Neural Encoding of Immediate and Instrumental Value During Planning

Daniil Luzyanin[a,b], Arkady Konovalov[a*]

[a] Centre for Human Brain Health, School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2UJ, United Kingdom

[b] Department of Pharmacology, School of Pharmacy, University of Reading, Whiteknights Campus, RG6 6UB, United Kingdom

[*] Corresponding author (a.konovalov@bham.ac.uk)

ABSTRACT

Planning is a key executive function that enables humans to anticipate future outcomes by mentally simulating action sequences and balancing immediate gains against long-term goals. While we developed a partial understanding of neural mechanisms of planning in spatial navigation and reward-based reinforcement learning, little is known about the relative neural encoding of the instantaneous and instrumental value of the same action. We used a novel fMRI task in which participants repeatedly chose between two options, each with an immediate monetary value and a known future-oriented instrumental value. Our results show that striatal activity is positively correlated with instrumental value, whereas activity in the dorsomedial prefrontal cortex and bilateral insula is negatively correlated with the instantaneous value. These dissociable patterns support specialization in valuation neural circuits during planning.

Keywords: planning, fMRI, value

INTRODUCTION

Planning is a key component of intelligent behavior, which allows us to anticipate future outcomes and choose actions that maximize long-term benefits[1–7]. When navigating a new city or strategizing during a chess game, we must mentally simulate sequences of actions and their consequences. In a way, planning enables "mental time travel", which provides evolutionary advantages by helping individuals prepare for or even influence upcoming events[8,9]. Neuroscience has long recognized planning as a higher-order executive function. Understanding how the brain mediates planning is not only crucial for models of decision-making but can also have clinical relevance. Classic neuropsychological work has shown that patients with prefrontal cortex damage struggle with multi-step problem-solving (for instance, the Tower of Hanoi task)[10]. Deficits in planning often contribute to impulsivity and maladaptive choices in disorders ranging from frontal dementia to substance abuse[11,12].

One important open question is how the brain encodes the value of actions during planning. Often, people must assess both the *instantaneous value* of a choice (its immediate, intrinsic reward or cost) and its *future value*, which is the benefit the choice provides as a step toward a larger goal; we will call this value *instrumental*. For example, taking a detour on a road trip may have a low immediate payoff (such as increased cognitive effort and longer travel time), but it can have high instrumental value if it helps avoid traffic and save time later.

Converging evidence shows that the brain encodes value across distributed networks, with different regions possibly representing immediate or future outcomes. Much previous research on decision-making has concentrated on how the brain values immediate rewards in simple, one-step choices using time-discounting paradigms that do not involve explicit planning but reflect an individual's intertemporal preferences or

patience[13,14]. These studies have identified a valuation network, including the ventromedial prefrontal cortex (vmPFC) and ventral striatum (particularly the nucleus accumbens), which encodes the subjective value of options at decision time[13,15–19]. Activity in the vmPFC and ventral striatum correlates with expected reward magnitude and probability, suggesting these regions may provide a common "currency" signal for comparing options[17]. This network is also active when choices involve trade-offs between immediate and delayed rewards. For example, the vmPFC and ventral striatum track the subjective value of future rewards discounted by delay, regardless of whether the reward is available now or later[13]. Recent research indicates that damage to the vmPFC can impair planning depth[20].

At the same time, there is evidence for functional specialization in how the brain processes immediate versus future rewards. Early neuroimaging studies of intertemporal choice reported a dissociation between choices of delayed and immediate rewards, showing stronger activation in the limbic system for immediate rewards, with PFC activity linked to delayed rewards. This evidence supports the idea of dual valuation systems: a fast, impulsive system that favors immediate gratification and a slower deliberative system for future outcomes[14,22].

Another region potentially implicated in future-oriented decisions is the dorsal anterior cingulate cortex (dACC), which is often involved in cognitive control and the selection of reward-guided actions[23,24]. The dACC has been suggested to compute the expected value of control, estimating whether the potential future payoff of a difficult or costly action justifies the effort[23]. This integrative role indicates that the dACC may be more active when the choice between immediate and future value becomes more difficult, a trade-off often present during planning. For example, single neurons in the

rodent and primate ACC encode cost–benefit variables and future reward expectations during decision-making[24,25].

In humans, dACC BOLD activity has been linked to foraging decisions that require comparing the value of staying versus exploring for potentially better future rewards[26]. The dACC and dorsomedial PFC (dmPFC) typically show increased activation during tasks involving difficult, future-oriented choices. This could reflect the mental effort or conflict involved in pursuing long-term goals[14,27].

Computational work has formalized the extent and direction of human planning. For example, Keramati and colleagues[28,29] and Sezener et al.[30] proposed that planners trade off the computational cost of simulating additional steps against the value gained by that foresight. These accounts suggest that planning depth increases when the instrumental value of doing so is high, essentially, when the future rewards justify the effort.

Model-based reinforcement learning studies have identified specific regions involved in planning (the caudate, lateral prefrontal cortex, and hippocampus) by correlating brain activity with computational model predictions[31–42]. These studies linked striatal activity to model-free (learned through reward) and model-based (learned through the structure of the planning tree) prediction errors[31,43–45] and OFC activity to a cognitive map of the task[46–49]. However, these paradigms often conflate planning with learning as participants must gradually learn (or even over-learn) state transitions or reward contingencies. As an example, the widely used two-step task[31], where a decision maker makes a first-stage choice between two options that can lead to four second-stage outcomes, dissociates model-based and model-free influences on behavior, but the first-stage choice does not have intrinsic value and is thus not separated from subsequent reward outcomes[41].

Some of this work dissociated the neural systems involved in forward-looking planning as opposed to simple reinforcement learning. Using a task with several planning steps,

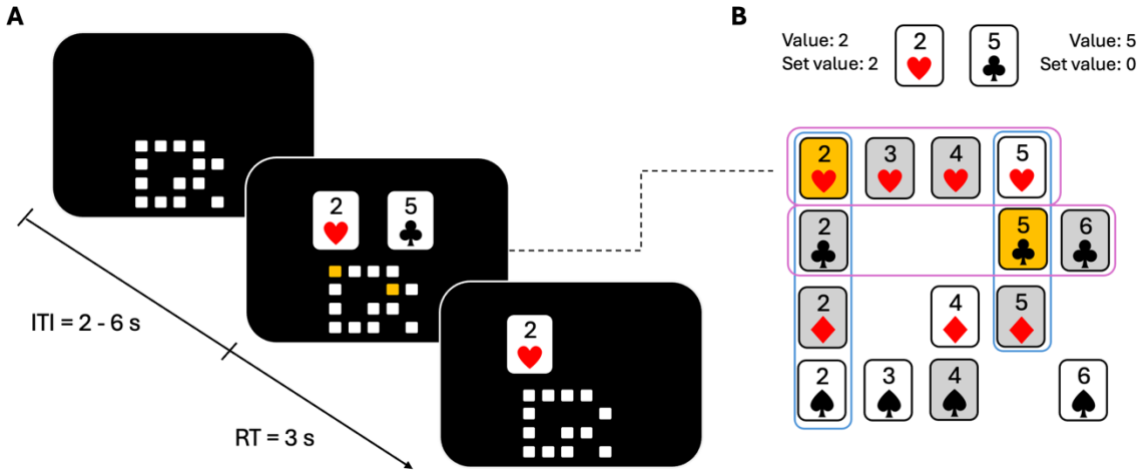

Figure 1. (**A**) Experimental task. In each game, the participant started with a deck of 20 cards, each representing four suits and five values (2 to 6), displayed in a grid at the lower part of the screen. All cards were grayed out at the start of the game. In each round, after a jittered inter-trial interval (ITI) of 2 to 6 seconds, two cards were randomly drawn from the deck, highlighted in the grid, and shown in full size above the grid. The participant chose one of the cards (left or right) using a button box. The response time was restricted to 3 seconds. The chosen card changed its color to white, and the discarded card was removed from the grid. The game proceeded for 10 rounds, until all 20 cards had either been chosen or discarded. The display then showed the total points earned in this game. (**B**) The foundations of the model. The model assumes that the value of each card includes a weighted sum of its digit value (2 to 6) and its value as a part of a set. In the shown example, no more sets can be possibly formed using the 5 of clubs, and two sets could be formed with the 2 of hearts card. Thus, the 2 of hearts has a lower instantaneous value, but a higher instrumental value.

