Selective Maintenance of Value Information Helps Resolve the Exploration/Exploitation Dilemma

Michael N. Hallquist*\textsuperscript{a,b}, PhD and Alexandre Y. Dombrovski*\textsuperscript{b}, MD

\textsuperscript{a} Penn State University, Department of Psychology, 309 Moore Building, Penn State University, University Park, PA 16801, USA. michael.hallquist@psu.edu

\textsuperscript{b} University of Pittsburgh, Department of Psychiatry, 3811 O'Hara St., BT 742, Pittsburgh, PA 15213, USA. dombrovskia@gmail.com

* Both authors contributed equally.

Corresponding author: Alexandre Y. Dombrovski, University of Pittsburgh, Department of Psychiatry, 3811 O'Hara St, BT742, Pittsburgh, PA 15213. Email: dombrovskia@gmail.com

Accepted for publication in Cognition
Abstract (220 words)

In natural environments with many options of uncertain value, one faces a difficult tradeoff between exploiting familiar, valuable options or searching for better alternatives. Reinforcement learning models of this exploration/exploitation dilemma typically modulate the rate of exploratory choices or preferentially sample uncertain options. The extent to which such models capture human behavior remains unclear, in part because they do not consider the constraints on remembering what is learned.

Using reinforcement-based timing as a motivating example, we show that selectively maintaining high-value actions compresses the amount of information to be tracked in learning, as quantified by Shannon’s entropy. In turn, the information content of the value representation controls the balance between exploration (high-entropy) and exploitation (low-entropy). Selectively maintaining preferred action values while allowing others to decay renders the choices increasingly exploitative across learning episodes.

To adjudicate among alternative maintenance and sampling strategies, we developed a new reinforcement learning model, StrategiC ExPloration/ExPloitation of Temporal Instrumental Contingencies (SCEPTIC). In computational studies, a resource-rational selective maintenance approach was as successful as more resource-intensive strategies. Furthermore, human behavior was consistent with selective maintenance; information compression was most pronounced in subjects with superior performance and non-verbal intelligence, and in learnable vs. unlearnable contingencies. Cognitively demanding uncertainty-directed exploration recovered a more accurate representation in simulations with no foraging advantage and was strongly unsupported in our human study.
Highlights

• When searching for the best among a multitude of response times, people cannot indefinitely maintain information about every experience.

• While previous work has focused on the choice policy, e.g. whether to explore more uncertain options preferentially, we argue that human sampling is constrained by the manner in which option values are maintained in memory.

• Uncertainty-directed exploration recovers a more accurate picture of simulated environments, but typically confers no advantage in foraging.

• The alternative approach of mapping the most valuable parts of the world accurately while having only approximate knowledge of the rest is just as successful, requires less representational capacity, and provides a better explanation of human behavior.

• A good strategy for someone with limited memory capacity is to selectively maintain a valuable subset of options and gradually forget the rest.
1. Introduction

*It is better to understand a little than to misunderstand a lot.*

- Anatole France

Laboratory studies of value-based decision-making typically involve choosing among a few actions according to their perceived subjective value (Glimcher & Fehr, 2013). In real life, however, we often face a multitude of options whose values may be unknown or poorly estimated. How can an organism with limited computational resources learn the most advantageous actions in the natural environment? Previous work on boundedly rational agents has considered the role of a limited-capacity working memory system (Collins & Frank, 2012) and the possibility that metareasoning (i.e., a policy guiding how to allocate resources) reduces the complexity of learning in large action spaces (Russell & Wefald, 1991). This study provides a new, complementary account highlighting how the selective maintenance of value information facilitates the search for the best among many actions.

