence
I felt I would fail and couldn't do it well. *	Competence
I felt my efforts were recognized.	Relatedness
I felt left behind by others. *	Relatedness
I learned what I needed to succeed.	Expectancy–Value
Succeeding in the task is useful.	Expectancy–Value

Note. Reverse-scored items are marked with "*."

S10. Descriptive Statistics of Neural Activity

Table S10.

Descriptive Statistics and 95% Credible Intervals of Neural Activity (*Liking*) by Reward Condition and Experiment Phase

Group	C		M		S	
ROI	Experiment phase 1	Experiment phase 2	Experiment phase 1	Experiment phase 2	Experiment phase 1	Experiment phase 2
VTA	0.12 ± 0.12	0.07 ± 0.06	0.15 ± 0.09	0.10 ± 0.10	0.10 ± 0.08	0.07 ± 0.07
	[0.09, 0.15]	[0.05, 0.10]	[0.12, 0.18]	[0.07, 0.13]	[0.09, 0.12]	[0.05, 0.09]
NAcc	0.17 ± 0.12	0.08 ± 0.11	0.16 ± 0.10	0.11 ± 0.10	0.12 ± 0.08	0.10 ± 0.07
	[0.14, 0.19]	[0.05, 0.11]	[0.14, 0.19]	[0.09, 0.14]	[0.10, 0.13]	[0.08, 0.11]
GPi	0.11 ± 0.09	0.09 ± 0.08	0.11 ± 0.08	0.10 ± 0.08	0.04 ± 0.05	0.06 ± 0.05
	[0.08, 0.14]	[0.06, 0.12]	[0.09, 0.14]	[0.07, 0.12]	[0.03, 0.05]	[0.05, 0.08]
GPe	0.11 ± 0.10	0.07 ± 0.08	0.09 ± 0.11	0.10 ± 0.11	0.02 ± 0.08	0.07 ± 0.05
	[0.09, 0.13]	[0.05, 0.09]	[0.06, 0.12]	[0.07, 0.12]	[0.00, 0.05]	[0.06, 0.09]
PUT	0.10 ± 0.11	0.05 ± 0.08	0.05 ± 0.11	0.06 ± 0.12	0.01 ± 0.08	0.06 ± 0.06
	[0.08, 0.13]	[0.03, 0.07]	[0.02, 0.07]	[0.03, 0.09]	[–0.01, 0.04]	[0.04, 0.08]
CN	0.20 ± 0.14	0.12 ± 0.12	0.15 ± 0.12	0.12 ± 0.10	0.09 ± 0.10	0.09 ± 0.07
	[0.18, 0.23]	[0.09, 0.15]	[0.12, 0.18]	[0.09, 0.15]	[0.07, 0.11]	[0.08, 0.11]

Note. Neural activity per trial (contrast estimates). C (control group, *n*=18), M (monetary reward group, *n* =18), S (social reward group, *n* =18), All participants (*N*=54). ROI = region of interest; Experiment phase 1 = reward presentation phase, Experiment phase 2 = reward withdrawal phase; Mean ± SD = mean ± standard deviation; CI = 95% confidence interval (Cousineau–Morey correction); *n* = 18.

S11. Supplementary Results for Subjective Measures

Changes in four components of intrinsic motivation (Autonomy, Competence, Relatedness, Expectancy–Value) and subjective *Liking* were evaluated between Experiment phase 1 and Experiment phase 2 for each group (see Table S11). Autonomy increased overall ($F_{(1, 51)} = 29.57$, $p < .001$, $\eta_p^2 = .37$), with significant increases in Group M ($F_{(1, 17)} = 24.16$, $p < .01$, $\eta_p^2 = .59$) and Group C ($F_{(1, 17)} = 4.99$, $p < .05$, $\eta_p^2 = .22$); no change occurred in Group S ($p = .09$). Competence did not change overall ($F_{(1, 51)} = 1.56$, $p = .22$) but increased post-withdrawal in Group M ($F_{(1, 17)} = 5.05$, $p < .05$, $\eta_p^2 = .23$); Groups C ($p = .63$) and S ($p = .76$) showed no change. Relatedness decreased overall ($F_{(1, 51)} = 9.12$, $p < .01$, $\eta_p^2 = .15$) and markedly in Group S ($F_{(1, 17)} = 14.93$, $p < .01$, $\eta_p^2 = .47$); no change was observed in Group M ($p = .43$) or C ($p = .74$). Expectancy–Value declined overall ($F_{(1, 51)} = 5.48$, $p < .05$, $\eta_p^2 = .10$); neither Group M ($p = .08$) nor Group S ($p = .09$) reached significance. Subjective *Liking* paralleled Relatedness, showing a decline in Group S but stability in the other groups. Full descriptive statistics are shown in Table S11.