Wunderlich et al. provided evidence for distinct neural systems in forward planning versus choices resulting from overtrained habitual learning[35]. During planning, the anterior caudate encoded step-by-step future values, whereas the putamen tracked values learned through extensive training. Lee et al. identified a prefrontal arbitration mechanism that determines when to engage model-based planning[50], showing that the inferior lateral PFC and frontopolar cortex monitor the reliability of model-based vs. model-free predictions and allocate control accordingly.

In summary, our understanding of neural value encoding during multi-step planning has been steadily improving but remains incomplete. First, as discussed above, most human studies on prospective decision-making have focused either on purely immediate-reward choices (single-step preference tasks) or on paradigms in which future value is learned through trial-and-error (reinforcement-learning tasks). These paradigms usually require participants to learn a decision tree, whereas in daily life, the decision tree is often

too complex to compute. Second, few experiments have specifically isolated neural signals of instrumental value for planning at the moment of choice, independent of confounding factors such as reward prediction errors or ongoing learning. Therefore, it is still unclear how the brain encodes immediate versus future value in decision scenarios where the individual already knows the task structure or when the structure is too complex. Do different neuronal populations or brain regions represent the immediate reward of an action versus its future value, even when no new learning is needed?

Here, we present an fMRI investigation of planning behavior that was designed to address this gap. We developed a novel card-collection task that dissociates immediate and future values within each decision, without requiring incremental learning. In this task, participants choose between two cards drawn from a deck; each card carries an *instantaneous value* (a point reward received immediately) and a potential *instrumental value* in helping complete a "set" that yields a bonus at the end of the game.

The task is structured such that subjects are fully informed of the rules and payoff structure in advance. There remains stochasticity in the task, as cards are drawn at random, but the space of possible outcomes is so vast that tracking the decision tree is not computationally tractable. However, participants still can engage in planning using simple counting models, weighing whether a given card's short-term points are worth more or less than its contribution to a future bonus, without trial-and-error learning about the card values. Importantly, the design orthogonalizes the two value components across trials, allowing us to use separate regressors for instantaneous and instrumental value in neuroimaging analyses.

RESULTS

Both instantaneous and instrumental values contribute to planning decisions

We designed a task to distinguish between the instantaneous and instrumental values of options (**Figure 1A**; see **Methods**). In this task, participants played a game using a deck of 20 cards, which included four standard suits and five distinct digit values (from 2 to 6). During each trial, the computer drew two cards from the deck without replacement. Participants had to choose one card and discard the other. They immediately received the point value of the selected card, which we will refer to as "instantaneous value." Depending on the luck of the draw, a participant could earn up to 44 points in a single game this way (e.g., four cards with a value of 5 and four cards with a value of 6).

They made ten decisions until the deck was exhausted. Then, they received an extra 20 points for each complete set of *at least four cards* sharing the same suit or point value (the set bonus was determined through pilot tests). Given the deck size, they could form up to three sets (for example, two suit sets and one set of digits), and the total possible set bonus could reach 60 points. Therefore, each card could have a "set value" (or, more generally, instrumental value) if it could be part of a set. The participant often faced a trade-off between the instantaneous and instrumental value (**Figure 1B**), with the latter being crucial for successful planning and earning the set bonuses. The game also featured long-term bonuses throughout the session (**Methods**), but since these did not influence participants' behavior, we did not include them in the analyses.

Always going for a higher-value card is a suboptimal strategy in the task, and participants recognized that. Participants did not select the card with the highest point value in approximately 30% of trials, with considerable variability across individuals (ranging from 16% to 54%). The participants also differed in their task performance, earning between 1,789 and 3,384 points across the session (mean = 2,734). Participants