One of the fundamental dilemmas in reinforcement learning is how to choose between exploiting an action with a known positive value and exploring alternatives in search of even more advantageous actions (Sutton & Barto, 1998). Global uncertainty about the overall distribution of action values motivates stochastic exploration (Achbany, Fouss, Yen, Pirotte, & Saerens, 2008). A much-debated question is whether humans are additionally motivated to explore highly uncertain options (Dayan & Daw, 2008). Below, we refer to exploration driven by local uncertainty about specific actions as ‘uncertainty-directed’. An influential idea from artificial intelligence is that agents may receive ‘exploration bonuses’ for exploring highly uncertain options (Sutton, 1990), yet studies using multi-armed bandit tasks have not found evidence of this (Daw, O’Doherty, Dayan, Seymour, & Dolan, 2006). Rather, humans appear to become averse to local uncertainty as the number of options increases (Payzan-LeNestour & Bossaerts, 2011) unless uncertainty-directed exploration is explicitly encouraged.
and the number of options is low (Gershman, 2018; Wilson, Geana, White, Ludvig, & Cohen, 2014). On the other hand, Frank and colleagues presented evidence of spontaneous uncertainty-directed exploration on an instrumental reinforcement-based timing task — the clock task — using their Time-Clock (TC) computational model (Badre, Doll, Long, & Frank, 2012; Frank, Doll, Oas-Terpstra, & Moreno, 2009). Thus, an important unanswered question is whether the large action space of a timing task elicits uncertainty-based exploration even if simpler discrete choice paradigms do not.

In reinforcement-based timing tasks, optimal timing is often uncertain, but can be learned by responding at different moments in time and evaluating the outcome. Unlike learning tasks with just a few actions, reinforcement-based timing requires one to explore a continuous action space to identify response times with high expected value. Assuming some degree of temporal generalization, the complexity of representing time-varying reinforcement can be reduced by a set of temporal basis functions (TBF) that approximate expected value as a function of time. A TBF representation has been validated in temporal difference (TD) models of Pavlovian conditioning (Ludvig, Sutton, & Kehoe, 2008, 2012), providing a parsimonious account of timing that limits the number of values maintained in memory. A further challenge, however, is that memory traces inevitably decay over time, particularly when many action values are kept online (Collins & Frank, 2012; Otto, Raio, Chiang, Phelps, & Daw, 2013). Thus, effective approaches to learning need to be robust to forgetting, ensuring that valuable information is selectively maintained (cf. Maxcey-Richard & Hollingworth, 2013). Building on models of working memory (Baddeley & Logie, 1999; Barrouillet, Bernardin, & Camos, 2004) and the dopamine system (Kato & Morita, 2016), we propose that in reinforcement-based timing, the values of recently sampled actions (response times) are selectively maintained, whereas more temporally distant action values decay. As we illustrate below, such selective maintenance can be resource-optimal (cf. Griffiths, Lieder, & Goodman, 2015), trading off a high-fidelity representation of all available rewards for the opportunity to exploit the best response time.
Algorithmic solutions to the intractable exploration/exploitation dilemma generally focus on the policy, which translates subjective value estimates into choices. The commonly used Boltzmann softmax policy probabilistically selects actions in proportion to their value, while controlling exploration with a temperature parameter (Sutton & Barto, 1998). The degree of exploration in a given state depends on the entropy of the Boltzmann distribution of action probabilities, a logarithmic measure of global uncertainty about which action to choose (Achbany et al., 2008). Entropy is maximal when all actions have equal probability and approaches zero as the probability of choosing one action approaches one. In turn, when the temperature parameter is high, entropy increases and actions are chosen with similar probabilities (exploration), whereas if it is low, the agent prefers high-value actions (exploitation).

While prior work has focused on controlling the softmax temperature or encouraging uncertainty-directed exploration, the similarity among the learned action values per se is a more fundamental determinant of the exploration/exploitation tradeoff. We quantify this similarity as Shannon’s entropy (Cover & Thomas, 2006) of the normalized vector of action values (cf. Payzan-LeNestour & Bossaerts, 2011), which represents the information content of the learned values. For example, if the expected value of one action is much greater than the alternatives, the entropy of action values will be low, increasing the likelihood of exploiting the valued action according to the softmax policy. Below, we show that in the case of reinforcement-based timing, selective maintenance of action values across learning episodes helps to resolve the exploration/exploitation dilemma by reducing entropy. Simulations and a human experiment demonstrate that selectively maintaining values of recently chosen actions while gradually forgetting temporally remote ones is generally superior to tracking the value of all available actions.