Table S11.

Descriptive Statistics of Questionnaire Scales Across Experiment Phases

Experiment phase		1	2	F	η_p^2
Scale	Group	Mean (SD)	Mean (SD)		
Subjective *Liking*	C	3.75 (0.83)	3.72 (0.83)	0.09	0.01
	M	4.22 (0.65)	3.83 (0.84)	7.37*	0.30
	S	4.19 (0.81)	3.39 (0.84)	23.17**	0.58
	All	4.06 (0.78)	3.65 (0.84)	26.00***	0.34
Autonomy	C	3.25 (0.88)	3.64 (0.74)	4.99*	0.22
	M	2.72 (0.79)	3.81 (0.73)	24.16**	0.59
	S	3.11 (0.80)	3.42 (0.70)	2.99	0.15
	All	2.99 (0.84)	3.62 (0.69)	29.57***	0.37
Competence	C	3.22 (1.06)	3.33 (0.88)	0.23	0.01
	M	3.25 (0.90)	3.64 (0.88)	5.05*	0.23
	S	3.61 (0.63)	3.56 (0.88)	0.10	0.01
	All	3.37 (0.88)	3.51 (0.88)	1.56	0.03
Relatedness	C	3.92 (0.62)	3.92 (0.74)	0.00	0.00
	M	3.86 (1.05)	3.81 (0.74)	0.62	0.04
	S	4.03 (0.58)	3.39 (0.80)	14.93**	0.47
	All	3.96 (0.78)	3.70 (0.79)	9.12**	0.15
Expectancy–Value	C	3.42 (0.84)	3.36 (0.95)	0.21	0.01
	M	3.61 (0.98)	3.33 (0.96)	3.10†	0.15
	S	3.39 (0.90)	3.19 (0.98)	3.24†	0.16
	All	3.49 (0.90)	3.30 (0.98)	5.48*	0.10

Note. C = control group (n = 18), M = monetary reward group (n = 18), S = social reward group (n = 18), All = all participants (N = 54). F-ratios for Groups C, M, and S have degrees of freedom (1, 17); those for All have degrees of freedom (1, 51). †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$; η_p^2 = partial eta-squared.

S12. Exploratory simple effects underlying the asterisks in Figure 2

This supplementary note provides a descriptive explanation of the asterisks shown in Figure 2. The asterisks indicate exploratory within-condition tests of the Experiment phase effect (Experiment phase 2 vs. Experiment phase 1) conducted separately for each ROI within each reward condition. These tests were performed without correction for multiple comparisons, and the reported p values are therefore unadjusted. Because these ROI-wise simple effects involve multiple comparisons across ROIs and reward conditions, the results are presented for reference purposes only and should be interpreted with caution. The primary inferences of the study are based on the main analyses reported in the main text and other Supplementary Materials, including difference-score analyses and planned contrasts.

In these exploratory analyses, Group C showed significant decreases in neural responses from Experiment phase 1 to phase 2 in NAcc ($p < .01$) and CN ($p < .05$). In contrast, Group S showed significant increases in GPe ($p < .05$) and PUT ($p < .05$). No other ROI-by-condition phase differences reached these significance thresholds.