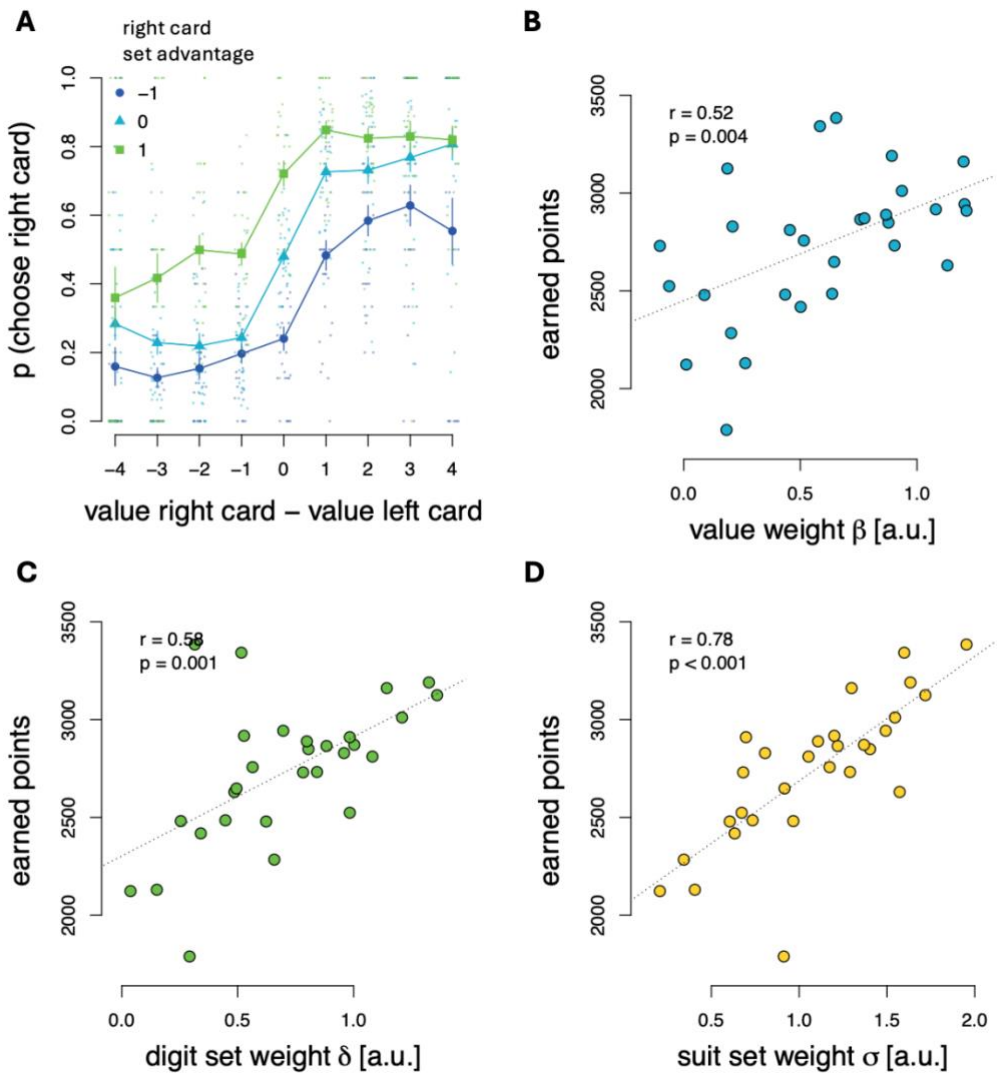

Figure 2. Behavioral results. (**A**) The probability of choosing the right card as a function of the difference between the digit value of the right card minus the left card, split by the right card set advantage (the difference between potential complete sets between the right and left cards). Differences of 2 are not shown due to the small amount of data in most bins. The individual subjects are shown as dots. The error bars denote s.e.m. (**B**) Individual earned points as a function of the participant's digit value model weight β. (**C**) Individual earned points as a function of the participant's suit set model weight σ. (**D**) Individual earned points as a function of the participant's digit set model weight δ. In panels B-D, each dot represents an individual participant, the line indicates the regression fit, and r and p-value denote the Pearson correlation. For all panels: N = 29.

who chose lower-value cards more often earned fewer points (r(27) = -0.48, CI = [-0.72, -0.13], p = 0.01), so overall choosing higher values was beneficial despite the benefits of set collection.

We found that participants' choices relied on both the point values of the cards and their set values (**Figure 2A**). To explain these choices, we fit a computational model that used a weighted sum of the differences in the point values of the two cards, as well as

their values as part of the suit and digit sets (see **Methods** for details). In short, each card's *set value* was initially set to 0 and increased by 1 (or 2) if there was still the potential for a complete set (or two sets) of that suit or digit to be collected in the current game.

The model provided a good fit to the data, correctly explaining (within-sample) 75±10% of participants' choices. The model estimated individual-level free parameters that determined the weights (β for the point value, σ for the suit set value, and δ for the digit set value). The suit set weight σ showed a higher contribution to overall performance, as it was easier to collect four cards of the same suit (since there were five of them) than four cards of the same value (since there were only four of them). We also estimated five competing models that showed worse fit to the data (**Methods**, **Supplementary Table 1**).

All three individual-level parameters were strongly linked to the participant's total earnings, demonstrating that each component contributed to task performance (**Figure 2B-D**). We also ran simulations of the model to see if higher weights are optimal for this task (**Supplementary Figure 1**). We found that the ideal point weight peaks at around 1, suggesting that relying too heavily on the instantaneous point value harms task performance. The simulations also revealed that participants, on average, used suboptimal set weights.

Higher value choices are linked to higher striatal and lower dorsomedial and insular activity

We then turned to neuroimaging to examine how choice values are encoded in the brain. First, we investigated model-free activity associated with choosing lower-value cards in general. Specifically, we ran a GLM predicting BOLD activity using a dummy

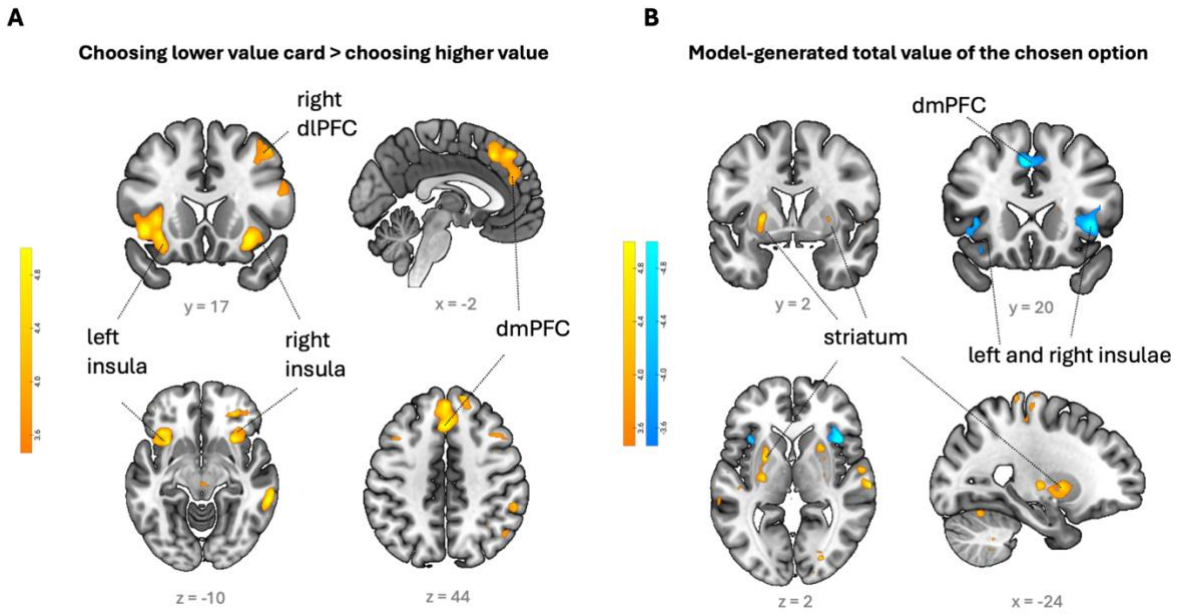

Figure 3. (**A**) Clusters of BOLD activity showing increased activity in the right dlPFC, dmPFC, and bilateral insula for trials where participants chose a card with a lower or equal digit compared to the trials where they chose a card with a higher digit value. See Supplementary Table 2 for the full table of activations. (**B**) Clusters of BOLD activity (positive in the striatum, negative in the bilateral insula and dmPFC) correlated with the model-generated value of the card chosen in a trial. See Supplementary Table 3 for the full table of activations. All images shown with threshold at $p < 0.05$, FWE-corrected ($t = 3.467$). For all panels: N = 25.

variable equal to 1 if the participant chose a card with a point value equal to or lower than that of the other card (GLM1, **Methods**).

We identified a network showing increased activity during these trials, including the right dlPFC (MNI: x = 51, y = 14, z = 30), dmPFC (MNI: x = 0, y = 23, z = 44), and bilateral insula (MNI: left x = -42, y = 17, z = -7, right x = 30, y = 17, z = -10) (**Figure 3A**, **Supplementary Table 2**). Interestingly, we did not detect any above-threshold activity with a negative sign (regions displaying higher activity when a higher point-value option was chosen), even though the literature suggests that higher received value usually correlates with increased activity in the vmPFC and striatum.

We then moved on to model-based estimation and computed the value of the selected option using our computational model (**Methods**) as a parametric modulator of BOLD activity on each trial (GLM2, **Methods**). As anticipated, we observed a positive association between this regressor and striatal activity, mainly in the putamen (MNI: x =

-24, y = 2, z = 2), along with clusters in the precuneus (MNI: x = -12, y = -70, z = 31), fusiform gyrus (MNI: x = 30, y = -64, z = -16), and right temporal pole (MNI: x = 63, y = -10, z = -1) (**Figure 3B**, **Supplementary Table 3**). Again, we found a negative correlation with bilateral insula (MNI: left x = -36, y = 14, z = -1; right x = 39, y = 17, z = 2) and the dmPFC (MNI: x = -3, y = 20, z = 41) (**Figure 3B**, **Supplementary Table 3**). These findings imply that the increased activity in the dmPFC and insula seen in GLM1 may be related to a decreased dependence on instantaneous value.