Altogether, these considerations motivated the hypothesis that humans selectively maintain value traces in a manner that compresses them, reducing memory load. Second, we hypothesized that the resulting entropy dynamics tune the exploration/exploitation tradeoff via the Boltzmann softmax, better
accounting for human sampling behavior than directed exploration toward uncertain options (Frank et al., 2009; Makarenko, Williams, Bourgault, & Durrant-Whyte, 2002; Sutton, 1990). To test these hypotheses, we developed a new reinforcement learning model, StrategiC ExPloration/ExPloitation of Temporal Instrumental Contingencies (SCEPTIC), that represents continuous action values using basis functions. We tested variants embodying alternative hypotheses, using the extensively validated temporal difference (TD) complete serial compound (CSC) model with timestep-level value representation and stochastic exploration as a benchmark (Sutton & Barto, 1990). Given its use in previous studies of exploration in reinforcement-based timing (Badre et al., 2012; Frank et al., 2009; Moustafa, Cohen, Sherman, & Frank, 2008), we selected Frank’s TC with a trial-level value representation and uncertainty-directed exploration as a comparator model. To evaluate whether tracking local uncertainty confers an advantage over an entropy-driven selective maintenance model, we also devised uncertainty-directed comparators within the SCEPTIC framework. In model comparisons, we used model and parameter identifiability and optimality as preliminary criteria and fit to behavior as the final criterion. We took particular care to rule out alternative accounts of behavior and information maintenance, such as exploration driven by local uncertainty, variable learning rate, and choice autocorrelation.

2. Material and Methods

2.1. Ethics Statement

Participants and/or their legal guardians provided informed consent or assent prior to participation in this study. Experimental procedures for this study complied with Code of Ethics of the World Medical Association (1964 Declaration of Helsinki) and the Institutional Review Board at the University of Pittsburgh (protocol PRO10090478). Participants were compensated $75 for completing the experiment.
2.2. Participants

We enrolled 76 normally developing youth and young adults, aged 14 to 30 (\( M = 21.32, SD = 5.10 \)). Thirty-nine participants (51.3%) were female. Prior to enrollment, participants were interviewed to verify that they had no history of neurological disorder, brain injury, pervasive developmental disorder, or psychiatric disorder (in self or first-degree relatives).

2.3. Procedure

Participants completed eight runs of a reinforcement-based timing task (hereafter called the “clock task”) during an fMRI scan in a Siemens Tim Trio 3T scanner. Runs consisted of fifty trials in which a darkened circle resembling a clock hand revolved 360° around a central stimulus over the course of four seconds (Figure 1a). Participants pressed a button on a button glove to end the trial and receive a probabilistic reward. Time-varying contingencies were taken from the paradigm developed by Moustafa and colleagues (2008) and included monotonically increasing expected value (IEV; reinforcing late responses), decreasing expected value (DEV; reinforcing early responses), constant expected value (CEV), and constant expected value–reversed (CEVR; see Figure 1b). The CEV and CEVR conditions had constant expected value for all response times, but varied in probability and magnitude. The central stimulus was a face with a happy expression or fearful expression, or a phase-scrambled version of face images intended to produce an abstract visual stimulus with equal luminance and coloration. Faces were selected from the NimStim database (Tottenham et al., 2009). All four contingencies were collected with scrambled images, whereas only IEV and DEV were also collected with happy and fearful faces. The emotion manipulation and fMRI results will be reported in separate manuscripts because they are not central for the validation of our model.
Figure 1. The clock paradigm and typical human behavior. a) The clock paradigm consists of decision and feedback phases. During the decision phase, a dot revolves 360° around a central stimulus over the course of four seconds. Participants press a button to stop the revolution and receive a probabilistic outcome. b) Rewards are drawn from one of four monotonically time-varying contingencies: increasing expected value (IEV), decreasing expected value (DEV), constant expected value (CEV), or constant expected value–reversed (CEVR). CEV and CEVR thus represent unlearnable contingencies with no true value maximum. Reward probabilities and magnitudes vary independently. c) Evolution of subjects’ response times (RT) by contingency and performance. Panels represent participants whose total earnings were above or below the sample median. d) Evolution of subjects’ response time swings (RT swings) by contingency and performance.

