TEMPORAL DYNAMICS OF MODEL-BASED CONTROL REVEAL ARBITRATION BETWEEN MULTIPLE TASK REPRESENTATIONS

Jungsun Yoo
Department of Cognitive Sciences
Center for Theoretical Behavioral Sciences
University of California, Irvine
Irvine, CA 92697
jungsun.yoo@uci.edu

Aaron M. Bornstein
Department of Cognitive Sciences
Center for the Neurobiology of Learning and Memory
Center for Theoretical Behavioral Sciences
University of California, Irvine
Irvine, CA 92697
aaron.bornstein@uci.edu

ABSTRACT

Previous work demonstrates that people arbitrate between control algorithms – e.g. model-based or model-free – according to their relative certainty at the given moment. Here, we examined whether a similar uncertainty-based arbitration could explain the relative pattern of reliance on distinct *representations*. We employ a novel variant of a standard, two-stage decision task. This task allows us to behaviorally capture the within- and across-trial dynamics of model-based planning. We jointly fit choices and response times with a new computational model that revealed how people select among multiple task representations during planning in environments of differing state-space complexity. In particular, we examined how the reliance on task representations changed both as a function of experience, within-subject, and task complexity, across-subjects (total $n = 426$). We show that both the complexity of the environment and experience with a given contingency structure inform the kinds of representations we use to make decisions: at the early stages of the task, people start with "conjunctive" representations (combining co-occurring first-stage states) in simpler environments, but a "separated" representation (splitting states according to their second-step outcomes) is preferred in more complex environments. With experience, this pattern is reversed. We show that this shift is likely to be governed by a change in objectives: initially, people focus on minimizing uncertainty, and once this is achieved, they transition to prioritizing efficiency. Taken together, we show that people not only arbitrate between different modes of control, but also between types of representations for efficient planning.

Keywords Planning · Representation · Reinforcement learning

1 Introduction

Planning, the process of using an internal representation of task contingencies to select actions on the basis of their anticipated outcomes, is essential for complex organisms to survive in complex environments. The complexity of an environment imposes various challenging conditions for an organism to plan effectively. Namely, they must choose between representing the environment in its full complexity, or choosing a simpler representation that may impose lower computational costs (Yoo, Chrastil, and Bornstein, 2024) – potentially at the expense of less flexibility in the face of change. In most real-world situations, an organism must perform this selection on the fly, and adjust as new information becomes available. How do organisms navigate these trade-offs? Previous work suggests that multiple kinds of representations are learned in parallel, and uncertainty guides the selection in a way that the representation with the least uncertainty is chosen at a given moment (Wang, Feng, and Bornstein, 2022; Lengyel and Dayan,

2007). For example, when individuals are navigating through a city that they have just moved to, they may rely on a more map-like representation ("allocentric" representation) rather than a series of paths they have previously taken ("egocentric" representation), because the latter carries high uncertainty due to lack of experience. However, as experience accumulates and certainty increases, there is a shift towards prioritizing efficiency (Lengyel and Dayan, 2007) – for example, once individuals become confident in a series of paths through repeated encounters, they may favor the more efficient representations (series of paths) over the more precise but resource-intensive representations (looking up a map).

A division of task representations that has been studied in reinforcement learning (RL) is *separated* vs. *conjunctive* (Niv et al., 2015). Learning based on separated representations involves assigning values to compound features via the sum of separated features, while the compositional features are treated as single units in conjunctive learning (Ballard, Wagner, and McClure, 2019). Humans learn both conjunctive and separated representations in reinforcement learning, in a manner such that values assigned to conjunctive representations (i.e., AB) are "spread" to individual elements (i.e., B; Ballard, Wagner, and McClure, 2019). An important question can be raised by this finding, which is whether the parallel acquisition of conjunctive and separated representations during model-free reinforcement learning could be generalized to *multi-step planning* or *model-based control*, considering that planning involves a sophisticated representation selection process (Ho et al., 2022).

A key tool for studying model-based control in humans is the two-stage task (Daw et al., 2011; Decker et al., 2016). In this task, participants are faced with two consecutive decisions, where first-stage decisions stochastically lead to distinct second-stages. Since reward is given following the second-stage decision, the best decision to take in the first stage involves using information about the contingency structure that leads to a given second stage – in other words, the model. Analysis of this task tends to focus on the relative reliance on model-based versus model-free algorithms. Theory and empirical findings support the idea that this reliance is informed by the relative uncertainty of each system (Daw, Niv, and Dayan, 2005; Lee, Shimojo, and O'Doherty, 2014; Kim et al., 2019), as well as the relative computational costs (Keramati, Dezfouli, and Piray, 2011; Kool, Cushman, and Gershman, 2016; Milli, Lieder, and Griffiths, 2021). Empirically, the dynamic of arbitration between model-based vs. model-free control takes the form of people first relying on goal-directed control but later using a more cost-efficient model-free control after extensive experience, and the switch of the dominant mode of control happens as a function of the trade-off between the uncertainty and cost-efficiency of each control mechanism (Daw, Niv, and Dayan, 2005; Lengyel and Dayan, 2007; Keramati, Dezfouli, and Piray, 2011; Wang, Feng, and Bornstein, 2022). In summary, our control systems are dependent on the trade-off between uncertainty and efficiency at the given moment.

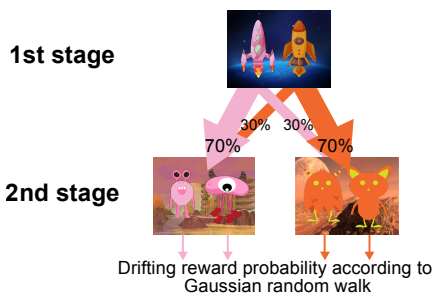
However, an aspect that has been uninterrogated by previous studies is how people construct and select *representations* for model-based control. Although extensive studies have established relations between model-based control and other cognitive functions (Hunter, Bornstein, and Hartley, 2018; Gillan et al., 2016; Vikbladh et al., 2019), as well as its sensitivity to stress (Wyckmans et al., 2022; Park, Lee, and Chey, 2017), and dependence on working memory capacity (Otto et al., 2013a,b) and episodic memory manipulations (Vikbladh, Shohamy, and Daw, 2017), how people decide to represent this structure remains unexplored. There are several reasons that could contribute to this gap. First, the standard task and analysis approach obscure representation selection – specifically, work in process-tracing suggests that subjects begin selecting actions *before* the trial onset (Konovalov and Krajbich, 2016, 2020), making it difficult to identify when this process begins and how long it takes – and, critically, how this time adjusts to the environment and experience. Another drawback is that the standard task has a fixed, minimal state-space complexity. Varying this aspect of the task is crucial to understanding how control dynamics change when task models grow exponentially larger. This is important because state-space complexity has been identified as a factor that interacts with state-space uncertainty to determine the arbitration between model-based vs. model-free control (Kim et al., 2019). We go one step further by examining how this interaction affects arbitration of representations *within* model-based control.

We posit that in model-based control, people mediate the arbitration of conjunctive and separated representations on the basis of their relative uncertainty and efficiency, which is a function of their variable experiences. Initially, individuals are likely to rely on representations that minimize uncertainty. In environments with fewer states, there are fewer conjunctive vs. separated states – and thus the former would accelerate the uncertainty minimization process. However, the opposite would hold true in environments with more states, where a separated representation would be more parsimonious. In all environments, as experience allows individuals to establish more precise representations, they should shift towards the representation that offer greater efficiency. To validate this idea, we investigate how behavioral signatures of model-based control develop over time under varying levels of environmental complexity.

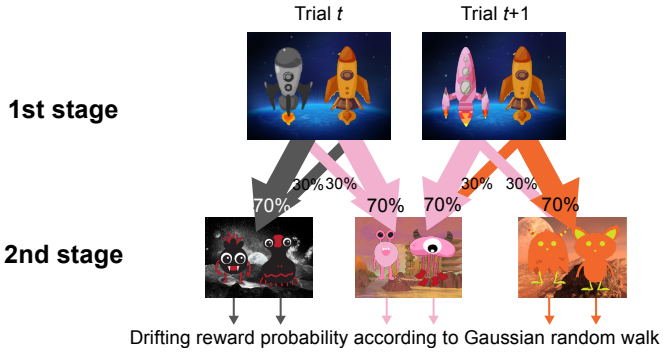
In summary, the unexplored questions in the previous literature could be characterized as 1) whether people employ both conjunctive and separated representations in model-based control as they do in model-free reinforcement learning (Ballard, Wagner, and McClure, 2019), and 2) how people construct and select representations as a function of uncertainty and cost-efficiency, as in arbitration of control (Lengyel and Dayan, 2007). To address these unanswered questions, we propose a novel variant of the two-stage task, the *multinomial* two-stage task (MTST), that manipulates the level of state-space complexity. This is achieved by increasing the number of possible first-stage options and presenting randomized combinations of the options at each first-stage choice (Fontanesi et al., 2019). Like the original version, two options are presented on first- and second-stage decisions, where the first-stage option stochastically leads to the second-stage. The increased number of possible states at the first stage renders it difficult for people to predict and plan the first-stage decisions beforehand, thereby aligning representation selection to the start of the trial. Time-locking the availability of decision-relevant information to the first-stage stimulus onset allows us to investigate the unexplored subprocesses of decisions such as representation selection. Critically, varying the number of possible first-stage options ($k \in \{2, 3, 4, 5\}$) also introduced a differential numerical balance between conjunctive ($N = \binom{k}{2}$) and separated ($N = k$) states. For example, in the canonical two-stage task, there are two separated states and one conjoined state (Fig. 1d). Increasing the number of first-stage options to three results in an equal number of conjoined and separated states (three), and beyond that, the number of potential conjoined representations outnumbers the number of separated states. We expect that people's reliance on conjunctive vs. separated state-space representations in varying state-space complexity will change at this "crossover point" – i.e., learning conjunctive representations would be faster in less complex environments. To capture the relative dependence on conjunctive vs. separated representations and how that selection changes as a function of learning, we examine the within- and across-trial temporal dynamics by a set of analyses that jointly takes into account people's choice and response time.

Using this novel task and analyses, we found that state-state complexity selectively increases participants' *first-stage* response time, where representation selection is assumed to take place. If representation selection were to take place during first-stage decisions, it is reasonable to assume that it happens before the valuation process. In other words, the action selection process could only happen after the representation of the two options have been loaded into working memory (Nunez et al., 2019; Kraemer and Gluth, 2023). Consistent with recent work showing that a portion of response time (i.e., *non-decision time*) is crucial for memory retrieval in advance of value-guided decisions (Kraemer and Gluth, 2023), results from our analyses showed that variability in the first-stage non-decision time can be explained by representational uncertainty, which suggests that first-stage non-decision time could encompass the representation selection process. Furthermore, we examined the relative use of conjunctive vs. separated representations across complexity (between-subjects) and experience (within-subject) by analyzing their relative contribution to explaining representational uncertainty. We found that people in complex environments (i.e., conditions where the number of conjoined states surpasses the separated states) first rely on separated representations but gradually switch to using conjunctive representations. The reverse pattern is observed for less complex environments where the number of separated states are equal to or greater than the number of conjunctive states. The results from our simulation and analyses explains this transition as a change in objectives: regardless of state-space complexity, subjects initially rely on representations that minimize uncertainty, but with repeated experience, they switch to representations that enhance *retrieval efficiency*, defined as the degree to which pre-accumulated evidence towards one option in the first-stage reduces second-stage non-decision time.

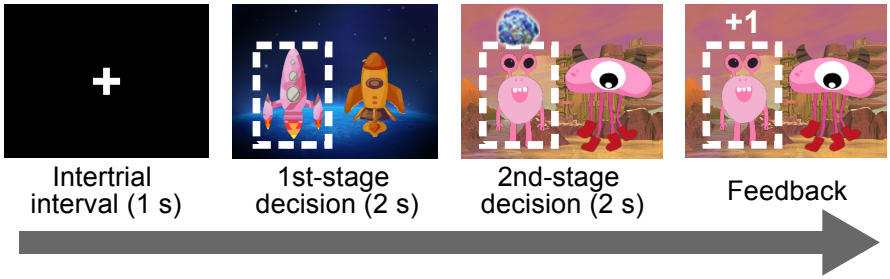
Taken together, our results suggest that individuals use representational uncertainty and efficiency to mediate the use of conjunctive versus separated representations, resulting in differential temporal dynamics as a function of environmental complexity. This finding establishes a new dimension of individual and environmental variation in model-based planning, which could be fruitful for the analysis of computations thought to be critical indicators of healthy cognitive function (Gillan et al., 2016).