Dissociation between instantaneous and instrumental value

Although the total value encoded in the striatum was not surprising, our main goal was to determine whether this signal was mainly influenced by the instantaneous or the instrumental component of the chosen option's value. To do that, we ran two GLM models (GLM 3 and 4, **Methods**) that included either the point value of the chosen option or its set value (combining the digit and suit values for simplicity, so this variable could take values of 0, 1, and 2 depending on the possible number of sets that could still be made with that specific card).

While the instantaneous value regressor showed a moderate correlation (approximately r = 0.5) with the dummy variable used in GLM1 (indicating selection of an option with a lower point value), it was not identical to it. The neural correlates, however, were qualitatively similar (**Figure 4A**). Again, contrary to our preregistration predictions, we found no significant clusters of positive activity in the standard value-encoding regions (including anatomical ROIs of the precuneus and caudate, **Figure 4C**), and only one significant positive cluster in the visual cortex (MNI: x = -6, y = -85, z = -7). Similarly to the total value of the chosen option, we observed negative correlations in the

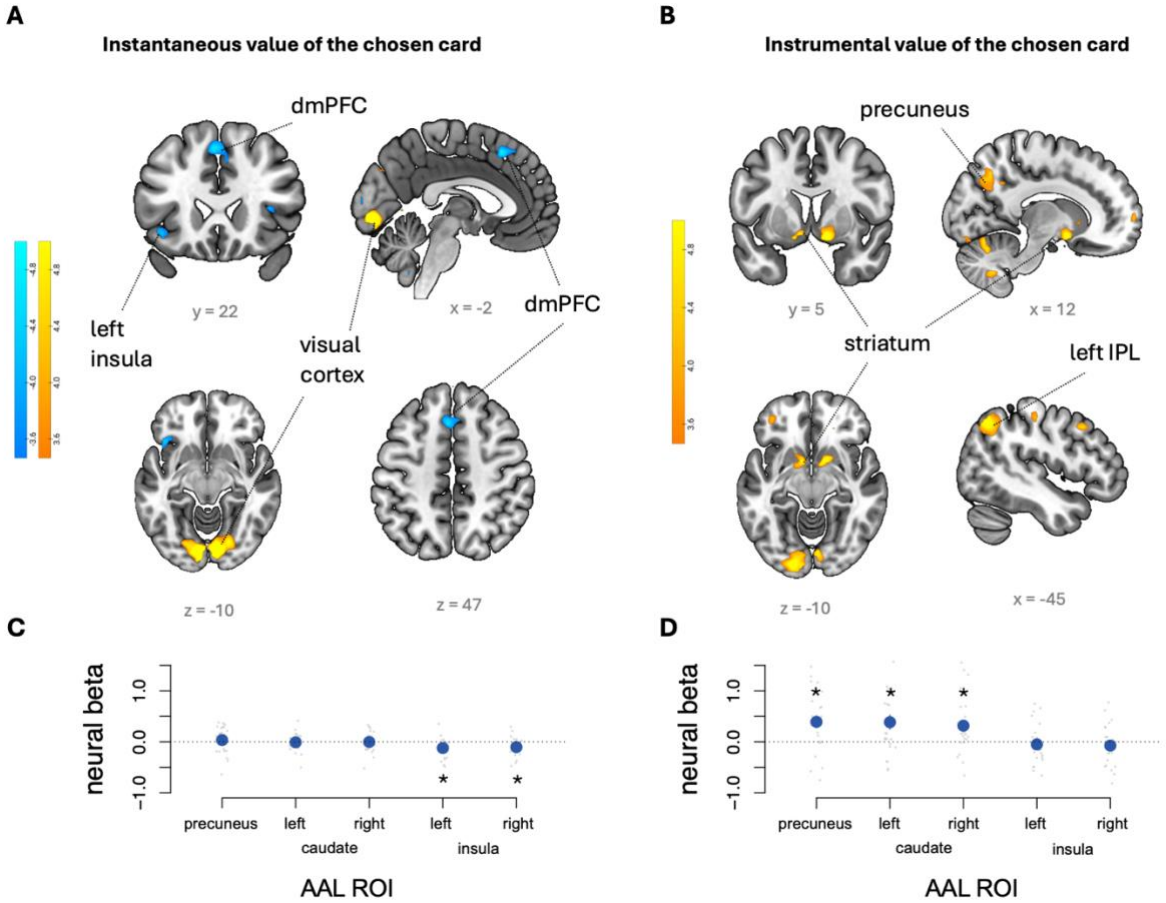

Figure 4. (**A**) Clusters of BOLD activity correlated negatively (in the right dlPFC, dmPFC, and bilateral insula) and positively (the visual cortex) with the digit value of a chosen card (instantaneous value). See Supplementary Table 4 for the full table of activations. (**B**) Clusters of BOLD activity (positive in the striatum, precuneus, and the left inferior parietal lobule (IPL)) correlated with the set value of the card chosen in a trial (instrumental value). See Supplementary Table 5 for the full table of activations. All images shown with threshold at $p < 0.05$, FWE-corrected ($t = 3.467$). (**C**) The average neural betas (GLM coefficients) for the instantaneous value of the chosen card in the specific regions of interest (ROI), defined anatomically using the AAL (Automated Anatomical Labeling atlas). (**D**) The average neural betas (GLM coefficients) for the instrumental value of the chosen card in the specific regions of interest (ROI), defined anatomically using the AAL (Automated Anatomical Labeling atlas). For panels C and D, gray dots denote individual subjects, error bars denote s.e.m., and stars label average activations in each ROI significantly different from 0 across subjects (Bonferroni-corrected). See Supplementary Figure 2 for the ROI images. For all panels: N = 25.

bilateral insula (MNI: left x = -30, y = 29, z = -4; right x = 36, y = 17, z = 8) and the dmPFC (MNI: x = -3, y = 20, z = 41) (**Supplementary Table 4**).We found a positive striatal signal linked to the instrumental value, mainly in the left putamen (MNI: x = -24, y = 17, z = 5) and right caudate (MNI: x = 9, y = 5, z = -13), as well as in the precuneus (MNI: x= -6, y = -37, z = 38), left angular gyrus (MNI: x = -45, y = -61, z = 44), and right inferior parietal lobule (IPL) (MNI: x = 45, y = -52, z = 47) (Figure 4B, Supplementary Table 5). This signal was also significant in an ROI analysis using anatomical masks of the precuneus and

bilateral caudate (**Figure 4D**). There were no significant clusters related to this regressor, and ROI analysis with the bilateral insula anatomical mask showed no significant activations (**Figure 4D**).

Taken together, these results strongly suggest that the positive activations (including the striatum) in the total chosen value GLM2 were primarily driven by the instrumental value component, while the negative activations (dmPFC and insula) were driven by a lower reliance on the instantaneous value.

Individual striatal activity is linked to task performance and planning ability

As an exploratory analysis, we hypothesized that participant-level BOLD activations in the striatum could be linked to the individual performance in the game and reliance on planning weights σ and δ.

Specifically, we identified two clusters of interest based on the analyses in GLM 2 (total value of the chosen option) and GLM 4 (set value of the chosen option). We created masks using the left and right striatum clusters (mainly including the putamen) from these analyses (**Figure 5**), then extracted individual GLM coefficients ("betas") that reflect the BOLD response to the value of the chosen option in these two clusters. For the left striatum, the analysis used the cluster identified for the same variable, while for the right striatum, the region was found to track a different variable.