S13. Difference Scores by ROI and Condition

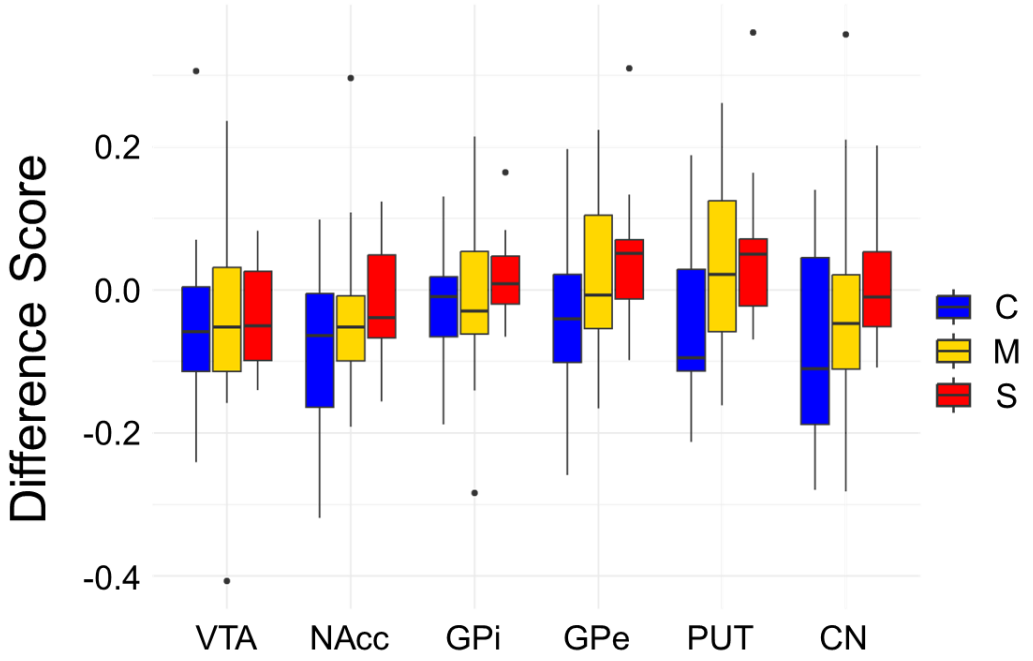

Figure S13. Box plot of difference scores between Experiment phases for each ROI. Difference scores per trial (change in *Liking*: Experiment phase 2 − Experiment phase 1) are displayed for each ROI. Groups: Group C = control group (n = 18), Group M = monetary reward group (n = 18), Group S = social reward group (n = 18). Positive difference scores indicate an increase in *Liking*, and negative difference scores indicate a decrease. The thick line in each box represents the median, and the upper and lower ends of the box represent the first quartile (Q_1) and third quartile (Q_3), respectively. The height of the box indicates the interquartile range ($IQR = Q_3 - Q_1$), and the whiskers connect the minimum and maximum values within the range from $Q_1 - 1.5 \times IQR$ to $Q_3 + 1.5 \times IQR$. Values outside the whisker range are displayed as individual points (outliers).

S14. Estimated Difference Scores by ROI and Condition (Bayesian LMM Analysis)

Table S14.

Estimated Difference Scores and 95% Credible Intervals (CrIs) from Bayesian LMM by ROI and Reward Condition

ROI	Group	Difference Score	95% CrI
VTA	C	–0.08	[–0.12, –0.03]
	M	–0.03	[–0.08, 0.01]
	S	–0.01	[–0.06, 0.03]
NAcc	C	–0.08	[–0.13, –0.04]
	M	–0.04	[–0.09, 0.01]
	S	–0.02	[–0.06, 0.03]
GPi	C	–0.04	[–0.09, 0.01]
	M	0.00	[–0.04, 0.05]
	S	0.03	[–0.02, 0.07]
GPe	C	–0.03	[–0.08, 0.02]
	M	0.01	[–0.03, 0.06]
	S	0.04	[–0.01, 0.09]
PUT	C	–0.03	[–0.07, 0.02]
	M	0.02	[–0.03, 0.06]
	S	0.04	[–0.01, 0.08]
CN	C	–0.07	[–0.12, –0.02]
	M	–0.03	[–0.07, 0.02]
	S	0.00	[–0.05, 0.04]

Note. Difference scores are estimated marginal means of (Experiment phase 2 − Experiment phase 1). 95% CrI = 95% credible interval. C = control group (n = 18), M = monetary reward group (n = 18), S = social reward group (n = 18).

S15. Correlation between Liking Change and WS play Count Change

Spearman's rank correlation coefficients (r) were calculated for changes in *Liking* and WS task play counts from Experiment phase 1 to Experiment phase 2. All correlations were non-significant ($|r| < .29$, $p > .10$), indicating no association in any group.