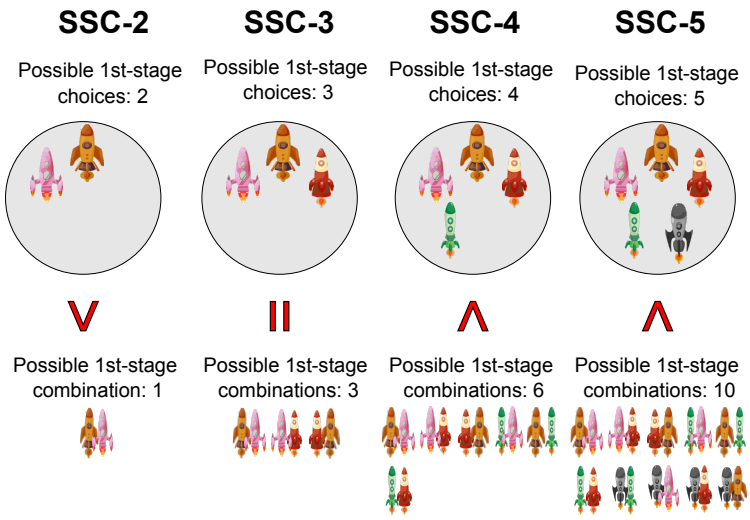
(a) Canonical two-stage task (SSC=2).



(b) Our proposed multinomial two-stage task (SSC>2).



(c) Time course of one trial.



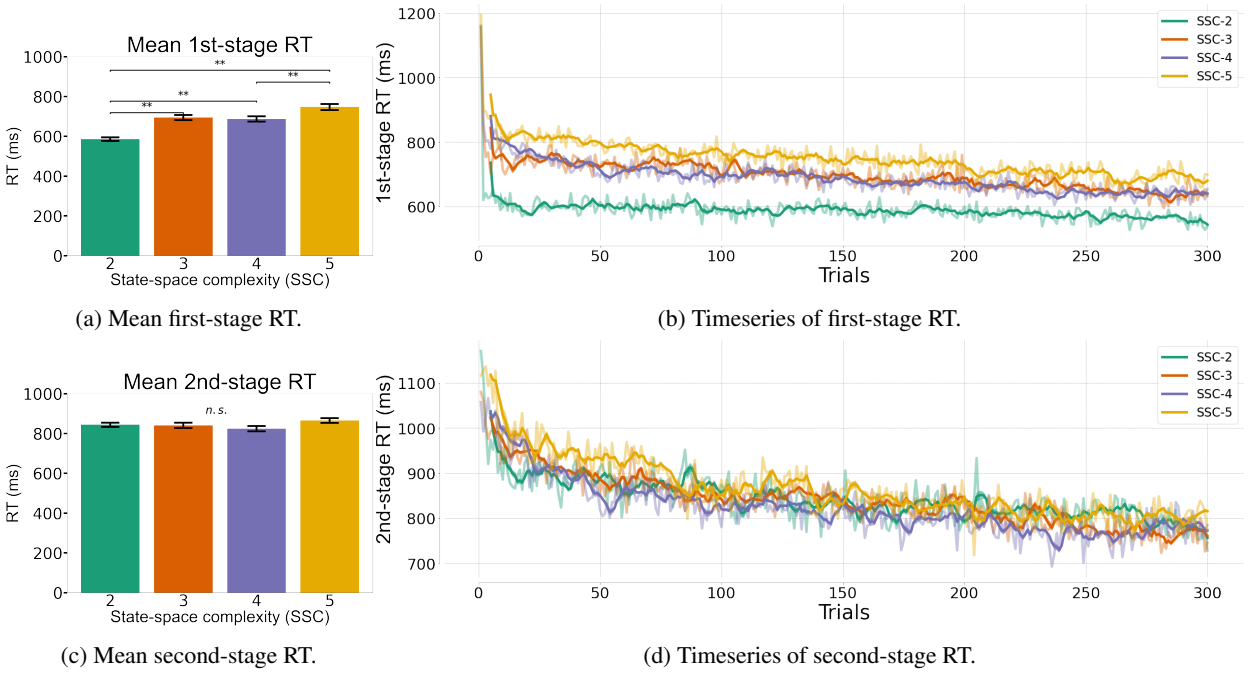
(d) Possible options in each condition.

Figure 1: Experimental task. (**a-b**) Each experiment (SSC-k, where SSC refers to state-space complexity and k refers to possible first-stage options, $k \in \{2, 3, 4, 5\}$) consists of 300 trials, with 5 "catch" trials randomly interspersed. (**a**) Experimental design of the canonical two-stage task (TST; equivalent to SSC-2). Note that the same first-stage state appeared on every trial, which allows for precomputation of plans in the intertrial interval (ITI). (**b**) Experimental design of our proposed multinomial TST (MTST). SSC corresponds to the number of spaceships in the first stage, where each spaceship mainly leads to its associated second stage (planet). Therefore, each first-stage state varies according to the combination of the spaceships. (**c**) Temporal order of (M)TST. Both the first and second stage decisions have a 2-second response window, followed by feedback. ITI lasted for 1 second with a fixation cross. (**d**) The 1st-stage options and their possible combinations are shown for each condition. Note that the "crossover point" occurs at SSC-3, after which the number of separated states (upper row) is fewer than the number of conjunctive states (bottom row).

2 Results

In this study, we investigated what the within- and across-trial dynamics of model-based control reveal about the format and use of individuals' internal representations of task contingencies. Specifically, we examined how an index of *representational uncertainty* informed decision dynamics, and how this relationship changed, within-subjects, as a function of experience with the task structure. We used this index to ask how individuals adjusted their representation use in response to greater computational demands, by manipulating *state-space complexity* (SSC) across-subjects (Fig. 1). To do this, we employed novel variants of the standard *two-stage task* (TST; Daw et al., 2011; Decker et al., 2016). These variants differed by the number of first-stage options possible on each trial, ranging from least complex (SSC-2: 2 possibilities, the standard task) to most (SSC-5: 5 possibilities). Critically, in the larger state spaces ($SSC > 2$), this manipulation overcame a known limitation of the standard task by preventing individuals from anticipating the next choice (Konovalov and Krajbich, 2016, 2020), thus giving us the ability to more robustly examine trial-locked decision dynamics.

2.1 State-space complexity selectively increases first-stage RT



(a) Mean first-stage RT.

(b) Timeseries of first-stage RT.

(c) Mean second-stage RT.

(d) Timeseries of second-stage RT.

Figure 2: Behavioral analyses across conditions. Error bars indicate standard error. Significant results of pairwise *t*-tests are indicated with asterisks ($* = p < .05, ** = p < .01, *** = p < .001$). **(a, c)** Mean first- and second-stage RT across different SSC. Prior to statistical tests such as one-way ANOVA and *t*-tests, each participant's mean RTs for first- and second-stage decisions were log-transformed. **(a)**. First-stage RT increased as a function of SSC. **(b)** First-stage RT significantly decreased over time for all conditions. SSC-2 showed less steeper decrease than other conditions. **(c)** A one-way ANOVA indicated no significant difference in mean second-stage RT. **(d)** Second-stage RT also significantly decreased as a function of experience. A significant difference across conditions was observed for slopes of second-stage RT.

First, utilizing a 2×2 analysis of variance (ANOVA), we examined the effects of within-task level (i.e., first- or second-stage) and across-task complexity on response times (RT). We identified a main effect for both stage (first- or second; $F_{(3,412)} = 309.14, p < .001$) and state-space complexity ($F_{(3,412)} = 16.57, p < .001$), along with a significant interaction effect between these two factors ($F_{(3,412)} = 13.49, p < .001$). The interaction was due to the differential effects of SSC within each stage; consistent with the idea that individual decision deliberation scales with task complexity, we found that higher SSC resulted in an increase in first-stage RT (mean and standard error for log first-stage RT: 6.31 (.02), 6.47 (.02), 6.45 (.02), 6.54 (.02); $F_{(3,412)} = 26, p < .001$, Fig. 2a). A post-hoc *t*-test

(Tukey's HSD) revealed that the log first-stage RT of SSC-2 was significantly lower than other conditions (SSC-2 vs. SSC-3: mean difference=.025, 95% CI [.01, .04], $p = .001$; SSC-2 vs. SSC-4: mean difference=.02, 95% CI [.01, .03], $p = .001$; SSC-2 vs. SSC-5: mean difference=.04, 95% CI [.02, .05], $p = .001$), and first-stage RT in SSC-5 were also significantly higher than SSC-2 and SSC-4 (SSC-4 vs. SSC-5: mean difference=.01, 95% CI [.002, .02], $p = .019$). No other pairwise comparisons were found to be statistically significant. However, supporting the idea that task complexity selectively increased planning costs, second-stage RT did not change between SSC variants (mean and standard error for log second-stage RT: 6.68 (.01), 6.67 (.02), 6.64 (.02), 6.7 (.01); $F_{(3,412)} = 1.86, p > .13$, Fig. 2c). Further, our data suggests that this additional planning time was well-spent, as increased task complexity did not result in diminished task performance: on average, participants' total reward did not differ between conditions ($F_{(3,412)} = 1.09, p > .35$; mean and standard error of total points earned out of 300 points: $155.7(2.02), 154.02(1.73), 151.25(1.9), 154.52(1.49)$ for SSC-2, SSC-3, SSC-4, SSC-5, respectively).

Supporting the idea that decision deliberation becomes less effortful as experience with the task increases, we found that both first- and second-stage RT decreased with experience (first-stage RT: mean and standard error: $-.19(.04), -.43(.04), -.47(.05), -.52(.05)$; *t*-tests against 0: $t_{(99)} = -4.73, p < .001, t_{(104)} = -10, p < .001, t_{(102)} = -8.78, p < .001, t_{(106)} = 10.3, p < .001$ for SSC-2,3,4, and 5, respectively, Fig. 2b; second-stage RT: mean and standard error: $-.44(.05), -.61, (.05), -.57(.06), -.68(.05)$; *t*-tests against 0: $t_{(99)} = -8.4, p < .001, t_{(104)} = -12.46, p < .001, t_{(102)} = -9.77, p < .001, t_{(106)} = -12.33, p < .001$ for SSC-2, SSC-3, SSC-4, and SSC-5, respectively, Fig. 2d). A 2×2 ANOVA conducted on the slopes of RT timeseries, considering both stage (first- and second-stage) and state-space complexity as factors, revealed significant main effects for both stage ($F_{(3,412)} = 23.87, p < .001$) and complexity ($F_{(3,412)} = 10.94, p < .001$), without an interaction between the two factors ($F_{(3,412)} = .72, p > .5$). The slope of the decrease in RT was greater for the higher-complexity conditions (SSC>2; first stage: $F_{(3,412)} = 9.25, p < .001$, second stage: $F_{(3,412)} = 3.23, p = .02$), suggesting that individuals' trial-by-trial learning in these conditions was more impactful on overall computation time in both first and second stages. In first-stage RT, SSC-2 showed a less steep decrease than other conditions (Tukey's HSD: SSC-2 vs. SSC-3: mean difference=-.238, 95% CI [-.411, -.065], $p = .002$; SSC-2 vs. SSC-4: mean difference=-.277, 95% CI [-.45, -.103], $p = .001$; SSC-2 vs. SSC-5: mean difference=-.325, 95% CI [-.497, -.154], $p = .001$, Fig. 2b). Unlike the gradual decrease in mean first-stage RT, the slope did not decrease as a function of SSC for higher states (i.e., there were no significant differences among SSC-3, SSC-4 or SSC-5). Despite a significant difference across conditions for slopes of second-stage RT, only SSC-2 differed from SSC-5 (Tukey's HSD: SSC-2 vs. SSC-5: mean difference=-.231, 95% CI [-.429, -.034], $p = .014$; no other conditions yielded statistically significant pairwise differences, Fig. 2d). The main effect of stage was observed such that second-stage RT, regardless of SSC, decreased more steeply overall compared to first-stage RT (mean and standard error of slopes: first-stage RT = $-.4(.024)$, second-stage RT = $-.58(.027)$).

Given that the second-stage RT did not exhibit substantial variance across SSC, we consequently ruled out second-stage RT and performance from further analyses of interest. These behavioral results suggest that planning-specific deliberation scales with complexity and experience, and these can be decoupled from other dynamics such as overall RT (vs. second-stage RT) or learning.

2.2 First-stage non-decision time changes with representational uncertainty

To determine the nature of the computations performed during complexity-sensitive deliberation, we jointly fit choices and RT with a custom, two-stage, variant of the Reinforcement Learning Drift-Diffusion Model (RLDDM; Pedersen and Frank, 2020, see *Methods* for details and Section 2.9 for response time recovery and model validation checks). To accomplish this, we decomposed the first-stage non-decision time (ndt_1) of each subject as a function of their representational uncertainty. We reasoned that, because this value-based decision task had identical perceptual demands across conditions (Nunez et al., 2019), cross-trial fluctuations in ndt_1 could potentially be explained by memory retrieval triggered by option presentation (Bornstein et al., 2017; Kraemer and Gluth, 2023). We operationalized the changing difficulty of memory-based option evaluation by modeling how individuals' uncertainty about the transition structure changed with experience, under the assumption that individuals attempted to identify the decision trees following from the presented options (Vikbladh, Shohamy, and Daw, 2017), and that this identification was more time-consuming when this structure was more uncertain (Wang, Feng, and Bornstein, 2022; Khoudary, Peters, and Bornstein, 2022).