Participants also completed the Reynolds Intellectual Screening Test (RIST), a brief inventory of verbal and nonverbal intelligence (Reynolds & Kamphaus, 2003) consisting of a verbal subtest measuring verbal reasoning and vocabulary, as well as a nonverbal subtest in which examinees identify which stimulus does not belong with the others in a series of progressively more abstract displays. The RIST has strong test-retest reliability and convergent validity, and correlates highly with full assessments of intellectual ability. The RIST was administered by personnel proficient in psychological testing and supervised by one of the authors (MNH). In our sample, the average RIST Index (a measure of overall intellectual ability) was 105.46 (SD = 9.39; range = 80–129).

2.4. StrategiC ExPloration/ExPloitation of Temporal Instrumental Contingencies (SCEPTIC)

2.4.1. Temporal basis representation.
The SCEPTIC model represents time using a set of unnormalized Gaussian radial basis functions (RBFs) spaced evenly over an interval $T$ in which each function has a temporal receptive field with a mean and variance defining its point of maximal sensitivity and the range of times to which it is sensitive, respectively (a conceptual depiction of the model is provided in Figure 2). The number, width, and spacing of these basis functions can be varied without loss of generality to more substantive parts of the reinforcement learning model, although a richer basis set can represent more fine-grained temporal information. A set of overlapping radial basis functions (RBFs) provides an efficient approximation of an arbitrary function, $f(T)$, over a finite interval (Boyd, 2010a; Buhmann, 2003). From a biological standpoint, the advantages of this approach are that 1) given a fixed number of basis functions, approximation imprecision scales with the length of the interval (Fiorillo, Newsome, & Schultz, 2008); and 2) it is compatible with accounts of response-sensitive neurons with distinct temporal tuning, for example in the medial premotor cortex (Merchant, Zarco, Pérez, Prado, & Bartolo, 2011).

The primary quantity tracked by the basis is the expected value of a given choice (response time). To represent time-varying value, the heights of each basis function are scaled according to a set of $b$ weights, $\mathbf{w} = [w_1, w_2, \ldots, w_b]$. The contribution of each basis function to the integrated value representation at the chosen response time, $t$, depends on its temporal receptive field:

$$
\varphi_b(t) = \exp \left[ -\frac{(t - \mu_b)^2}{2 s_b^2} \right]
$$

where $\mu_b$ is the center (mean) of the RBF and $s_b^2$ is its variance. And more generally, the temporally varying expected value function on a trial $i$ is obtained by the multiplication of the weights with the basis:

$$
V(i) = \mathbf{w}(i) \varphi
$$

2.4.2. Parameterization of temporal basis functions in SCEPTIC.
In order to represent temporal decision-making during the clock task, where the probability and magnitude of rewards varied over the course of four-second trials, we spaced the centers of 24 Gaussian RBFs evenly across the discrete interval and chose a fixed width, $s^2_b$, to represent the temporal variance (width) of each basis function. Because of the challenges of approximating functions over discrete intervals (the Runge phenomenon; see Boyd, 2010a, 2010b), two additional modifications of the basis were required to obtain an accurate function approximation: (1) the time interval represented by the basis was extended 10% beyond the bounds of the trial, and (2) weight updates were performed using a truncated Gaussian basis such that the area under the curve within the time interval of interest was equal for all basis functions (additional details provided in Supplementary Methods).

Although it is plausible that a system coding time-dependent information about rewards may update its temporal width, height, or center of maximal sensitivity on the basis of reinforcement, here we updated only the heights of each basis function based on the reinforcement history. For treatments of basis function adaptation in reinforcement learning models, see Menache, Mannor, and Shimkin (2005) and Mahadevan, Giguere, and Jacek (2013). For the widths of the RBFs, $s^2_b$, we chose a moderate degree of overlap between adjacent basis functions in order to provide reasonable coverage of each moment within the time interval. More specifically, $s^2_b$ was chosen such that the distribution of adjacent RBFs overlapped by approximately 50%, but overlap between 30% and 70% provided similar results. The parameterization of the temporal basis is not a crucial component of the SCEPTIC model, and another temporal basis (e.g., piecewise polynomial splines or discrete cosine transform) would likely yield similar results (see Supplementary Table 2 for basis parameter values).