We found an overall pattern reflecting a lateral dissociation. Higher individual activations in the left striatum predicted *lower* game performance and lower reliance on set values (as parameterized by σ and δ) (**Figure 5A**). Conversely, higher activations in the right striatum were *positively* correlated with performance and set weights σ and δ (**Figure 5B**). Some of these correlations were weak, and only three analyses survived the Bonferroni correction for multiple comparisons: the negative correlations in the left

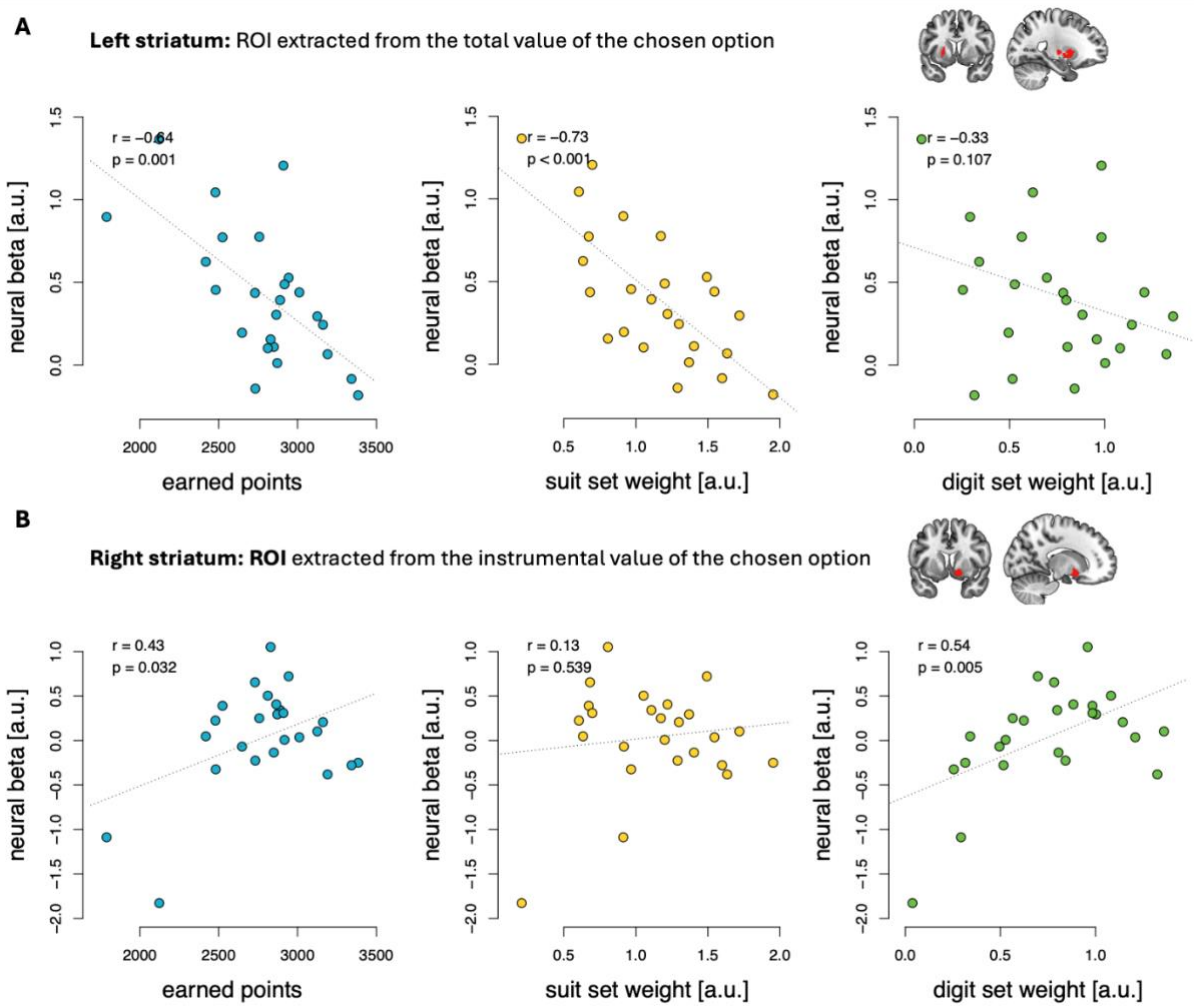

Figure 5. (**A**) Correlations between the participant's earned points and their neural beta (GLM coefficient) reflecting response to the total model-based generated value of the chosen option extracted from the *left* striatum mask that was calculated based on the BOLD correlates of the same variable (shown in red, also in Figure 3B). (**B**) Correlations between the participant's earned points and their neural beta (GLM coefficient) reflecting response to the total model-based generated value of the chosen option extracted from the *right* striatum mask that was calculated based on the BOLD correlates of the instrumental value of the chosen option (shown in red, also in Figure 4B). For all panels, each dot represents an individual participant, the line indicates the regression fit, and r and p-value denote the Pearson correlation (uncorrected). For all panels: N = 25.

striatum for earned points and suit weight, and the positive correlation in the right striatum for digit set weight. One can also note that two outliers could potentially drive the latter result (**Figure 5B**, right panel). However, overall, the results clearly point to the role of striatal circuit activation in successful planning at the individual level.

DISCUSSION

Valuation network and striatal encoding

In this study, we aimed to identify neural correlates of the instantaneous and future values of the same options using a new planning task that cannot be solved by simple reinforcement learning or exhaustive forward-looking tree search. Our results largely align with canonical analyses of brain valuation networks[15,16]. However, they also reveal a dissociation between instantaneous- and instrumental-value signals: we found that striatal activity scaled with instrumental (future-oriented) value rather than instantaneous point value. Our results align with some model-based reinforcement learning findings, as studies on multi-step tasks have shown that striatal BOLD reflects not only model-free rewards but also model-based (future) value prediction[31,35]. Previous studies have shown that regions such as the vmPFC and ventral striatum encode a common "currency" of subjective reward[13,15], typically responding positively to greater expected or received rewards. At the same time, work on intertemporal choice has highlighted dual systems: limbic-striatal areas respond to immediate gains, whereas lateral frontal/parietal regions are more involved regardless of immediacy[14].

Interestingly, we did not observe strong vmPFC encoding of either instantaneous or instrumental value at choice. One possibility is that the vmPFC may be engaged in evaluating or consolidating outcomes rather than in the planning process itself. Future studies could aim to dissociate the roles of this area using a more elaborate design. Previous fMRI studies suggest that the vmPFC integrates subjective value at outcome time[17] and may reflect the eventual chosen value after learning[51]. The absence of vmPFC activation here might also indicate that our task engaged the striatum as the primary encoder of predictive value during the anticipation phase.

The hippocampus and non-spatial planning

We observed no significant (above-threshold) hippocampal activation tied to value in our task. The hippocampus is often implicated in planning through its role in episodic simulation and cognitive mapping[52,53]. Prior work shows that hippocampal patterns reflect future goals[54] and support the prospective evaluation of a multi-step route[2]. Lesion studies also indicate that hippocampal damage impairs goal-directed planning[55]. Animal models reveal "preplay" sequences in CA3 that simulate upcoming paths[56–58]. Together, these findings suggest the hippocampus contributes to internal models of future states and instrumental outcomes.