To do so, we devised a *dynamic & hybrid RLDDM* that decomposes ndt_1 into components that are explained by representational uncertainty (specifically, U_{rep}, the variance of the subjects' estimated distribution over possible

transition probabilities; details are in *Methods* Section 4.4.3) and all other components. In other words, we regressed out an index of representational uncertainty (U_{rep}) from ndt_1, such that

$$ndt_1 = \beta_{ndt1} \times U_{rep} + t_{res} \quad (1)$$

We refer to the proportion of ndt_1 accounted for by U_{rep} as the *representation selection time* ($t_{rep} = \beta_{ndt1} \times U_{rep}$), and label the remaining unexplained component as *residual time* (t_{res}; Figure 3). We found that representational uncertainty significantly contributed to ndt_1; the proportion of time explained by representational uncertainty (i.e., β_{ndt1}) was significantly greater than 0 for all conditions (*t*-tests against 0: $t_{(99)} = 4.28, p < .001, t_{(104)} = 6.28, p < .001, t_{(102)} = 9.66, p < .001, t_{(106)} = 10.27, p < .001$ for SSC-2, SSC-3, SSC-4, and SSC-5, respectively; Fig. 4a; the model-fitting results of other parameters are described in Supplementary Figure 1). Consistent with the idea that SSC-2 was easier to learn and options could be anticipated before trial start, β_{ndt1} was significantly lower for SSC-2 than other conditions ($F_{(3,412)} = 10.73, p < .001$; Tukey's HSD: SSC-2 vs. SSC-3: mean difference=2.59, 95% CI [.89, 4.28], $p = .001$; SSC-2 vs. SSC-4: mean difference=2.66, 95% CI [.96, 4.36], $p = .001$; SSC-2 vs. SSC-5: mean difference=3.56, 95% CI [1.87, 5.24], $p = .001$; no other significant pairwise differences were found).

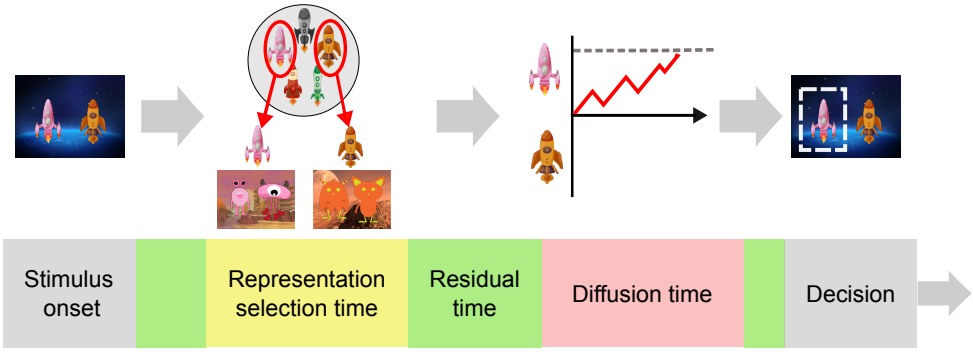
We next examined how these dynamics changed with experience, by fitting the dynamic & hybrid RLDDM to a sliding window of 30 trials. The proportion of representational uncertainty in ndt_1 did not change with experience in SSC-2, while it significantly decreased in other conditions (mean and standard error of the slope of β_{ndt1}: $-.001$ (.001), $-.008$ (.0002), $-.013$ (.002), $-.016$ (.002) for SSC-2, SSC-3, SSC-4, and SSC-5, respectively; Fig. 4b; the sliding-window model-fitting results of other parameters are described in Supplementary Figure 2). While higher-complexity conditions (SSC>2) showed a gradual decrease in β_{ndt1} over experience, this was not observed for SSC-2 (*t*-tests against 0: $t_{(99)} = -1.24, p = .22, t_{(104)} = -4.86, p < .001, t_{(102)} = -8.53, p < .001, t_{(106)} = -7.04, p < .001$ for SSC-2, SSC-3, SSC-4, and SSC-5, respectively). The slopes generally decreased (i.e., became steeper) with SSC ($F_{(3,412)} = 15.31, p < .001$; Tukey's HSD: SSC-2 vs. SSC-3: mean difference=-.007, 95% CI [-.013, -.001], $p = .03$; SSC-2 vs. SSC-4: mean difference=-.01, 95% CI [-.018, -.006], $p = .001$; SSC-2 vs. SSC-5: mean difference=-.015, 95% CI [-.021, -.009], $p = .001$; SSC-3 vs. SSC-5: mean difference=-.008, 95% CI [-.014, -.002], $p = .002$; no other pairwise comparisons yielded statistically significant difference). These results suggest that the complexity-related increase in first-stage RT is related to the time spent selecting among decision trees in more complex variants of the task.

2.3 Environmental complexity and experience modulate the dominant form of state-space representation

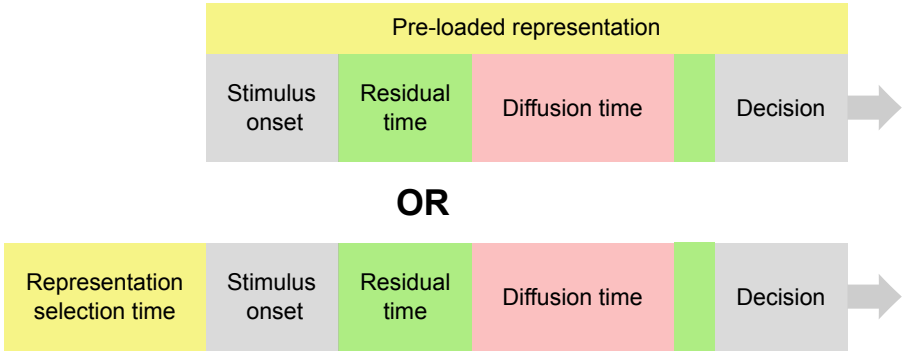
Having identified the representation selection component of ndt_1, we next examined whether decision dynamics could adjudicate hypotheses about the *content* of representations (decision trees) used to inform decisions. Specifically, previous work has suggested that individuals in repeated reinforcement learning tasks can learn to develop *conjunctive* states and assign action values to these new states (Ballard, Wagner, and McClure, 2019). Here, we examine how the learning and use of decision trees involving these conjunctive representations evolves alongside that of more standard *separated* state representations. Specifically, the representation selection process (t_{rep}) assumes these *two* sets of decision trees, which is formalized by letting ndt_1 vary as a linear function of the uncertainty in each (Fig. 5; *Methods* Section 4.4.3). Normatively, individuals should select the representation that has lower uncertainty, and use this to guide their choices (Wang, Feng, and Bornstein, 2022). Critically, the use of these representations should vary both with the relative uncertainty, within-session, as a function of experience, and also between conditions, as a function of task complexity. This is because in the two lower SSC conditions (2&3), there are an equal or lower number of possible decision trees (and, thus, lower uncertainty about *which* decision tree to select) in the conjunctive state space, while the opposite is true in the higher SSC conditions.

Our results suggest that individuals were sensitive to this distinction, initially favoring the separated states when they were fewer in number in SSC-4 and SSC-5, the two conditions where conjunctive and separated state spaces implied diverging numbers of potential decision trees (Fig. 6b). Interestingly, in these conditions, participants' ndt_1 grew to more strongly reflect the uncertainty in the conjunctive representation with experience. This may be due to uncertainty in the separated state spaces reaching asymptote in fewer trials, thus leaving ndt_1 to be affected largely by still-developing conjunctive representations.

Specifically, the weighting parameter between the uncertainty measure of conjunctive vs. separated representations (w_{rep}) generally increased with SSC (one-way ANOVA: $F_{(3,412)} = 19.2, p < .001$; Tukey's HSD: SSC-2 vs. SSC-4: mean difference=-1.84, 95% CI [-3.45, -.24], $p = .017$; SSC-2 vs. SSC-5: mean difference=2.76, 95% CI [1.17,

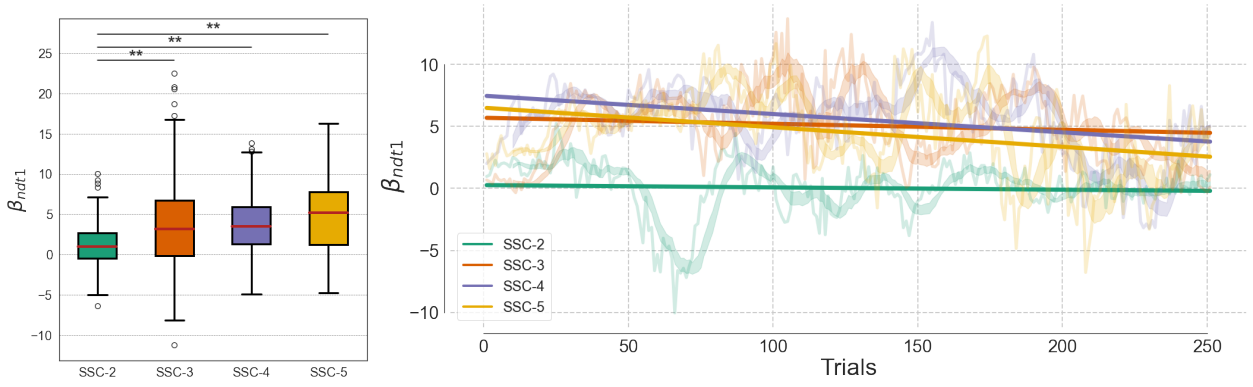


(a) Our proposed within-trial temporal dynamics of online planning.



(b) A schematic of the within-trial computations within the canonical TST.

Figure 3: Schematics of dynamics within a decision. **(a)** In our model, *representation selection time* and *residual time* add up to define what has been previously captured by non-decision time. Representation selection time refers to the stage where people identify which representations (models) to use for action selection; in our model, it is defined as $\beta_{ndt1} \times w_{rep}$. *Diffusion time* captures the action evaluation and selection process, once the two options have been loaded during the representation selection time. Like the previous DDM literature, diffusion time ends as soon as the evidence for one option over another reaches the decision threshold. Residual time is the component of the RT that is not explained by either representation time or diffusion time, and could include processes such as motor execution or stimulus encoding. Note that while the temporal sequence of the residual time vs. representation or diffusion time is unspecified, we assume sequential precedence of representation selection activity over action selection, such that the amount of time spent in representation selection could affect the starting point of the diffusion time (Bornstein et al., 2023). **(b)** For the canonical, two-state version of the task, we argue that due to having one possible combination of first-stage decisions, the representation of the forthcoming trials can be anticipated at any time besides the presentation of the options. This selection process could either be represented throughout the trial (upper), or participants could be planning during the intertrial interval, as shown in (Konovalov and Krajbich, 2016, 2020).



(a) Distribution

(b) Timeseries

Figure 4: β_{ndt1}. Significant results of pairwise *t*-tests are indicated with asterisks ($* = p < .05, ** = p < .01, *** = p < .001$). **(a)** The distribution of posterior means of β_{ndt1}. SSC-2 yielded significantly lower values than other conditions. **(b)** Timeseries of β_{ndt1} from the sliding-window dynamic & hybrid RLDDM, according to conditions. Thin lines indicate the mean timeseries of participants, which is overlaid by a rolling average of window=10 represented as thick lines. Straight lines indicate least squares regression lines, and shaded areas represent standard errors.

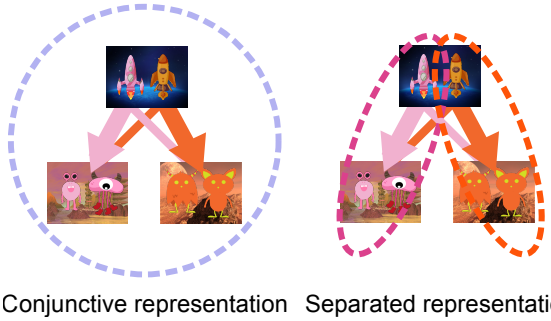
4.35], $p = .001$; SSC-3 vs. SSC-4: mean difference=-2.37, 95% CI [-3.95, -.78], $p = .001$; SSC-3 vs. SSC-5: mean difference=2.23, 95% CI [.66, 3.8], $p = .002$; SSC-4 vs. SSC-5: mean difference=4.6, 95% CI [3.02, 6.18], $p = .001$; Fig. 6a). Also, w_{rep} were significantly smaller than 0.5 (i.e., general tendency to use separated representations) except for SSC-5 (*t*-tests against 0.5: $t_{(99)} = -8.26, p < .001, t_{(104)} = -3.82, p < .001, t_{(102)} = -6.68, p < .001, t_{(106)} = .77, p = .44$ for SSC-2, SSC-3, SSC-4, and SSC-5, respectively).