Table S15.

Spearman's *r* between *Liking* Change and WS Task Plays Change

ROI	VTA	NAcc	GPi	GPe	PUT	CN
Group C	–.13	–.27	–.27	–.04	–.06	–.29
Group M	.10	.08	.01	–.12	–.16	.12
Group S	–.21	–.18	–.21	–.08	–.17	–.22
All	–.05	–.06	–.07	.04	.02	–.05

Note. Correlations are Spearman's *r* between WS play count change and BOLD response difference scores (*Liking* change) for each ROI: VTA, NAcc, GPi, GPe, PUT, CN. Group C = control group (n = 18), Group M = monetary reward group (n = 18), Group S = social reward group (n = 18), All = all participants (N = 54). All correlations were non-significant ($p > .10$).

S16. Correlation Analysis of Behavioral and Subjective Changes

Spearman's rank correlation coefficients (r) were computed between changes in task engagement (SW and WS task play counts) and changes in subjective measures (subjective *Liking* and related factors) from Experiment phase 1 to Experiment phase 2 (see Table S15). WS task plays correlated negatively with subjective *Liking* overall ($r = –.32$, $p < .05$) and in Group C ($r = –.58$, $p < .01$). Competence correlated positively with SW plays overall ($r = .25$, $p < .10$) and in Group C ($r = .50$, $p < .05$). Relatedness exhibited the highest correlation with subjective *Liking* overall ($r = .50$, $p < .001$) and in Group S ($r = .49$, $p < .01$). Expectancy–Value correlated with subjective *Liking* overall ($r = .47$, $p < .001$) and strongly in Group M ($r = .74$, $p < .001$) and Group S ($r = .62$, $p < .01$). These results complement the main-text finding that *Wanting*–subjective *Liking* correlations were small and non-significant across conditions ($|r| < .31$, $p > .10$).

Table S16.

Spearman's *r* for Changes in Task Plays and Questionnaire Scale Scores

Variables	Mean	SD	1	2	3	4	5	6	7
1. *Wanting* (SW play count)	-1.13	3.19	-	-.03 .12 -.39	.20 -.12 .21	.14 -.13 -.30	.50* -.20 .24	.20 .41** .21	.38 -.01 .29
2. WS play count	-0.19	1.03	-.12	-	-.58** -.21 .07	-.16 .29 -.22	-.24 .41** -.13	.06 .26 .01	-.22 -.13 .20
3. Subjective *Liking*	-0.41	0.66	.12	-.32*	-	.09 -.46† -.31	.16 -.03 .24	.00 .39 .49*	-.15 .74*** .62**
4. Autonomy	0.63	0.93	-.12	-.04	-.20	-	.56* .50* -.02	.12 .11 -.22	.45† -.24 -.20
5. Competence	0.14	0.82	.25†	-.10	.15	.40**	-	.30 .13 .37	.71** -.01 .57*
6. Relatedness	-0.26	0.68	.22	-.04	.50***	.07	.27*	-	.59** .51* .62**
7. Expectancy-Value	-0.19	0.61	.21	-.10	.47***	-.12	.33*	.49***	-

Note. Spearman's rank correlation coefficients. Variables represent changes from Experiment phase 1 to Experiment phase 2. SW play count corresponds to *Wanting*. The lower triangle shows coefficients for all participants ($N = 54$). The upper triangle shows, for each reward condition group, three coefficients per cell arranged vertically: top row = Control group ($n = 18$), middle = Monetary reward group ($n = 18$), bottom = Social reward group ($n = 18$). $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

S17. Participant Factors and Neural Change Correlations

Table S17.