This absence of hippocampal activation above threshold likely reflects task-specific factors. Our task was abstract and fully instructed; there was neither a need to learn spatial layouts nor uncertain transitions, nor a necessity for memory retrieval or spatial navigation. Recent work emphasizes that hippocampal engagement depends on relational or spatial demands[59,60]. Our participants could rely on simple rules rather than memory-guided search and use other areas, such as the dlPFC, for retention of rules in short-term memory. Another factor is the hippocampus's role in learning contingencies versus in online evaluation. Potentially, the hippocampus encodes learned associations that can later be retrieved for decision-making[36,38], and hippocampal planning functions emerge most strongly when building and using cognitive maps of an environment[1,3].

The dmPFC and insula: negative correlates of instantaneous value

Our results extend this framework in the planning context. We found that dmPFC and bilateral insula activity was inversely related to the chosen card's immediate value. This pattern is consistent with the idea that low immediate value creates a conflict or demand for control. When an option's immediate reward is low (i.e., a "bad deal" in the short term), the insula and dACC may signal increased salience or drive the recruitment of control to

pursue the long-term plan. Indeed, similar dmPFC/insula responses have been reported in human fMRI when participants forgo a sure reward to acquire information or pursue future goals[23,63].

The dorsomedial prefrontal cortex (dmPFC)and anterior cingulate cortex (aACC) are often described as part of the cognitive control or salience networks. They are often recruited when difficult trade-offs or cognitive conflicts arise[23,24]. Critically, these regions are suggested for computing the expected value of control**,** weighing whether it is worth exerting effort now to secure future payoff[23,26]. Typically, the dmPFC/dACC activity increases when an immediate reward is forgone for a larger, delayed, or uncertain outcome. Likewise, the anterior insula co-activates with the ACC when individuals face risk, loss, or negative affective states, possibly signalling aversive aspects of a decision[61,62].

Thus, in our task, these regions might reflect the neural effort or conflict required to down-regulate immediate reward in favor of long-term benefit. Notably, these regions did not correlate with instrumental value, suggesting they specifically encode an inverse or salience-weighted version of instantaneous reward.

Implications for planning and model-based decision-making

Our findings have implications for theories of planning and value-based choice. First, the dissociation between striatal and dmPFC/insula signals indicates that the brain may use separate valuation channels for different temporal or instrumental components of a choice. The striatum appears to accumulate expected total payoff (including future rewards), consistent with forward-looking model-based evaluation[4,31]. At the same time, the dmPFC and insula encode the relative disadvantage of acting impulsively, effectively biasing the decision system toward the instrumentally rewarding option when immediate rewards are potentially small. This complementary pattern aligns with the idea of parallel

decision systems[64] but suggests a potentially more integrated implementation. Rather than one module devoted to model-free impulses and another to model-based plans, our results could support a model in which the canonical reward areas (striatum/vmPFC) encode values on the full decision tree, while control regions modulate behavior according to expected gains[23]. Our approach, however, differs from the common two-step and related paradigms[31,35], which infer model-based activity through choices and prediction errors. In our task, participants did not need to compute prediction errors or update values, so any neural correlates we observe (e.g., striatal activation by instrumental value) reflect valuation during deliberative planning rather than learning.

Although dopamine is typically associated with immediate reward prediction errors[65], some work[35,66] shows that increasing dopamine can enhance model-based control. Potentially, the striatum is not limited to habitual "learned" values but can flexibly incorporate planned outcomes[39,49]. This suggests that "model-free" circuits may, in fact, have access to predictive representations (the successor representations or learned maps) that enable them to propagate future value without exhaustive tree search.

MATERIALS AND METHODS

Preregistration

The study, including the task code, study design, and main hypotheses, was preregistered on OSF (https://osf.io/f75gx). The target sample size was chosen based on a previous study from our lab[38], aiming to detect a moderate correlation between individual behavioral and neural measures ($r = 0.5$, $p = 0.05$, 80% power), resulting in a target sample of 28-30 participants.

Participants

We recruited 29 adult participants (18 female; average age 24.6 years, ranging from 18 to 35) from the Centre for Human Brain Health participant pool at the University of Birmingham. All participants were free of neurological or psychiatric disease, right-handed, and had normal or corrected-to-normal vision. Participants received a fixed fee of £30 plus a bonus of up to £10 based on their overall performance in the task, with an average earning of £36. We used all 29 participants for behavioral analyses and model fits. A physician identified one participant's brain as abnormal. Three participants showed excessive head movement during the task (more than 3 mm in any direction). We excluded the imaging data of these four participants, leaving 25 participants for the fMRI analyses. The University of Birmingham Ethics Board approved the experiment, and all subjects provided written informed consent.

Procedures

Each experimental session lasted less than 2 hours, including 30 minutes of document procedures and training, 10 minutes for scanner setup, 1 hour in the scanner, and 10

minutes of post-scanner procedures. Before entering the scanner, participants completed the consent forms and took a 15-minute tutorial to learn the task and practice several games. During behavioral training, subjects used two arrow keys on a laptop keyboard (left and right). Participants were then changed into MR-safe clothes, given earplugs, screened for metal, and positioned in the 3T Siemens MRI scanner. Inside the scanner, they used an fMRI-compatible 2-button box to respond. After completing the task, participants saw their total earnings before leaving the lab.

Task

We programmed the task in MATLAB using Psychtoolbox-3 (http://psychtoolbox.org/). We conducted three scanner runs. Each run consisted of 10 games, and each game included 10 choice trials (resulting in 300 trials per participant). Before the first trial of each run (1, 2, and 3), the computer displayed a "GET READY" screen for 5 seconds and waited for a scanner trigger to begin the run.

In each game, a participant played with a new 20-card deck made up of four standard suits (hearts, clubs, spades, and diamonds) and five digits (2 to 6), with exactly one card for each suit-digit combination. At the bottom of the screen, the computer displayed a grid showing the entire deck and highlighted the two current options (**Figure 1**). On each trial, the computer randomly drew two cards without replacement from the remaining deck and presented them as large stimuli on the left and right sides of the top part of the screen, while also indicating their positions in the grid.

Participants pressed the keys on the button box with their index and middle fingers to select the corresponding card (left or right) within a 3-second deadline. We defined response time (RT) as the interval from stimulus onset to keypress. After a valid choice, we highlighted the selected option and kept the display for a duration equal to 1 second

plus the remaining time in the 3-second response window (i.e., feedback duration = 1 second + [3 seconds − RT]), ensuring that the choice and feedback periods were the same across trials. The chosen card was then highlighted in white, and the unchosen card was removed from both the choice set and the grid.

If participants missed the deadline, the computer removed both cards, displayed "TOO LATE" for 1 second, and assigned zero points for that trial. We jittered the inter-trial interval from 2 to 6 seconds using a pre-generated schedule that was independently shuffled for each of the three runs.

The participants scored each selected card based on its digit value (2 to 6 points). After each game (10 trials), the computer calculated the total points for the round and awarded two set bonuses. It also tallied suit sets and digit sets of at least four cards, giving 20 points for each suit set and each digit set, then displayed the round summary and cumulative set counts. Throughout the session, the computer tracked the total number of suit sets and digit sets. It then calculated a session bonus by multiplying 20 points by the larger of these two cumulative counts.

We calculated the final score as the sum of round totals plus the total session bonus. We converted points to a monetary bonus at a rate of 500 points per £1 and paid a £30 show-up fee in addition to any performance-based bonus.