The timeseries analyses of w_{rep} further revealed more sophisticated dynamics (Fig. 6b): in complex environments, people initially use separated representations but gradually favor conjunctive representations with experience (mean and standard error of the slope of w_{rep}: $-.01$ (.0002), $-.008$ (.0006), .01 (.001), .2 (.001) for SSC-2, SSC-3, SSC-4, and SSC-5, respectively; one-way ANOVA: $F_{(3,412)} = 561.92, p < .001$). The slopes were all significantly increasing or decreasing (*t*-tests against 0: $t_{(99)} = -44.74, p < .001, t_{(104)} = -12.69, p < .001, t_{(102)} = 18.92, p < .001, t_{(106)} = 23.96, p < .001$ for SSC-2, SSC-3, SSC-4, and SSC-5, respectively). Importantly, a qualitative difference was observed between SSC-2,3 vs. SSC-4,5 in that the former steadily decreased in w_{rep} with the number of trials, while the latter showed an increase: slopes significantly increased with SSC (SSC-2 vs. SSC-3: mean difference=.003, 95% CI [.001, .006], $p = .004$; SSC-2 vs. SSC-4: mean difference=.026, 95% CI [.023, .028], $p = .001$; SSC-2 vs. SSC-5: mean difference=.034, 95% CI [.032, .037], $p = .001$; SSC-3 vs. SSC-4: mean difference=.02, 95% CI [.02, .025], $p = .001$; SSC-3 vs. SSC-5: mean difference=.03, 95% CI [.028, .033], $p = .001$; SSC-4 vs. SSC-5: mean difference=.009, 95% CI [.006, .01], $p = .001$).

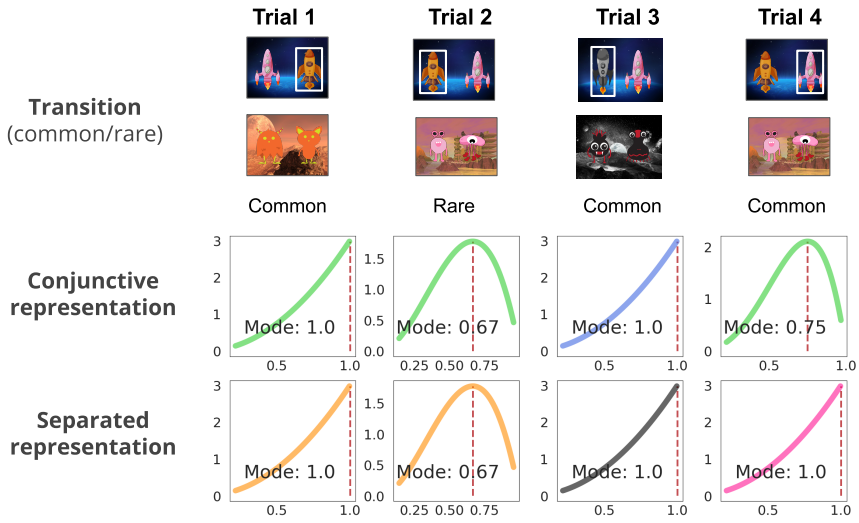
Unlike SSC-2, SSC-3, and SSC-4 where the majority of the subjects predominantly employed separated representations, the subjects in SSC-5 exhibited a more even split when divided into two groups using the threshold value of $w_{rep} = 0.5$ (Fig. 6a). This raises the possibility of more than two distinct groups of subjects exhibiting qualitatively different across-trial dynamics in representation selection. Consequently, we searched for potential differences in the use of representations over time between groups. Despite the highly divergent point-estimate, the timeseries of w_{rep} did not significantly differ between groups, even when the groups were further subdivided into three (Supplementary Figure 3). In summary, despite individual differences in overall representational use, a consistent pattern was observed in complex environments: subjects began with separated representations and transitioned to conjunctive representations with experience. These results suggest that individuals adapt their strategies for selecting representations in response to environmental complexity and experience.

2.4 Closed- vs. open-loop planning in the two-stage task

Two strategies are relevant to planning – closed-loop control, which require adjustments of plans based on changes in the environment, and open-loop control, where sequences of actions pre-defined at the start of planning are executed



(a) Conjunctive and separated representations of transition function. Here, we assume the "common" transition probability would be updated differently for conjunctive and separated representations: while the unit of update is the combination of two choice options in the conjunctive representation, each option is updated separately in the separated representation upon a common transition.

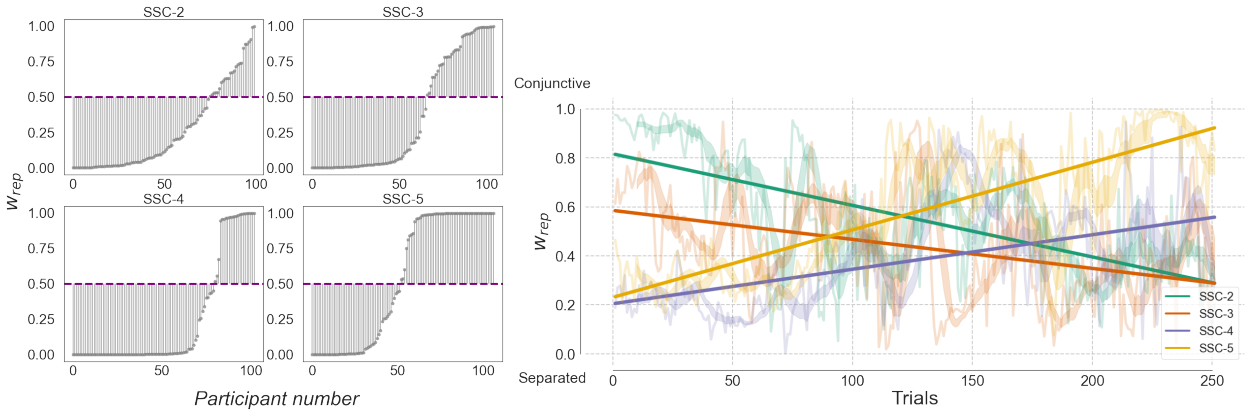


(b) Sample Bayesian updates of transition functions. The example of four trials illustrates that, although an agent experiences the same sequence of events, conjunctive and separated representations are updated differently.

Figure 5: Mixture of representations of the transition function.

(Hansen, Barto, and Zilberstein, 1996). The critical difference is how much attention the subject pays to the external environment during the action; for instance, in open-loop control, the agent implicitly assumes that there is no need to adjust in response to environmental changes. Applied to TST, we defined the rare transitions as the environmental changes that could reflect subjects' use of open vs. closed loop control. People who plan in a closed-loop manner, or subjects who attend more to the environmental changes, will slow down more after rare transitions to adjust their behavior. In contrast, open-loop planners will show less of this effect. Thus, we operationalized the degree of closed- vs. open-loop planning as the "surprise" after an uncommon transition, which was measured as the difference between post-rare and post-common RT in the second-stage decisions (Decker et al., 2016). Incorporating this concept to our dynamic & hybrid RLDDM framework, we investigated the relationship between parameters and surprisal (the degree of closed-loop control). To account for possible systematic differences due to different levels of experience, as we have seen in the previous timeseries analyses, we grouped the trials into "early" (trials 1-150) and "late" (trials 151-300) phases.

We found that **learning rate** (α; early SSC-2: $r = .58, p < .001$, early SSC-3: $r = .52, p < .001$, early SSC-4: $r = .55, p < .001$, early SSC-5: $r = .58, p < .001$; late SSC-2: $r = .68, p < .001$, late SSC-3: $r = .61, p < .001$, late SSC-4: $r = .58, p < .001$, late SSC-5: $r = .6, p < .001$), **first-stage drift rate** (v_1; early SSC-2: $r = .2, p = .046$, early SSC-3: $r = .22, p = .026$, early SSC-4: $r = .25, p = .01$, early SSC-5: $r = .38, p < .001$; late SSC-2:



(a) Distribution ("between-condition" development of w_{rep})

(b) Timeseries ("within-condition" development of w_{rep})

Figure 6: w_{rep}. **(a)** The distribution of posterior means of w_{rep}. w_{rep} generally increased with SSC. **(b)** Timeseries of w_{rep} from the sliding-window dynamic & hybrid RLDDM, according to conditions. Thin lines indicate the mean timeseries of participants, which is overlaid by a rolling average of window=10 represented as thick lines. Straight lines indicate least squares regression lines, and shaded areas represent standard errors.

$r = .25, p = .013$, late SSC-3: $r = .16, p = .1$, late SSC-4: $r = .31, p = .001$, late SSC-5: $r = .41, p < .001$), **second-stage drift rate** (v_2; early SSC-3: $r = .38, p < .001$, early SSC-4: $r = .43, p < .001$, early SSC-5: $r = .42, p < .001$; late SSC-3: $r = .47, p < .001$, late SSC-4: $r = .38, p < .001$, late SSC-5: $r = .37, p < .001$), 1st-stage residual time (early SSC-4: $r = .41, p < .001$, early SSC-5: $r = .51, p < .001$, late SSC-4: $r = .33, p = .001$, late SSC-5: $r = .55, p < .001$), and **representation selection** (β_{ndt}; early SSC-4: $r = .45, p < .001$, early SSC-5: $r = .49, p < .001$, late SSC-4: $r = .41, p < .001$, late SSC-5: $r = .53, p < .001$) were positively correlated with surprisal. On the other hand, parameters that were inversely related to surprisal were **memory decay** (γ: early SSC-3: $r = -.29, p = .003$, early SSC-4: $r = -.22, p = .026$, early SSC-5: $r = -.245, p = .011$, late SSC-3: $r = -.209, p = .033$, late SSC-4: $r = -.18, p = .069$, late SSC-5: $r = -.21, p = .033$) and **first-stage starting point bias** (z_1: early SSC-3: $r = -.42, p < .001$, early SSC-4: $r = -.47, p < .001$, early SSC-5: $r = -.46, p < .001$, late SSC-3: $r = -.39, p < .001$, late SSC-4: $r = -.4, p < .001$, late SSC-5: $r = -.53, p < .001$). In summary, surprisal was correlated with parameters indicative of sensitivity to change within the task, such as learning rate, drift rate (the rate of information accumulation during online deliberation), and representation selection, and negatively correlated with parameters associated with prepotent responses, such as pre-accumulated evidence (first-stage starting-point bias).

Lastly, we searched for any possible macro-level differences of post-common and post-rare RT according to SSC. To do so, we performed a 2×2 ANOVA on second-stage RT, introducing state-space complexity and common vs. rare transition as factors. We found a main effect of common vs. rare transition ($F_{(3,412)} = 139.61, p < .001$), a marginal main effect of state-space complexity ($F_{(3,412)} = 2.33, p > .07$), and no interaction between the two factors was observed ($F_{(3,412)} = .01, p > .9$). These results suggests that post-common and post-rare RT per se does not differ across environmental complexity.

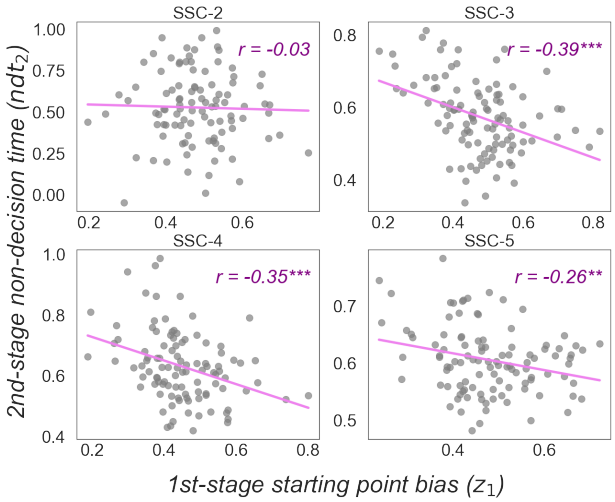
2.5 Representational format affects retrieval efficiency

We have postulated that pre-accumulated evidence for first-stage decisions (z_1) is associated with commitment to the initial plan. Building on this premise, we explored whether this could affect representation loading for *second* stage decisions. The intuition is that in planning, accumulating evidence for either option in the first stage involves the loading of information related to future decisions after the first-stage (Solway and Botvinick, 2015), and the amount of time spent on representation selection could lead to the pre-accumulation of evidence before the diffusion process (Bornstein et al., 2023). We refer to the degree to which second-stage information is already accumulated in the first-stage as *retrieval efficiency*. This concept was supported by a negative correlation between 1st-stage starting point bias (z_1) and second-stage non-decision time (t_2) in the more complex versions of the task (SSC > 2; Fig. 7a). In other words, subjects who have pre-accumulated more evidence towards an option at the first stage also tended to show reduced time

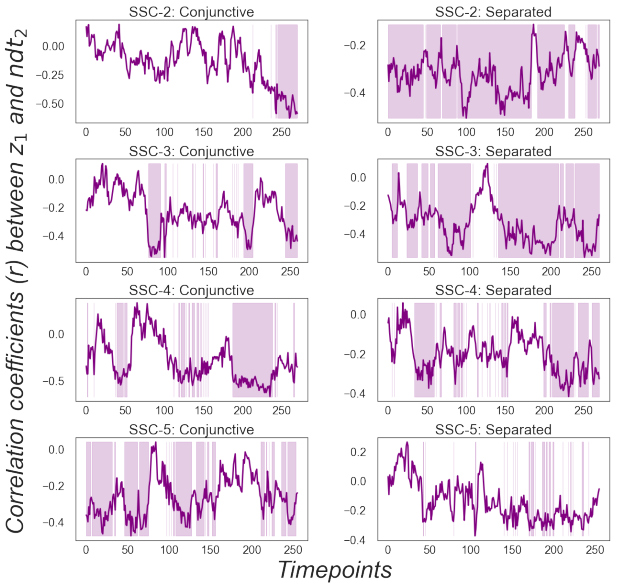
required for loading representations in the second stage. The lack of this effect in SSC-2 reiterates the idea that the loading of representations is not time-locked to the start of the trial in this condition.