2.4.3. **Updating expected value on the basis of reinforcement.**

A straightforward model of temporal instrumental learning can be specified by combining the delta learning rule (Bush & Mosteller, 1955) with the temporal representational structure defined above. More specifically, the weight for a basis function $b$ can be updated according to the equation:
where \( i \) is the current trial in the task, \( t \) is the observed response time, and \( \text{reward}(i|t) \) is the reward obtained on trial \( i \) given the choice \( t \). The effect of prediction error is scaled according to the learning rate \( \alpha \) and the temporal generalization function \( e_b \). Of note, this learning rule updates the weight of each basis function, \( w_b(i) \), individually without assuming knowledge of the integrated value representation, \( V(i) \). Thus, the value function approximation does not converge absolutely on the temporal distribution of rewards, instead preserving relative differences (i.e., the rank ordering) of alternative values. We note that to make the model converge on the underlying value function, the learning rule of SCEPTIC variants can be altered to compute prediction errors as the difference between actual reward and the integrated value representation, \( V(i) \). Implementing such a function-wise update does not qualitatively change any of our substantive results. This alternative, however, suffers from lower identifiability and provides a poorer fit to subjects’ behavior. It also requires a stronger assumption about value representation, namely that basis functions query each other during learning to update their weights (details available from authors upon request). We acknowledge that Ludvig and colleagues’ TBF model of Pavlovian learning (2012) applies function-wise value updates, and the reasons for the apparent superiority of elementwise updates here remain to be investigated.

To avoid tracking separate value estimates for each possible moment, it is crucial that feedback obtained at a given response time \( t \) is propagated to adjacent times. Thus, to represent temporal generalization of expected value updates, we used a Gaussian RBF centered on the response time \( t \), having width \( s^2 \) and normalized to have an area under the curve of unity. The eligibility of a basis function \( \varphi_b \) to be updated by prediction error is defined by the area under the curve of its product with the temporal generalization function:
This parameterization leads to a scalar value for each RBF between zero and one representing the proportion of overlap between the temporal generalization function and the receptive field of the RBF. In the case of perfect overlap, where the response time is perfectly centered on a given basis function and the width of the generalization function matches the basis (i.e., $s_g^2 = s_b^2$), $e_b$ will reach unity, resulting a maximal weight update according to the learning rule above. Conversely, if there is no overlap between an RBF and the temporal generalization function $e_b$ will be zero and no learning will occur in the receptive field of that RBF.

2.4.4. Choice rule.

Finally, having defined a framework for tracking and updating estimates of expected value over time, SCEPTIC variants select an action based on a softmax choice rule, analogous to simpler reinforcement learning problems (e.g., two-armed bandit tasks; Sutton & Barto, 1998). For computational speed, we arbitrarily discretized the interval into 100ms time bins such that the agent selected among 40 potential responses. The agent chose responses in proportion to their expected value:

$$p(rt(i + 1) = j \mid V(i)) = \frac{\exp \left( \frac{V(i)_j}{\beta} \right)}{\sum_{t=0}^{T} \exp \left( \frac{V(i)_t}{\beta} \right)}$$

where $j$ is a specific response time and the temperature parameter, $\beta$, controls the sharpness of the decision function (at higher values, actions become more similar in selection probability).

The softmax function has two potentially desirable properties in the temporal instrumental learning context. First, if several actions are associated with similar expected value, even if substantially separated in time, they will be selected with similar probability. Second, by virtue of the temporal basis
representation (where reinforcement information is generalized in time), response times adjacent to the global maximum of learned expected value are more likely to be selected, promoting temporally local exploration of advantageous areas. To highlight the advantages of the softmax policy, we contrasted it with a $\epsilon$-greedy choice rule, which was predictably inferior in optimality simulations (details provided in Supplementary Materials).