Imaging methods

We acquired functional images using a 3 T Siemens Prisma scanner with a 32-channel head coil at the Centre for Human Brain Health (University of Birmingham). We obtained a T1 anatomical scan and three runs, each with 700 acquisitions.

For the structural scan, we used a 3D inversion-recovery prepared gradient-echo sequence (Siemens MPRAGE) with these parameters: TR = 2.0 s, TE = 2.03 ms, TI = 880 ms, flip angle = 8°, GRAPPA = 2 with 24 reference lines, bandwidth = 240 Hz/pixel, phase oversampling = 5%, base matrix = 256 × 256 (phase-encoding steps = 269), slice/partition thickness = 1.0 mm.

For the functional scans, we used the CMRR multiband gradient-echo EPI sequence (multiband factor = 3; GRAPPA = 2) with 60 oblique slices (thickness = 2.4 mm; no gap; obliquity ≈ −40°). Imaging parameters included TR = 1.254 s, TE = 30 ms, flip angle = 65°, bandwidth = 1735 Hz/pixel, effective echo spacing = 0.33 ms, and total readout time = 29.37 ms.

We preprocessed the data with SPM12 (Wellcome Centre for Human Neuroimaging). First, we performed rigid-body realignment (Estimate & Reslice; two-pass register-to-mean) and saved the resliced images along with the mean functional image. Next, we applied slice-timing correction to the resliced images using the study's slice order, TR, TA, and the specified reference slice. Then, we co-registered the mean EPI (source) to the T1-weighted structural image (reference) with normalized mutual information, and we ran SPM's Unified Segmentation on the T1 to compute forward and inverse deformation fields. We normalized the slice-timing–corrected functional images to MNI space by applying the forward deformation field, resampling to 3×3×3 mm within the standard MNI bounding box, and finally smoothed the normalized images with a 6 mm FWHM Gaussian kernel.

Computational model

We used a computational model that assumed that participants made their choices on each trial by computing weighted values for each of the two cards, including that card's

point value V (that varied from 2 to 6), its suit set value V^{SUIT} (= 1 if the card can still make a complete set of 4 cards of the same suit), and its digit set value V^{DIGIT} (= 1 if the card can still make a complete set of 4 cards with the same digit):

$$P_R = \frac{1}{1 + e^{-\beta(V_R - V_L) - \sigma\,(V_R^{SUIT} - V_L^{SUIT}) - \delta(V_R^{DIGIT} - V_L^{DIGIT})}},$$

where P_R is the probability of choosing the right card, and R and L index the corresponding values for the right and left card. The model includes three free parameters: value weight β, suit set weight σ, and digit set weight δ. We extracted the values of these parameters for each participant for across-subject analyses.

The model assumed a standard logit softmax choice function and was fitted with a mixed-effects logit regression model that includes fixed effects, as well as random intercepts and random slopes, treating subjects as random effects (using the lme4 (1.1-35.5) R package).

We compared this model to five alternative models that excluded or combined some of these variables using the Bayesian Information Criterion (BIC), and found that this full model provided the best fit to the data (**Supplementary Table 1**).

To compute the trial-by-trial parametric modulator for the fMRI analyses, we generated the model-predicted value for each trial based on the model fit ($\beta(V_R - V_L) + \sigma\,(V_R^{SUIT} - V_L^{SUIT}) + \delta(V_R^{DIGIT} - V_L^{DIGIT})$) if the right card was chosen, and the negative of this expression if the left card was chosen.

fMRI analysis

We included three runs in the same analysis with constants added to the GLM to account for run differences, such as mean activation and scanner drift. We used SPM12 for the first-level analyses and SnPM13(using the standard settings: 5,000 permutations, cluster-forming threshold of 0.001, cluster-wise family-wise error rate of 0.05) for the second-level contrasts. We only report clusters with an FWE-corrected value lower than 0.05, as indicated by SnPM, with extent varying across analyses (see Supplementary Materials). Data visualization was performed using the MRIcroGL toolbox (https://www.nitrc.org/projects/mricrogl), and regions were labeled with the Harvard-Oxford Atlas and the Anatomy Toolbox distributed with bspmview (http://www.bobspunt.com/bspmview/).

For each GLM, we included an intercept regressor at trial onset (duration = response time) to model overall task activation on each trial, following standard fMRI GLM practices, and convolved the BOLD signal with the canonical HRF. We included six motion parameters as nuisance regressors. In each GLM, we set the regressors as parametric modulators of the HRF at the time of stimulus presentation. Automatic orthogonalization was turned off.

GLM1 included the following regressors: a dummy variable indicating if the participant chose a lower value card (= 1 if the participant chose a card with a value less than or equal to the unchosen card), trial number within the game (1-10), response time on the specific trial, and the overall session trial number.

GLM2 included the following regressors: the model-generated value of the chosen option, trial number within the game (1-10), response time on the specific trial, and the trial number across the whole session.

GLM3 included the following regressors: the digit value of the chosen option (V_R or V_L), trial number within the game (1-10), response time on the specific trial, and the trial number across the whole session.

GLM4 included the following regressors: the set value of the chosen option combining both suit and digit set value ($V_R^{DIGIT} + V_R^{SUIT}$ or $V_L^{DIGIT} + V_L^{SUIT}$), trial number within the game (1-10), response time on the specific trial, and the trial number across the whole session.

Region of interest (ROI) extraction. For ROI analyses, we extracted neural betas from the GLM results using the Marsbar toolbox (http://marsbar.sourceforge.net/). We used five standard masks from the AAL atlas (precuneus, left and right insula, and left and right caudate, as provided by the xjView toolbox, **Supplementary Figure 2**), and two striatal ROIs generated by GLM2 and GLM4 (**Figure 5**).

Data availability

The behavioral data are available on OSF (https://osf.io/95ucg/). The fMRI data are available from the corresponding author upon request.

Code availability

The analysis code is available on OSF (https://osf.io/95ucg/).

Acknowledgments

We thank Romy Froemer for helpful feedback and conversations.

Author contributions

A.K. designed the experiment and analyses. A.K. programmed the experiment. D.L. and A.K. collected the data. A.K. performed the data analysis and wrote the initial draft of the paper. D.L. co-wrote the paper. A.K. supervised the project.

Competing interests

The authors declare no competing interests.

REFERENCES

1. Mattar, M. G. & Lengyel, M. Planning in the brain. *Neuron* **110**, 914–934 (2022).
2. Balaguer, J., Spiers, H., Hassabis, D. & Summerfield, C. Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. *Neuron* **90**, 893–903 (2016).
3. Kuperwajs, I., Russek, E. M., Mattar, M. G., Ma, W. J. & Griffiths, T. L. Looking deeper into the algorithms underlying human planning. *Trends Cogn. Sci.* **0**, (2025).
4. Simon, D. A. & Daw, N. D. Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *J. Neurosci.* **31**, 5526–5539 (2011).
5. Jensen, K. T., Hennequin, G. & Mattar, M. G. A recurrent network model of planning explains hippocampal replay and human behavior. *Nat. Neurosci.* **27**, 1340–1348 (2024).
6. Kuperwajs, I., Schütt, H. H. & Ma, W. J. Using deep neural networks as a guide for modeling human planning. *Sci. Rep.* **13**, 20269 (2023).
7. van Opheusden, B. *et al.* Expertise increases planning depth in human gameplay. *Nature* **618**, 1000–1005 (2023).
8. Ingvar, D. H. ‘ Memory of the future’: an essay on the temporal organization of conscious awareness. *Hum. Neurobiol.* **4**, 127–136 (1985).
9. Suddendorf, T. & Corballis, M. C. The evolution of foresight: What is mental time travel, and is it unique to humans? *Behav. Brain Sci.* **30**, 299–313 (2007).
10. Shallice, T. Specific impairments of planning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **298**, 199–209 (1982).
11. Passini, R., Rainville, C., Marchand, N. & Joanette, Y. Wayfinding in dementia of the Alzheimer type: Planning abilities. *J. Clin. Exp. Neuropsychol.* **17**, 820–832 (1995).