Given the highly bimodal distribution of w_{rep} in the most complex state (SSC-5; Fig. 6a), we examined whether the type of representation (transition matrix) used to pre-assess the second-stage options could show differential temporal dynamics of the relationship between z_1 and t_2, or retrieval efficiency (Fig. 7b). To test this, we divided participants into two groups according to their w_{rep} point estimate: conjunctive ($w_{rep} \geq 0.5$) and separated ($w_{rep} < 0.5$). This resulted in 25, 39, 23, and 55 conjunctive participants and 75, 66, 80, and 52 separated participants in SSC-2, SSC-3, SSC-4, and SSC-5, respectively. We then returned to a sliding-window analysis approach to calculate the temporal evolution of the correlation between z_1 and t_2 for each subgroup across all windows (270 windows total).

Our findings revealed that increasing SSC gradually increased the impact of conjunctive representations while diminishing the effect of separated representations on retrieval efficiency. In other words, as SSC increased, more time points were significant for conjunctive representations, and fewer for separated representations. The results again indicate a shift based on SSC; in low complexity conditions (SSC-2 and SSC-3; $n_{sep} \geq n_{conjunctive}$), separated, rather than conjunctive, representations predominantly exhibited a negative correlation between first-stage bias and second-stage non-decision time (Chi-square test: SSC-2: $\chi^2 = 322.37, p < .001$; SSC-3: $\chi^2 = 144.4, p < .001$). In SSC-4, there was no notable difference between conjunctive and separated participants ($\chi^2 = 2.22, p = .14$). This was reversed in SSC-5, where people using conjunctive representations showed a tendency for pre-loading of representations ($\chi^2 = 80.46, p < .001$). These results suggest that considering transitions of individual options enhances the off-loading of second-stage representations onto the first-stage, which is reversed in the most complex state (SSC-5), where conjunctive representations drive this dynamic. Importantly, this pattern remained consistent when only the latter half of the trials were considered (SSC-2: $\chi^2 = 89.5, p < .001$; SSC-3: $\chi^2 = 106.16, p < .001$; SSC-4: $\chi^2 = .06, p = .81$; SSC-5: $\chi^2 = 4.1, p = .04$). This, taken together with the timeseries results of w_{rep}, suggests that people rely on retrieval-efficient representations at the later phase of experience.



(a) The correlation between first-stage starting-point bias (z_1) and second-stage non-decision time (ndt_2) in each condition. ** indicates $p < .01$, and *** indicates $p < .001$.



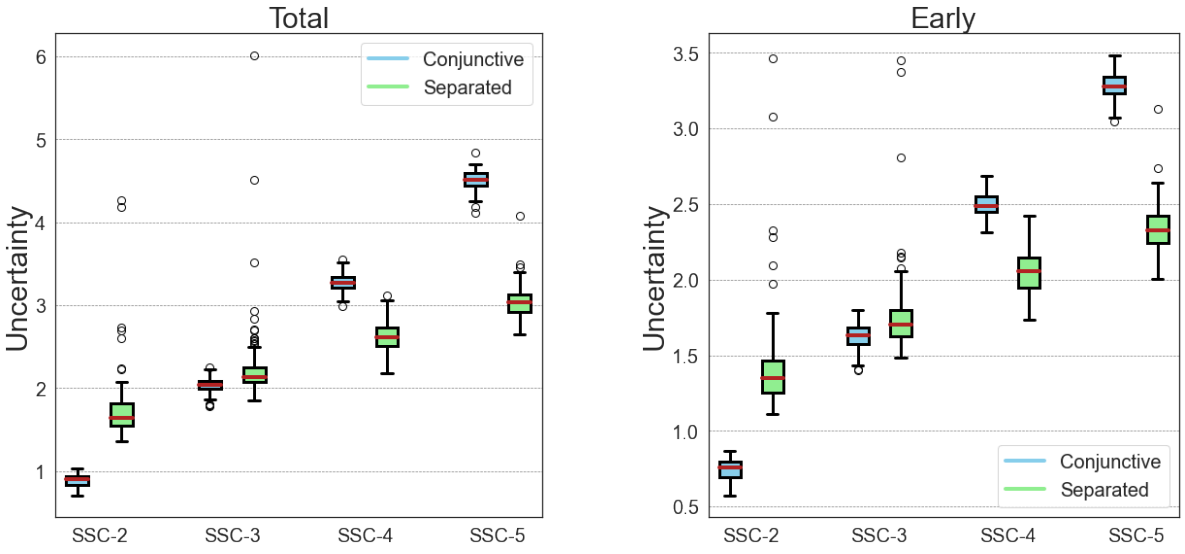
(b) Sliding-window correlation of z_1 and ndt_2, as a function of type of representation used (conjunctive and separated). Shaded areas indicate the window with a significant correlation. Number of windows with significant correlations: SSC-2 Conjunctive=30, SSC-2 Separated=240, SSC-3 Conjunctive=63, SSC-4 Separated=201, SSC-4 Conjunctive=100, SSC-4 Separated=118, SSC-5 Conjunctive=135, SSC-5 Separated=38.

Figure 7: The effect of first-stage evidence pre-accumulation on second-stage representation loading (non-decision time).

2.6 Uncertainty determines the initial choice of representation at early stage of learning

Inspired by a previous study demonstrating arbitration of control according to uncertainty (Lengyel and Dayan, 2007), we investigated how overall uncertainty changes as a function of the representations used. To maximize the differences, we performed two simulations using extremely conjunctive ($w_{rep} = 1$) and separated representations ($w_{rep} = 0$). We obtained the timeseries of uncertainty, where uncertainty was defined as the variance of the transition probability distribution updated trial-by-trial (see *Methods* Section 4.4.3 and Equation 20). For parameters other than subjects' w_{rep}, we used subjects' estimated parameters for simulation. Thus, this simulation shows a "what-if" scenario where actual subjects had extremely high and low values of w_{rep} (i.e., not simulations based on purely synthetic agents).

We calculated the cumulative sum of the timeseries of uncertainty as an estimate of the total amount of uncertainty for each participant in the task. Overall (i.e., for all 300 trials), uncertainty increased with SSC (Conjunctive: $F_{(3,412)} = 26824, p < .001$, mean and standard error of the sum of uncertainty: .89 (.007), 2.04 (.009), 3.28 (.01), 4.51 (.01); separated: $F_{(3,412)} = 253.14, p < .001$, mean and standard error of the sum of uncertainty: 1.75 (.043), 2.26 (.048), 2.62 (.017), 3.05 (.019) for SSC-2, SSC-3, SSC-4, and SSC-5, respectively; Fig. 8a). We found that overall uncertainty explains the actual subjects' initial preference of representations (Fig. 6b). In SSC-2 and SSC-3, where the number of separated states is greater than or equal to the number of conjunctive states, conjunctive representations produced lower uncertainty (paired t-tests: SSC-2: $t_{(99)} = -19.87, p < .001$; SSC-3: $t_{(104)} = -4.68, p < .001$), aligning with subjects' use of conjunctive representations in the early phase of learning. In contrast, for SSC-4 and SSC-5, where the number of conjunctive states exceed the number of separated states, separated representations led to lower uncertainty (paired t-tests: SSC-4: $t_{(102)} = 43.64, p < .001$; SSC-5: $t_{(106)} = 87.92, p < .001$). The assumption that uncertainty accounts for the subjects' initially preferred representation is reinforced by the fact that this pattern holds when confining this analysis to the first half of the trials (Conjunctive: $F_{(3,412)} = 21394.3, p < .001$, mean and standard error of the sum of uncertainty: .74 (.007), 1.62 (.007), 2.5 (.008), 3.28 (.008); separated: $F_{(3,412)} = 260.8, p < .001$, mean and standard error of the sum of uncertainty: 1.42 (.035), 1.77 (.029), 2.05 (.014), 2.34 (.014) for SSC-2, SSC-3, SSC-4, and SSC-5, respectively; Fig. 8b). The early trials also exhibited the pattern such that SSC-2 and SSC-3 yielded lower uncertainty in the conjunctive vs. separated (paired t-tests: SSC-2: $t_{(99)} = -19.77, p < .001$; SSC-3: $t_{(104)} = -5.26, p < .001$), and all other conditions showing lower uncertainty for separated vs. conjunctive (paired t-tests: SSC-4: $t_{(102)} = 35.75, p < .001$; SSC-5: $t_{(106)} = 73.19, p < .001$). Taken together, this provides an explanation for people's favored representation during early learning: people prefer representations that provide more certainty in the early phases of learning, regardless of the environmental complexity.



(a) The distribution of uncertainty for all trials (n=300).

(b) The distribution of uncertainty for the first half of trials ("early").

Figure 8: The distribution of the participant-wise sum of uncertainty, measured by the variance of the beta distribution (see *Methods* Section 4.4.3 and Equation 20).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
SSC-2	21533	21620	21749	20972	21108	21058	20568	**19254**
SSC-3	39911	40035	40122	38345	38490	**38071**	39093	38357
SSC-4	39504	39584	39839	37827	38214	**37577**	38972	38885
SSC-5	47050	47093	47215	44976	44998	**44450**	46912	46802

Table 1: **The DIC values of each model.** Lower values indicate a better model fit to the data. The model with the best fit to the data is boldfaced for each condition. **Model 1**: baseline RLDDM, **Model 2**: RLDDM with dynamic separated transition function, **Model 3**: RLDDM with dynamic conjunctive transition function, **Model 4**: uncertainty-regression RLDDM with dynamic separated transition function, **Model 5**: uncertainty-regression RLDDM with dynamic conjunctive transition function, **Model 6**: uncertainty-regression RLDDM with weighted sum of dynamic separated and conjunctive transition function (dynamic & hybrid RLDDM; our model of interest), **Model 7**: RLDDM with MB vs. MF weighted by *w*, **Model 8**: RLDDM with MB vs. MF weighted by *w* with separate learning rates for stage 1 and 2.

To examine whether the preference for more certain representations is solely due to uncertainty or can also be explained as optimal reward-seeking, we simulated the overall reward obtained with these extreme values of w_{rep}. Specifically, the simulation was performed with participants' empirical parameters, except for the w_{rep} parameters. Each simulation was run 50 times under different random seeds, and the average reward rate of these 50 simulations was used for statistical analyses. Our simulation indicates no difference of performance (reward rate) according to level of w_{rep} for conditions other than SSC-5 (paired t-test between conjunctive vs. separated: all $p > .29$). In SSC-5, however, separated representations produced a slightly higher reward rate than fully conjunctive representations ($t_{(206)} = 2.45, p = .016$). The average reward rates were all significantly above chance (SSC-2 conjunctive: .56 ($t_{(99)} = 17.12, p < .001$); SSC-2 separated: .56 ($t_{(99)} = 17.06, p < .001$); SSC-3 conjunctive: .55 ($t_{(104)} = 17.15, p < .001$); SSC-3 separated: .55 ($t_{(104)} = 17.24, p < .001$); SSC-4 conjunctive: .54 ($t_{(102)} = 13.12, p < .001$); SSC-4 separated: .54 ($t_{(102)} = 13.12, p < .001$); SSC-5 conjunctive: .53 ($t_{(106)} = 14.31, p < .001$); SSC-5 separated: .53 ($t_{(106)} = 14.3, p < .001$)). Taken together, these results suggest that uncertainty minimization is a primary driving principle behind representation selection, though more complex environments may further reward the selection of appropriate representations.