Correlation between Various Measures and Neural Activity Changes

Variable	VTA	NAcc	GPi	GPe	PUT	CN	Ave
Age	–.24	–.17	.02	–.08	–.07	.06	–.10
	.13	.12	.34	.37	.20	.04	.22
	.07	.08	–.07	–.14	–.20	–.01	–.07
	.00	.02	.17	.10	.00	.04	.06
Success rate	.16	.46†	.36	.36	.34	.30	.39
	.05	.08	–.01	–.11	–.17	.12	.00
	–.19	–.10	.03	.15	.19	–.02	.03
	.02	.13	.05	.08	.08	.12	.09
Wanting (SW task play count)	–.31	–.17	–.20	–.18	–.31	–.21	–.21
	.02	.16	.12	.07	.02	.06	.06
	–.21	.07	–.01	.00	–.16	–.10	–.10
	–.18	.02	–.06	–.06	–.16	–.16	–.10
WS task play count	–.13	–.27	–.27	–.04	–.06	–.29	–.21
	.10	.01	.01	–.12	–.16	.12	.01
	–.21	–.21	–.21	–.08	–.17	–.22	–.21
	–.13	–.06	–.07	.04	.02	–.05	–.03
Subjective *Liking*	.26	.36	.24	.11	.07	.26†	.26
	.34	.11	.20	.05	–.13	–.28	.05
	–.70**	–.70**	–.03	.07	–.01	–.44†	–.34
	.00	–.21	.02	–.10	–.20	–.27	–.16
Autonomy	.08	–.13	–.06	–.16	–.10	–.04	–.07
	–.11	–.14	–.10	–.09	–.06	.02	–.09
	.55*	.48*	.22	.15	.04	.40	.35
	.05	.01	–.07	–.05	–.02	.07	.00
Competence	.38	–.04	.23	.14	.17	.02	.17
	–.10	–.14	–.13	–.23	–.32	–.24	–.22
	.07	.02	.52*	.57*	.57*	.32	.43
	.13	–.07	.09	.10	.11	–.03	.07
Relatedness	.34	.01	–.10	–.11	–.18	–.20	–.05
	.47*	.31	.39	.09	–.10	–.03	.21
	–.45	–.53*	–.36	–.27	–.30	–.52*	–.47*
	.14	–.12	.00	–.18	–.29*	–.26†	–.15
Expectancy–Value	.30	–.23	–.06	–.19	–.18	–.25	–.12
	.66**	.42*	.65**	.41†	.18	.10	.45†
	–.54*	–.52*	.03	.09	.04	–.29	–.22
	.35*	.00	.36**	.12	–.01	–.10	.12

Note. Spearman's rank correlation coefficients. Variables: Change scores from Experiment phase 1 to Experiment phase 2. In each cell, rows (top to bottom) are Group C, Group M, Group S, and all groups. Ave = average across the six ROIs. $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Supplementary References

[1] F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, *Behav. Res. Methods*, 39 (2007) 175–191. https://doi.org/10.3758/bf03193146

[2] J. W. Atkinson, Motivational determinants of risk-taking behavior, *Psychol. Rev*. 64 (1957) 359–372. https://doi.org/10.1037/h0043445

[3] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. HarperPerennial, 1990/2008

[4] K. Murayama, M. Matsumoto, K. Izuma, K Matsumoto., Neural basis of the undermining effect of monetary reward on intrinsic motivation, *Proc. Natl. Acad. Sci. U. S. A.* 107 (2010) 20911–20916. https://doi.org/10.1073/pnas.1013305107

[5] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian Data Analysis,* Chapman and Hall/CRC, 2013. https://doi.org/10.1201/b16018

[6] A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Stat. Comp.* 27 (2017) 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

[7] R. M. Ryan, E. L. Deci, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being, *Am. Psychol*. 55 (2000) 68–78. https://doi.org/10.1037/0003-066x.55.1.68

[8] J. S. Eccles, A. Wigfield, Motivational beliefs, values, and goals, *Annu. Rev. Psychol*. 53 (2002) 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

[9] E. McAuley, T. Duncan, V. V. Tammen, Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis, *Res. Q. Exerc. Sport*. 60 (1989) 48–58. https://doi.org/10.1080/02701367.1989.10607413

[10] R. Pekrun, T. Goetz, W. Titz, R. P. Perry, Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research, *Educ. Psychol*. 37 (2002) 91–105. https://doi.org/10.1207/s15326985ep3702_4

[11] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice-Hall, 1986

[12] S. Harter, The perceived competence scale for children, *Child Dev*. 53 (1982) 87. https://doi.org/10.2307/1129640

[13] R. F. Baumeister, M. R. Leary, The need to belong: Desire for interpersonal attachments as a fundamental human motivation, *Psychol. Bull*. 117 (1995) 497–529. https://doi.org/10.1037/0033-2909.117.3.497