12. Valls-Serrano, C., Verdejo-García, A. & Caracuel, A. Planning deficits in polysubstance dependent users: Differential associations with severity of drug use and intelligence. *Drug Alcohol Depend.* **162**, 72–78 (2016).

13. Kable, J. W. & Glimcher, P. W. The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci.* **10**, 1625–1633 (2007).

14. McClure, S. M. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science* **306**, 503–507 (2004).

15. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76**, 412–427 (2013).

16. Clithero, J. A. & Rangel, A. Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014).

17. Levy, D. J. & Glimcher, P. Comparing apples and oranges: Using reward-specific and reward-general subjective value representation in the brain. *J. Neurosci.* **31**, 14693–14707 (2011).

18. Peters, J. & Büchel, C. Neural representations of subjective reward value. *Behav. Brain Res.* **213**, 135–141 (2010).

19. Cox, K. M. & Kable, J. W. BOLD subjective value signals exhibit robust range adaptation. *J. Neurosci.* **34**, 16533–16543 (2014).

20. Holton, E. *et al.* Disentangling the Component Processes in Complex Planning Impairments Following Ventromedial Prefrontal Lesions. *J. Neurosci.* **45**, (2025).

21. McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. Time Discounting for Primary Rewards. *J. Neurosci.* **27**, 5796–5804 (2007).

22. O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, Reward, and Decision Making. *Annu. Rev. Psychol.* **68**, 73–100 (2017).

23. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240 (2013).

24. Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J. & Rushworth, M. F. S. Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* **9**, 940–947 (2006).

25. Cowen, S. L., Davis, G. A. & Nitz, D. A. Anterior cingulate neurons in the rat map anticipated effort and reward to their associated action sequences. *J. Neurophysiol.* **107**, 2393–2407 (2012).

26. Kolling, N., Behrens, T. E. J., Mars, R. B. & Rushworth, M. F. S. Neural mechanisms of foraging. *Science* **336**, 95–98 (2012).

27. Figner, B. *et al.* Lateral prefrontal cortex and self-control in intertemporal choice. *Nat. Neurosci.* **13**, 538–539 (2010).

28. Keramati, M., Dezfouli, A. & Piray, P. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLoS Comput. Biol.* **7**, e1002055 (2011).

29. Keramati, M., Smittenaar, P., Dolan, R. J. & Dayan, P. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proc. Natl. Acad. Sci.* **113**, 12868–12873 (2016).

30. Sezener, C. E., Dezfouli, A. & Keramati, M. Optimizing the depth and the direction of prospective planning using information values. *PLOS Comput. Biol.* **15**, e1006827 (2019).

31. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).

32. Daw, N. D. Model-based reinforcement learning as cognitive search: neurocomputational theories. *Cogn. Search Evol. Algorithms Brain*

http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.216.209 (2012).

33. Daw, N. D. Are we of two minds? *Nat. Neurosci.* **21**, 1497 (2018).

34. Daw, N. D. & Dayan, P. The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130478–20130478 (2014).

35. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron* **75**, 418–424 (2012).

36. Bornstein, A. M. & Daw, N. D. Cortical and Hippocampal Correlates of Deliberation during Model-Based Decisions for Rewards in Humans. *PLoS Comput. Biol.* **9**, e1003387 (2013).

37. Bornstein, A. M. & Daw, N. D. Dissociating hippocampal and striatal contributions to sequential prediction learning: Sequential predictions in hippocampus and striatum. *Eur. J. Neurosci.* **35**, 1011–1023 (2012).

38. Konovalov, A. & Krajbich, I. Neurocomputational Dynamics of Sequence Learning. *Neuron* **98**, 1282-1293.e4 (2018).

39. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Comput. Biol.* **13**, e1005768 (2017).

40. Momennejad, I., Otto, A. R., Daw, N. D. & Norman, K. A. Offline replay supports planning in human reinforcement learning. *eLife* **7**, e32548 (2018).

41. Konovalov, A. & Krajbich, I. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nat. Commun.* **7**, 12438 (2016).

42. Konovalov, A. & Krajbich, I. Mouse tracking reveals structure knowledge in the absence of model-based choice. *Nat. Commun.* **11**, 1893 (2020).

43. Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. Model-based choices involve prospective neural activity. *Nat. Neurosci.* https://doi.org/10.1038/nn.3981 (2015) doi:10.1038/nn.3981.

44. Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W. & Rangel, A. Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *J. Neurosci.* **28**, 5623–5630 (2008).

45. Feher da Silva, C., Lombardi, G., Edelson, M. & Hare, T. A. Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nat. Hum. Behav.* 1–14 (2023) doi:10.1038/s41562-023-01573-1.

46. Moneta, N., Garvert, M. M., Heekeren, H. R. & Schuck, N. W. Task state representations in vmPFC mediate relevant and irrelevant value signals and their behavioral influence. *Nat. Commun.* **14**, 3156 (2023).

47. Garvert, M. M., Saanum, T., Schulz, E., Schuck, N. W. & Doeller, C. F. Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nat. Neurosci.* 1–12 (2023) doi:10.1038/s41593-023-01283-x.

48. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron* **91**, 1402–1412 (2016).

49. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).

50. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron* **81**, 687–699 (2014).

51. Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).

52. Addis, D. R. & Schacter, D. The Hippocampus and Imagining the Future: Where Do We Stand? *Front. Hum. Neurosci.* **5**, (2012).

53. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).

54. Brown, T. I. *et al.* Prospective representation of navigational goals in the human hippocampus. *Science* **352**, 1323–1326 (2016).

55. Vikbladh, O. M. *et al.* Hippocampal Contributions to Model-Based Planning and Spatial Memory. *Neuron* **102**, 683-693.e4 (2019).

56. Johnson, A. & Redish, A. D. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12176–12189 (2007).

57. Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).

58. Wikenheiser, A. M. & Redish, A. D. Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* **18**, 289–294 (2015).

59. Crivelli-Decker, J. *et al.* Goal-oriented representations in the human hippocampus during planning and navigation. *Nat. Commun.* **14**, 2946 (2023).

60. Kaplan, R. *et al.* The Neural Representation of Prospective Choice during Spatial Planning and Decisions. *PLOS Biol.* **15**, e1002588 (2017).

61. Menon, V. & Uddin, L. Q. Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* **214**, 655–667 (2010).

62. Paulus, M. P. & Stein, M. B. An insular view of anxiety. *Biol. Psychiatry* **60**, 383–387 (2006).

63. Cogliati Dezza, I., Maher, C. & Sharot, T. People adaptively use information to improve their internal states and external outcomes. *Cognition* **228**, (2022).

64. Redish, A. D. *The Mind within the Brain: How We Make Decisions and How Those Decisions Go Wrong*. (Oxford University Press, 2013).

65. Schultz, W. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).

66. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).

67. Charpentier, C. J., Bromberg-Martin, E. S. & Sharot, T. Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proc. Natl. Acad. Sci.* **115**, E7255–E7264 (2018).

68. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).