2.7 Model comparison

To verify that our model (dynamic & hybrid RLDDM) provides the most parsimonious description of participant behavior, we compared it with other variants of the RLDDM. We tested five additional models that varied in incorporating representational uncertainty (details under Section 4.4.4). The models varied in how the representations (transition functions) were represented and updated, and whether the representational uncertainty entered as a factor explaining ndt_1. Specifically, the models could differ according to the following dimensions: 1) whether the regression of ndt_1 with regard to representational uncertainty was performed, 2) whether the static and ground-truth representations (e.g., common transition probability of 0.7) or dynamic transition functions updated each trial by a Beta function were used for the transition function, and 3) if the dynamic transition function was used, whether the representation updated was conjunctive, separated, or both (Figure 5a), using the fact that conjunctive vs. separated representations use different updating rules (Figure 5b). In addition to these models, two models that combine the standard dual-system RL model (Daw et al., 2011) with RLDDM were fitted to the data, in order to allow direct comparison of the dual-system (model-based vs. model-free) RL models with our models that only took model-based control into account. Model fits to the data were compared using the Deviance Information Criterion (DIC; Table 1), a standard criterion to compare hierarchical Bayesian models (Spiegelhalter et al., 2002). Comparing DIC values revealed that the dynamic & hybrid RLDDM, which took the weighted sum of the conjunctive and separated representations of the dynamic transition function, best explains the data for all conditions except SSC-2. SSC-2, the original version of the TST, was best explained by the reduced standard model.

2.8 Dual-system RL model: no group-wise difference of MB planning

In addition to the RLDDM, we explored whether our results could be interpreted in light of the models traditionally used for TST (i.e., "dual-system RL model"; Daw et al., 2011). Applying the dual-system RL model could provide an indicator of the general tendency for participants to rely on model-based planning vs. model-free strategies. Building

on models previously used to analyze TST data (Daw et al., 2011; Kool, Cushman, and Gershman, 2016), we selected a model that uses the following eight parameters: learning rate (α), inverse temperature for the first-stage decision (β_1), inverse temperature for the second-stage decision (β_2), eligibility trace decay (λ), model-based planning (w), memory decay (γ), choice stickiness (κ), and response stickiness; see *Methods* Section 4.4.1 for details. Only the distributions of learning rate ($F_{(3,412)} = 4.11, p = .007$), eligibility trace decay ($F_{(3,412)} = 7.47, p < .001$), and memory decay parameters ($F_{(3,412)} = 3.47, p = .02$) were significantly different across SSC (the distributions and the statistics of post-hoc comparisons are in Supplementary Figure 4). The learning rate and eligibility trace parameters were significantly lower for SSC-2 than for other conditions, but differences across SSC-3,4,5 were not observed. Although memory decay resulted in statistical differences across groups, the only group-wise difference was found between SSC-2 and SSC-5. These results suggest that SSC did not yield a systematic difference in model-based vs. model-free strategy usage or degree of planning in general (i.e., ruling out the possibility that subjects in SSC-2 resulted in faster first-stage RT due to using a more model-free strategy).

2.9 Qualitative assessment of the model

Our model (dynamic & hybrid RLDDM) outperforms other models in explaining the data from MTST, but is the model fit qualitatively reliable in itself? To address this concern, our model's qualitative fits were assessed in four ways. First, the Gelman-Rubin statistic ($\hat{R}$) was smaller than 1.1 for all group-level parameters in the dynamic & hybrid RLDDM, indicating that the converged posterior values are reliable for interpret (Gelman and Rubin, 1992). Second, the convergence of Monte-Carlo Markov Chains were confirmed by a visual inspection of the indicators of convergence (Supplementary Figure 5). Third, the joint posterior distribution indicated that there were no spurious correlations between parameters that could alternatively explain the results (Supplementary Figure 6). Fourth, we assessed whether the posterior distributions of the parameter estimates from our model are reliable enough to reproduce the observed data. Specifically, we performed a series of posterior predictive checks where we extracted the subject-level posterior means of the parameters and used them to simulate the RT data. Supplementary Figure 7 shows the comparison between the synthetic and observed RT data, and indicates that the mean of the synthetic data aligns with the mean of the empirical data. It is worth noting two observable patterns. First, the tails in SSC-2 are heavier than the more complex conditions (SSC>2). This is in accordance with the result that our model (dynamic & hybrid RLDDM) is suitable for explaining MTST but not canonical TST data. Second, the second-stage RT simulations yield heavier tails than the first-stage RTs in general, implying that accounting for the representational uncertainty in the first-stage RLDDM effectively captures the behavior.

Taken together, these results support the use of the dynamic, hybrid RLDDM for measuring individual and condition-level differences in choices and response times.

3 Discussion

Multiple task representations coexist and evolve with experience, potentially prompting a transformation in the predominant form of representation at different time points. Despite extensive studies on the arbitration of control, a comprehensive examination of the evolution of the dominant *representations* has been absent in the literature. In this research, we investigated the nature of representations that emerge across experience, under varying levels of state-space complexity. Our novel analyses revealed a component of first-stage RT that is proportional to representational uncertainty, implying a representation selection process being embedded in the first-stage RT. We further dissociated the content of the representations involved in here. We confirmed the previous findings that people simultaneously maintain at least two different kinds of representations – conjunctive and separated (Duncan et al., 2018; Ballard, Wagner, and McClure, 2019). Our study advances preexisting knowledge by showing that environmental complexity modulates the use of these representations *over experience*. In simple environments where the number of conjunctive states is equal to or less than the separated (SSC-2,3), participants initially rely on conjunctive representations, but less so with increasing experience and decreasing uncertainty. In more complex environments where the number of conjunctive states outnumber the separated states (SSC-4,5), an opposite effect is observed such that people start with separated representations but gradually employ a conjunctive representation.

We account for the gradual shift in the preferred form of representation based on two objectives: uncertainty minimization and retrieval efficiency. In the early phase of learning, regardless of state-space complexity, subjects select

representations that minimize the overall uncertainty. For SSC-2 and SSC-3, where the number of separated states exceeds the number of conjunctive states, conjunctive representations are favored at first. For SSC-4 and SSC-5, which have higher complexity, separated representations are initially preferred. In the later phase of the experiment, representations that enhance retrieval efficiency – defined as the extent to which first-stage starting-point bias (z_1) reduces the second-stage non-decision time (ndt_2) – become favored: separated representations for relatively simple environments (SSC-2 and SSC-3), and conjunctive representations for relatively complex environments (SSC-4 and SSC-5). This indicates that, initially, people opt for representations that minimize immediate uncertainty, but as the other kind of representation (with a greater number of states) gains certainty, they switch to those.

Our explanation that uncertainty governs the form of representation chosen at the early phase of learning is in line with previous literature. Uncertainty-weighted arbitration of control across time has been predicted through simulations (Daw, Niv, and Dayan, 2005). Model-based and model-free systems alternate across time according to their relative uncertainty. Another simulation study adds a third mode – episodic control – to this continuum, where episodic control is favored in the early stages of learning before sufficient experience has been collected to the extent that the models are reliably constructed (Lengyel and Dayan, 2007). These studies indicate that agents' trade-off of uncertainty and accuracy of control modes explains the continuum of favored modes of control across experience. Two aspects that have been uninterrogated are the empirical demonstration of uncertainty-weighted arbitration over time, and whether the dominant form of *representation* also arbitrates over time, given that representations are shaped with experience. We address these gaps by connecting the idea of uncertainty-weighted arbitration to empirical findings regarding coexistence of and selection among multiple task representations when people learn (Duncan et al., 2018; Ballard, Wagner, and McClure, 2019; Wang, Feng, and Bornstein, 2022; Luettgau et al., 2023; Correa et al., 2023). Drawing from the idea that multiple representations could be maintained at the same time, we showed that people first rely on more certain representations, but eventually turn to representations that are more efficient, after these have been refined with experience.

Unlike uncertainty-weighted arbitration, which explains the first half of representation learning in our experiment, retrieval efficiency is a relatively new concept that we introduce. This foreshadows the results of the simulation study that have shown that after episodic control and model-based control, agents eventually settle down to habitual control, which is more efficient (Lengyel and Dayan, 2007). This suggests that once uncertainty is resolved, agents shift their focus to efficiency. Our finding provides a representational counterpart of this continuum. This aligns with recent studies showing that model-based control comprises both a representation selection (task construal) stage and a planning stage, which relaxes the previous assumption that agents plan with a fixed representation of the environment (Ho et al., 2022; Correa et al., 2023). In this framework, an additional objective for resource-limited agents is to construct a parsimonious (efficient) representation of the world, so that the utility of planning with the representation is maximized, but, at the same time, the complexity of the representation is minimized.

Taking this work a step further, we for the first time decompose the portion of time dedicated for resolving representational uncertainty, by utilizing another dimension of behavior – response time, both within- and across-trials. This represents an important step towards field or clinical applications which will require small numbers of trials (Donegan et al., 2023) and thus may benefit from leveraging response times which can improve quality of model fits (Ballard and McClure, 2019; Fontanesi et al., 2019) and reveal important dynamics with potential relevance to clinical applications (Bornstein and Pickard, 2020; Banavar et al., 2024; Chwiesko et al., 2023; Copeland, Stafford, and Field, 2024). Our discussion on open-loop vs. closed-loop planning is also a novel contribution to the two-stage task paradigm. The parameters that correlate with the concept of closed-loop planning (learning rates, representation selection time, etc) introduce a new perspective of interpreting RLDDM parameters in TST.

In summary, although the idea of using multiple representations for planning has been proposed before, we show here for the first time how they are temporally entangled in a single trial of decision making, and what kind of representations are developed across experience in different levels of environmental complexity. These findings establish a useful tool for future work investigating the neural substrates that give rise to the temporal dynamics of planning in complex environments.

4 Methods

4.1 Materials and participants

We recruited a total of 448 participants between age 18-40 - 112, 108, 108, and 110 participants for 2-, 3-, 4-, and 5-SSC version of MTST tasks, respectively - via Amazon Mechanical Turk. All versions of the experiment were approved by the Ethics Committee of the University of California, Irvine. All participants were compensated with a base payment of $8 and an additional reward that ranged between $4-$8, depending on their performance of four randomly drawn trials.

We excluded 8, 4, 5, and 5 subjects from each experiment who met at least one of the following criteria: responding with the same key on more than 95% of the trials, responding implausibly fast (RT below 150 ms) on more than 10% of the trials (Shahar et al., 2019), choosing more than 90% to either option, failing to respond within the response window on more than 20% of the trials, pressing a certain button or choosing a certain option consecutively for more than 10% of the trials, responding with RT of 0 for more than 5 trials (this indicates that the button was pressed before the onset of the trial), and scored less than 2 out of 5 catch trials.

4.2 Task design

In order to encourage planning to begin at the presentation of the first-stage options, we increased the number of possible first-stage options (or second-stage states) from 2 (canonical TST) to 3, 4, and 5 (Figure 1). These different conditions are referred to as SSC-2, SSC-3, SSC-4, and SSC-5, and the study was conducted as a between-participant study such that one participant experienced 300 trials of one condition. In the canonical TST, a single decision tree (Figure 1a) being used throughout the experiment does not necessitate planning to occur only after the first-stage stimuli onset. However, for k-SSC TST ($k > 2$), one decision tree is randomly constructed from a set size of $N = \binom{k}{2}$ possible trees for each trial, thereby making it difficult to select a plan without observing the first-stage options.

Other than the set size of decision trees, the procedural details are the same across all versions of TST (Figure 1b). Participants are stochastically led to the second stage based on their first-stage choice, where the first-stage choice (spaceship) leads to the second-stage state (planet) of the same color on 70% of trials (i.e., common transition) but results in the planet of the alternative choice for 30% of the trials (i.e., rare transition). The reward probability of the second stage decision drifted according to a zero-centered Gaussian random walk, with reflecting bounds of [.25 .75]. The standard deviation of the Gaussian noise added at each step was .025, allowing for gradual and stochastic variations in the reward probability over time. The initial reward probability of each planet was set to 60% and 40% or 75% and 25%, with planets randomly assigned to these probabilities in equal proportions. The reward was either 1 (represented by a graphic of glittering space treasure) or 0 (represented by an empty circle), determined by the reward probability on the current trial (Figure 1c). After the reward feedback, an inter-trial interval (ITI) of 1s was represented by a fixation cross in the middle of the screen.

4.3 Experimental procedure

Prior to performing the main TST, participants practiced 10 trials each of first- and second-stage decision making, followed by 10 trials of the full canonical TST. After training, they took a comprehension quiz that asked (1) the total number of spaceships/aliens in the experiment, (2) whether the spaceship of a certain color always led to the same colored planet (stochasticity of transition), and (3) whether the reward probability drifted over time. The participants could proceed to the main experiment only after answering correctly all questions to ensure that every subject had a correct model of the environment to begin with, with an intention to reduce the variance of meta-knowledge of the model as much as possible.

Each trial was aborted if participants failed to respond within the 2-second window for either stage. The sequence and timing of events within a trial is described in Figure 1c. Our main task employed 300 trials to allow for sufficient number of trials per state, particularly for experiments with higher SSC. Throughout the experiment, five catch trials that asked the participants to press a certain letter on the keyboard ("Please press the letter Z on your keyboard") appeared intermittently in order to exclude inattentive participants' data from analyses.

4.4 Computational modeling

4.4.1 Dual-process RL

We used a MATLAB implementation of the dual-process RL model based on that provided by Kool, Cushman, and Gershman (2016). Prior to investigating the temporal dynamics of planning, we performed a manipulation check to verify that the participants were using at least some degree of model-based planning and not relying solely on model-free tactics. To do so, the dual-system RL model was used to capture the general tendency to plan, where parameter *w* directly weighs the tendency to use the model-based (MB) system against the model-free (MF) system (Daw, Niv, and Dayan, 2005; Daw et al., 2011; Gläscher et al., 2010). Both systems estimate the state-action value, $Q(s, a)$, that represents how much value the current state-action pair carries considering the future discounted return. The MF system uses temporal-difference (TD) learning to update cached values, acquired through direct experience, for each encountered state-action pair. To elaborate, for a given trial *t*, the reward prediction error (RPE) of the first-stage decision ($\delta_{t,1}$) can be represented as:

$$\delta_{t,1} = Q_{MF}(s_2, a_2) - Q_{MF}(s_1, a_1) \tag{2}$$

, where s_i and a_i represent states and actions in the ith stage ($i \in \{1, 2\}$). $\delta_{t,1}$ is then used to update the MF Q-value of the first-stage state-action pair, weighted by the learning rate parameter $\alpha \in [0, 1]$:

$$Q_{MF}(s_1, a_1) = Q_{MF}(s_1, a_1) + \alpha\delta_{t,1} \tag{3}$$

The RPE upon receiving feedback (r_t) after the second-stage decision is calculated like the following:

$$\delta_{t,2} = r_t - Q_{MF}(s_2, a_2) \tag{4}$$

and this second-stage RPE is used to update the MF Q-value of the second-stage state-action pair:

$$Q_{MF}(s_2, a_2) = Q_{MF}(s_2, a_2) + \alpha\delta_{t,2} \tag{5}$$

.

TD learning is achieved by the eligibility trace parameter $\lambda \in [0, 1]$ that governs the extent to which the updated Q-value for the second-stage state-action pair would be backed up to the first-stage state-action pair:

$$Q_{MF}(s_1, a_1) = Q_{MF}(s_1, a_1) + \lambda\alpha\delta_{t,2} \tag{6}$$

Finally, in addition to the parameters in the original dual-system RL model, we introduce a memory decay factor γ to take into account that unseen state-action pairs in a given trial may be devalued over time (Ito and Doya, 2009). We assumed that for each trial, $Q_{MF}(s, a)$ that are not experienced will be discounted by the factor of $\gamma \in [0, 1]$.

The gist of the MB system lies in utilizing the transition probability and second-stage Q-values to select first-stage actions. The MB Q-value of the first stage is derived using the Bellman equation (Sutton and Barto, 2018), such that

$$Q_{MB}(s_1, a_1) = \sum_{s_2} p(s_2|s_1, a_1) \times max(Q_{MF}(s_2)) \tag{7}$$

Here, $p(s_2|s_1, a_1)$ corresponds to the common or rare transition probability, and $max(Q_{MF}(s_2))$ corresponds to the larger MF Q-value in the successor state. To derive the net Q-value, $Q_{MB}(s_1, a_1)$ is multiplied by the planning parameter w, and then added to the MF Q-value multiplied with $1 - w$:

$$Q(s_1, a_1) = wQ_{MB}(s_1, a_1) + (1 - w)Q_{MF}(s_1, a_1) \tag{8}$$

For both stages, the softmax function determines the choice probability of each action, with the net Q-value derived from Equation 8 as the first-stage input and $Q(s_2, a_2)$ as the second-stage input, respectively:

$$P(a_i = a|s_i) = \frac{exp(\beta Q(s_i, a))}{\sum_{a\prime} exp(\beta Q(s_i, a\prime))} \tag{9}$$

.

Here, the inverse parameter $\beta \in [0, 20]$ dictates the greediness of the choice and $a\prime$ represents all possible actions in state s_i.

4.4.2 Baseline RLDDM

To explore possible computations within the first-stage RT that differ by SSC, we constructed our model based on the reinforcement learning diffusion decision model (RLDDM; (Pedersen and Frank, 2020; Fontanesi et al., 2019)) to inspect both temporal and choice dimensions of our data. We employed the Hierarchical Drift-Diffusion Model (HDDM) software for our RLDDM analyses, which enables the estimation of individual and group-level parameters simultaneously (Wiecki, Sofer, and Frank, 2013; Pedersen and Frank, 2020). We report the individual-level parameter estimates in the results. For Monte Carlo Markov Chain (MCMC) sampling, we utilized 3 chains, each consisting of 20,000 samples, with a burn-in period of 12,500 samples. Our models inherited RLDDM by combining parameters that capture drift diffusion processes – drift rates (v_1, v_2), non-decision times (ndt_1, ndt_2), and starting-point bias (z_1, z_2) respectively for each stage (1st- and 2nd stage) – and a learning-rate parameter, α, to capture the learning process in RL. To adapt to our new task design, we additionally incorporated a parameter (γ) that accounts for the decay of values for unseen state-action pairs. How these parameters were used with respect to diffusion-decision or learning process in our model is explicated below, followed by how we integrated additional components to introduce our proposed model.

Reinforcement learning To estimate the Q-values of the first-stage options, we use the Bellman equation (Sutton and Barto, 2018) under the Markov decision process (MDP) framework. Decisions in the first-stage choice require forward planning. In other words, participants use the model, consisting of transition and reward function, to estimate the value of the current choice in light of the future consequences. The state-action value function of the first-stage option, $Q(a, s)$ is defined as the optimal Bellman equation, which assumes that the agent will follow the optimal policy π^* after state and action s, a:

$$Q^*(s, a) = \sum_{s\prime} p(s\prime|a, s) \times max_{a\prime} Q^*(s\prime, a\prime) \tag{10}$$

. The choice rule between two choices in the first and second stage follows a diffusion decision modeling approach discussed below in **Diffusion decision modeling**. Upon receiving a reward after the second-stage choice, the second-stage state-action value (Q-value) is updated by the learning rate α:

$$Q(s, a) = Q(s, a) + \alpha \times (r - Q(s, a)) \tag{11}$$

. Also, the value of the state-value pairs in the second stage decays with a factor of $(1 - \gamma)$ if it is unseen (Ito and Doya, 2009):

$$Q(s, a) = Q(s, a) \times (1 - \gamma) \tag{12}$$

Diffusion decision modeling For all models of interest, the choice rule for both first-stage and second-stage decisions is modeled as following a Wiener first-passage time distribution with three effective parameters (Navarro and Fuss, 2009):

$$rt_i \sim wfpt(a_i, v_i, ndt_i, z_i) \tag{13}$$

, where $i \in [1, 2]$ represents the stage of respective parameters. In our study, the decision threshold (a) was fixed at 1. The drift rate is denoted by v, the non-decision time by ndt, and the starting-point bias by z.

4.4.3 Dynamic and hybrid RLDDM

The dynamic & hybrid RLDDM incorporates several additional components to the baseline RLDDM. First, a parameter that links the uncertainty of representations to the first-stage non-decision time – (β_{ndt1}) – is added (see ***Estimation of***

representational uncertainty below). Also, we dynamically model the transition function as a function of experience, using a Beta distribution, given that the choices are binary options (see Section ***Dynamic transition function*** below). A parameter that weighs the relative contribution of conjunctive vs. separated representation in estimating the source of representational uncertainty (w_{rep}) is added too, which is elaborated in the ***Representational units: conjunctive and/or separated*** section.

Dynamic transition function Instead of hard-coding a fixed, ground-truth probability for the transition function ($P(s\prime|s,a)$), we postulated a dynamic and subjective transition function that is updated as a function of subjects' observations. Specifically, the transition function is represented as a transition matrix (TM) with common (s_c) and rare (s_r) transition as its elements:

$$TM_i = \begin{bmatrix} P(s_c|s_i,a_1) & P(s_r|s_i,a_2) \\ P(s_c|s_i,a_2) & P(s_r|s_i,a_1) \end{bmatrix} \tag{14}$$

, where the transition function TM_i is multiplied to second-stage Q-values to derive the action score of the first-stage options:

$$Q_{MB}(s,a) = TM_i \times \begin{bmatrix} Q^*(s_1,a) \\ Q^*(s_2,a) \end{bmatrix} \tag{15}$$

.

Common transition probability ($P(s_c|s,a)$) stands for the probability of transitioning into a "common" successor state following a state-action pair (i.e., in a first-stage state where yellow and purple spaceships are presented, the common transition of choosing a yellow spaceship would be landing on a yellow planet). Likewise, the rare transition probability corresponds to the probability of a state-action pair leading to an uncommon transition (i.e., choosing a yellow spaceship leading to the purple planet).

We assume that participants' subjective estimate of the common transition probability at a given trial t is a continuous random variable $[0,1]$ that follows a Beta distribution:

$$P_t(s_c|s_i) \sim Beta(\alpha_i, \beta_i) \tag{16}$$

,

with each representational unit (first-stage state) $i \in \mathbb{S}$, where $\mathbb{S}$ is a set of states.

The shape parameters $\alpha > 0$ and $\beta > 0$ stand for the number of common and rare transitions observed at each first-stage state, respectively, and gets updated each trial: if a common transition is observed, $\alpha \leftarrow \alpha + 1$, and $\beta \leftarrow \beta + 1$ otherwise. The mode of the Beta distribution ($Mode(\alpha,\beta)$) is used as a point estimate of common transition probability for each first-stage state ($P_t(s_c|s_i)$):

$$Mode(\alpha,\beta) = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{17}$$

when $\alpha > 1$ and $\beta > 1$. Thus, the transition function at first-stage state combination of (i,j) at trial t is derived by

$$TM_{t,i,j} = \begin{bmatrix} Mode(\alpha_{t,i},\beta_{t,i}) & 1 - Mode(\alpha_{t,i},\beta_{t,i}) \\ 1 - Mode(\alpha_{t,j},\beta_{t,j}) & Mode(\alpha_{t,j},\beta_{t,j}) \end{bmatrix} \tag{18}$$

, where $\alpha_{t,i}$ and $\beta_{t,i}$ represents the number of common and rare transition observed for first-stage state i up to trial t. Since common and rare transitions are mutually exclusive events, we set the rare transition probability as $P(s_r|s_i) = 1 - P(s_c|s_i)$.

Estimation of representational uncertainty For models that regress ndt_1 as a function of representational uncertainty, ndt_1 is defined as a combination of the product of uncertainty (U_{rep}) weighted by the regressor (β_{ndt1}) and the residual (t_{res}):

$$ndt_1 = \beta_{ndt1} \times U_{rep} + t_{res} \tag{19}$$

.

As described above, we model the participants' subjective estimate of the common transition probability ($P(s_c|s_i)$) via a Beta posterior update function. Here, the point estimate of the common transitional probability for each first-stage state could be defined as the mode of the Beta distribution, and the representational uncertainty (U_{rep}) could be estimated via the variance of the Beta distribution ($Var()$):

$$U_{rep} = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \tag{20}$$

.

Representational units: conjunctive and/or separated Acknowledging previous research that has shown people use conjunctive vs. feature-based representations during learning (Kikumoto and Mayr, 2020; Ballard, Wagner, and McClure, 2019), we modeled parallel updates of two Beta distributions, reflecting the conjunctive and separated transition structures (Fig. 5b).

For illustration, when a subject chooses option j upon seeing a combination of options (j, k) and faces a common transition, the Beta update rule for the separated representation follows:

$$\alpha_j \leftarrow \alpha_j + 1 \tag{21}$$

,

while the conjunctive representation is updated by the following:

$$\alpha_{j,k} \leftarrow \alpha_{j,k} + 1 \tag{22}$$

. The variance of the separated representation is derived by

$$var(separated) = \frac{(\alpha_j + y)(\beta_j + n_j - y)}{(\alpha_j + \beta_j + n)^2(\alpha_j + \beta_j + n_j + 1)} \tag{23}$$

,

and the variance of the conjunctive representation is obtained through

$$var(conjunctive) = \frac{(\alpha_{j,k} + y)(\beta_{j,k} + n_{j,k} - y)}{(\alpha_{j,k} + \beta_{j,k} + n)^2(\alpha_{j,k} + \beta_{j,k} + n_{j,k} + 1)} \tag{24}$$

. The parameter $w_{rep} \in [0, 1]$ links these two representations to derive the net representational uncertainty, such that

$$U_{rep} = w_{rep} \times var(conjunctive) + (1 - w_{rep}) \times var(separated) \tag{25}$$

. In other words, high w_{rep} indicates that a subject is relying more on conjunctive rather than separated representations.

4.4.4 Model specification

We defined eight models according to different combinations of the transition function (static vs. dynamic), whether uncertainty was regressed from ndt_1, and the representational format of the transition function (conjunctive and/or separated). The models we compared are summarized in Table 1 as well as the following: **Model 1**: baseline RLDDM as described in Section 4.4.2. **Model 2**: Dynamic RLDDM with separated representations: uses trial-by-trial beta updates for the transition function (as described in Section ***Dynamic transition function***), but only assumes separated representations. This model does not regress the representational uncertainty from ndt_1. **Model 3**: Dynamic RLDDM with conjunctive representations: uses trial-by-trial beta updates for the transition function (as described in Section ***Dynamic transition function***), but only assumes conjunctive representations. This model does not regress the representational uncertainty from ndt_1. **Model 4**: Dynamic RLDDM with separated representations, which also

regresses representational uncertainty from ndt_1. **Model 5**: Dynamic RLDDM with conjunctive representations, which also regresses representational uncertainty from ndt_1. **Model 6**: Dynamic and hybrid RLDDM, as described in Section 4.4.3. This is our model of interest. **Model 7**: An incorporation of the dual-system RL model's w parameter into the baseline RLDDM, such that both the MB and MF values are estimated to obtain net Q-values. **Model 8**: An extension of Model 7 such that separate learning rates are used for stage 1 and 2.

5 Conclusion

Despite extensive work on the arbitration of control in model-based reinforcement learning, less work has investigated how people arbitrate between representations for efficient planning. We introduce a novel experiment that increases the complexity of the canonical two-stage task, creating an environment suitable for observing the temporal dynamics of representation use across varying levels of complexity, both within and across trials. We found that, regardless of environmental complexity, people first employ representations that offer more certainty. However, as repeated experience reduces uncertainty, they transition to kinds of representations that foster efficiency. This study empirically demonstrates, for the first time, that individuals dynamically arbitrate between representations to meet different objectives, a phenomenon previously shown only through simulations.

6 Acknowledgements

Funding was provided by NIA R21AG072673 and NINDS R01NS119468 (PI: ER Chrastil) to AMB. We thank Kevin Miller, Joachim Vandekerckhove, and Frederick Callaway for engaging in insightful discussions.

7 Citation diversity statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field (Mitchell, Lange, and Brus, 2013; Dion, Sumner, and Mitchell, 2018; Caplar, Tacchella, and Birrer, 2017; Maliniak, Powers, and Walter, 2013; Dworkin et al., 2020; Bertolero et al., 2020; Wang et al., 2021; Chatterjee and Werner, 2021; Fulvio, Akinnola, and Postle, 2021). Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020; Zhou et al., 2020). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 12.01% woman(first)/woman(last), 22.58% man/woman, 22.44% woman/man, and 42.97% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (Ambekar et al., 2009; Sood and Laohaprapanon, 2018). By this measure (and excluding self-citations), our references contain 7.28% author of color (first)/author of color(last), 15.20% white author/author of color, 14.79% author of color/white author, and 62.73% white author/white author. This method is limited in that a) names and Florida Voter Data to make the predictions may not be indicative of racial/ethnic identity, and b) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.

References

Ambekar, Anurag et al. (2009). "Name-ethnicity classification from open sources". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 49–58.

Ballard, Ian C and Samuel M McClure (2019). "Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models". In: *Journal of Neuroscience Methods* 317, pp. 37–44.

Ballard, Ian C, Anthony D Wagner, and Samuel M McClure (2019). "Hippocampal pattern separation supports reinforcement learning". In: *Nature Communications* 10.1, p. 1073.

Banavar, Nidhi V et al. (2024). "A response time model of the three-choice Mnemonic Similarity Task provides stable, mechanistically interpretable individual-difference measures." In.

Bertolero, Maxwell A. et al. (2020). "Racial and ethnic imbalance in neuroscience reference lists and intersections with gender". In: *bioRxiv*.

Bornstein, Aaron M and Hanna Pickard (2020). ""Chasing the first high": memory sampling in drug choice". In: *Neuropsychopharmacology* 45.6, pp. 907–915.

Bornstein, Aaron M et al. (2017). "Reminders of past choices bias decisions for reward in humans". In: *Nature Communications* 8.1, pp. 1–9.

Bornstein, Aaron M et al. (2023). "Associative memory retrieval modulates upcoming perceptual decisions". In: *Cognitive, Affective, & Behavioral Neuroscience* 23.3, pp. 645–665.

Caplar, Neven, Sandro Tacchella, and Simon Birrer (2017). "Quantitative evaluation of gender bias in astronomical publications from citation counts". In: *Nature Astronomy* 1.6, p. 0141.

Chatterjee, Paula and Rachel M Werner (2021). "Gender Disparity in Citations in High-Impact Journal Articles". In: *JAMA Netw Open* 4.7, e2114509.

Chwiesko, Caroline et al. (2023). "Parsing memory and nonmemory contributions to age-related declines in mnemonic discrimination performance: a hierarchical Bayesian diffusion decision modeling approach". In: *Learning & Memory* 30.11, pp. 296–309.

Copeland, Amber, Tom Stafford, and Matt Field (2024). "The influence of alcohol-specific episodic memory and cue exposure on value-based decision-making and its role in ad libitum drinking". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46.

Correa, Carlos G et al. (2023). "Humans decompose tasks by trading off utility and computational cost". In: *PLOS Computational Biology* 19.6, e1011087.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature neuroscience* 8.12, pp. 1704–1711.

Daw, Nathaniel D et al. (2011). "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6, pp. 1204–1215.

Decker, Johannes H et al. (2016). "From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning". In: *Psychological science* 27.6, pp. 848–858.

Dion, Michelle L, Jane Lawrence Sumner, and Sara McLaughlin Mitchell (2018). "Gendered citation patterns across political science and social science methodology fields". In: *Political Analysis* 26.3, pp. 312–327.

Donegan, Kelly R et al. (2023). "Using smartphones to optimise and scale-up the assessment of model-based planning". In: *Communications Psychology* 1.1, p. 31.

Duncan, Katherine et al. (2018). "More than the sum of its parts: a role for the hippocampus in configural reinforcement learning". In: *Neuron* 98.3, pp. 645–657.

Dworkin, Jordan D. et al. (2020). "The extent and drivers of gender imbalance in neuroscience reference lists". In: *bioRxiv*. DOI: 10.1101/2020.01.03.894378. eprint: https://www.biorxiv.org/content/early/2020/01/11/2020.01.03.894378.full.pdf.

Fontanesi, Laura et al. (2019). "A reinforcement learning diffusion decision model for value-based decisions". In: *Psychonomic bulletin & review* 26.4, pp. 1099–1121.

Fulvio, Jacqueline M, Ileri Akinnola, and Bradley R Postle (2021). "Gender (Im)balance in Citation Practices in Cognitive Neuroscience". In: *J Cogn Neurosci* 33.1, pp. 3–7.

Gelman, Andrew and Donald B Rubin (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical science*, pp. 457–472.

Gillan, Claire M et al. (2016). "Characterizing a psychiatric symptom dimension related to deficits in goal-directed control". In: *elife* 5, e11305.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Hansen, Eric, Andrew Barto, and Shlomo Zilberstein (1996). "Reinforcement learning for mixed open-loop and closed-loop control". In: *Advances in Neural Information Processing Systems* 9.

Ho, Mark K et al. (2022). "People construct simplified mental representations to plan". In: *Nature* 606.7912, pp. 129–136.

Hunter, Lindsay E, Aaron M Bornstein, and Catherine A Hartley (2018). "A common deliberative process underlies model-based planning and patient intertemporal choice". In: *bioRxiv*, p. 499707.

Ito, Makoto and Kenji Doya (2009). "Validation of decision-making models and analysis of decision variables in the rat basal ganglia". In: *Journal of Neuroscience* 29.31, pp. 9861–9874.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray (2011). "Speed/accuracy trade-off between the habitual and the goal-directed processes". In: *PLoS computational biology* 7.5, e1002055.

Khoudary, Ari, Megan AK Peters, and Aaron M Bornstein (2022). "Precision-weighted evidence integration predicts time-varying influence of memory on perceptual decisions". In: *Cognitive Computational Neuroscience*.

Kikumoto, Atsushi and Ulrich Mayr (2020). "Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection". In: *Proceedings of the National Academy of Sciences* 117.19, pp. 10603–10608.

Kim, Dongjae et al. (2019). "Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning". In: *Nature communications* 10.1, p. 5738.

Konovalov, Arkady and Ian Krajbich (2016). "Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning". In: *Nature communications* 7.1, pp. 1–11.

— (2020). "Mouse tracking reveals structure knowledge in the absence of model-based choice". In: *Nature communications* 11.1, pp. 1–9.

Kool, Wouter, Fiery A Cushman, and Samuel J Gershman (2016). "When does model-based control pay off?" In: *PLoS computational biology* 12.8, e1005090.

Kraemer, Peter M and Sebastian Gluth (2023). "Episodic memory retrieval affects the onset and dynamics of evidence accumulation during value-based decisions". In: *Journal of Cognitive Neuroscience* 35.4, pp. 692–714.

Lee, Sang Wan, Shinsuke Shimojo, and John P O'Doherty (2014). "Neural computations underlying arbitration between model-based and model-free learning". In: *Neuron* 81.3, pp. 687–699.

Lengyel, Máté and Peter Dayan (2007). "Hippocampal contributions to control: the third way". In: *Advances in neural information processing systems* 20.

Luettgau, Lennart et al. (2023). "Decomposing dynamical subprocesses for compositional generalization". In.

Maliniak, Daniel, Ryan Powers, and Barbara F Walter (2013). "The gender citation gap in international relations". In: *International Organization* 67.4, pp. 889–922.

Milli, Smitha, Falk Lieder, and Thomas L Griffiths (2021). "A rational reinterpretation of dual-process theories". In: *Cognition* 217, p. 104881.

Mitchell, Sara McLaughlin, Samantha Lange, and Holly Brus (2013). "Gendered citation patterns in international relations journals". In: *International Studies Perspectives* 14.4, pp. 485–492.

Navarro, Daniel J and Ian G Fuss (2009). "Fast and accurate calculations for first-passage times in Wiener diffusion models". In: *Journal of mathematical psychology* 53.4, pp. 222–230.

Niv, Yael et al. (2015). "Reinforcement learning in multidimensional environments relies on attention mechanisms". In: *Journal of Neuroscience* 35.21, pp. 8145–8157.

Nunez, Michael D et al. (2019). "The latency of a visual evoked potential tracks the onset of decision making". In: *Neuroimage* 197, pp. 93–108.

Otto, A Ross et al. (2013a). "The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive". In: *Psychological science* 24.5, pp. 751–761.

Otto, A Ross et al. (2013b). "Working-memory capacity protects model-based learning from stress". In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20941–20946.

Park, Heyeon, Daeyeol Lee, and Jeanyung Chey (2017). "Stress enhances model-free reinforcement learning only after negative outcome". In: *PLoS One* 12.7, e0180588.

Pedersen, Mads L and Michael J Frank (2020). "Simultaneous hierarchical bayesian parameter estimation for reinforcement learning and drift diffusion models: a tutorial and links to neural data". In: *Computational Brain & Behavior* 3.4, pp. 458–471.

Shahar, Nitzan et al. (2019). "Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling". In: *PLoS computational biology* 15.2, e1006803.

Solway, Alec and Matthew M Botvinick (2015). "Evidence integration in model-based tree search". In: *Proceedings of the National Academy of Sciences* 112.37, pp. 11708–11713.

Sood, Gaurav and Suriyan Laohaprapanon (2018). "Predicting race and ethnicity from the sequence of characters in a name". In: *arXiv preprint arXiv:1805.02109*.

Spiegelhalter, David J et al. (2002). "Bayesian measures of model complexity and fit". In: *Journal of the royal statistical society: Series b (statistical methodology)* 64.4, pp. 583–639.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Vikbladh, Oliver, Daphna Shohamy, and Nathaniel Daw (2017). "Episodic contributions to model-based reinforcement learning". In: *Annual conference on cognitive computational neuroscience, CCN*.

Vikbladh, Oliver M et al. (2019). "Hippocampal contributions to model-based planning and spatial memory". In: *Neuron* 102.3, pp. 683–693.

Wang, Shaoming, Samuel F Feng, and Aaron M Bornstein (2022). "Mixing memory and desire: How memory reactivation supports deliberative decision-making". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 13.2, e1581.

Wang, Xinyi et al. (2021). "Gendered citation practices in the field of communication". In: *Annals of the International Communication Association*. DOI: 10.1080/23808985.2021.1960180.

Wiecki, Thomas V, Imri Sofer, and Michael J Frank (2013). "HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python". In: *Frontiers in neuroinformatics*, p. 14.

Wyckmans, Florent et al. (2022). "The modulation of acute stress on model-free and model-based reinforcement learning in gambling disorder". In: *Journal of behavioral addictions* 11.3, pp. 831–844.

Yoo, Jungsun, Elizabeth Chrastil, and Aaron Bornstein (2024). "Cognitive graphs: Representational substrates for planning". In: *Decision*.

Zhou, Dale et al. (Feb. 2020). *Gender Diversity Statement and Code Notebook v1.0*. Version v1.0. DOI: 10.5281/zenodo.3672110.