2.4.5. Overcoming cognitive constraint on value representation: the selective maintenance model.

We tested a selective maintenance model (Figure 3) in which basis weights reverted toward zero in inverse proportion to the temporal generalization function:

$$ w_b(i + 1) = w_b(i) + e_b(i|t)\alpha[\text{reward}(i|t) - w_b(i)] - \gamma(1 - e_b(i|t))(w_b(i) - h) \quad (6) $$

where $\gamma$ is a selective maintenance parameter between zero and one that scales the degree of reversion toward a point $h$, which is taken to be zero here, but could be replaced with an alternative, such as a prior expectation.

2.4.6. Representing value and uncertainty according to the Kalman filter.

The Kalman filter (KF) is a classic Bayesian approach to estimating the expectation (mean) and uncertainty (variance) of a Gaussian process that unfolds in discrete time (for a classic example of Kalman filters in reinforcement learning models, see Dayan, Kakade, & Montague, 2000). We also note that Frank and colleagues tested a Kalman filter variant of their model to track expected value, rather than the probability of prediction error, and our work built on this useful insight. In the SCEPTIC model, each basis function can be reconceptualized as a Kalman filter that tracks information about both the expected value of a response (i.e., the mean) in its temporal receptive field as well as uncertainty about expected value. Crucially, integrating across basis functions, KF variants of the SCEPTIC model represent
both time-varying value and uncertainty functions ($V$ and $U$, respectively), enabling policies that integrate information from both sources.

Compared to SCEPTIC variants described above that rely on a fixed learning rate to update basis weights, there are three major differences for KF variants: 1) the effective learning rate (gain) declines with experience such that early outcomes have the greatest effect on learning, 2) the model tracks the evolution of uncertainty about expected value, and 3) the choice rule (policy) for some models involves a tradeoff between exploratory and exploitative influences. The learning rule for KF SCEPTIC variants is:

$$
\mu_b(i + 1) = \mu_b(i) + e_b(i|t)k_b(i)[\text{reward}(i|t) - \mu_b(i)]
$$

(7)

where $\mu_b(i)$ represents the contribution of basis function $b$ on trial $i$ to the expected value function, $V(i)$. The gain (learning rate) for a given basis function on trial $i$ is defined as

$$
k_b(i) = \frac{\sigma_b(i)^2}{\sigma_b(i)^2 + \sigma_{rew}^2}
$$

(8)

where $\sigma_{rew}^2$ represents the expected volatility (measurement noise) of the environment. Here, we provided the model the variance of returns from a typical run of the experiment as an initial estimate of measurement noise, although other priors led to similar model performance. We also initialized prior estimates of uncertainty for each basis function to be equal to the measurement noise, $\sigma_b^2(0) := \sigma_{rew}^2$, leading to a gain of 0.5 on the first trial (as in Frank et al., 2009).

Under the KF, the contribution of each basis function to uncertainty about expected value is represented as the standard deviation of its Gaussian distribution. Likewise, posterior estimates of uncertainty about responses proximate to the basis function $b$ decline in inverse proportion to the gain according to the following update rule:
\[ \sigma_b(i + 1) = [1 - e_b(i|t)k_b(i)]\sigma_b(i) \]  

(9)

Note that the temporal generalization function \( e_b(i|t) \) is parameterized identically across SCEPTIC variants and is used in KF variants to update both value and uncertainty estimates. Extending the temporal representation described above, for KF variants, estimates of the time-varying value and uncertainty functions are provided by the evaluation of the basis over time:

\[ V(i) = \mu(i)\varphi \]  

(10)

\[ U(i) = \sigma(i)\varphi \]  

(11)

2.4.7. Integrating uncertainty and expected value in response selection under KF SCEPTIC variants.

Early in learning, expected value will be low for most responses and uncertainty will be high, whereas the converse will be true late in learning. Thus, a policy that combines value and uncertainty may confer particular advantages because uncertainty-directed responses early in learning would facilitate more robust and efficient sampling. The KF \( U + V \) policy represents a decision function, \( Q(i) \), as a weighted sum of the value and uncertainty functions according to a free parameter, \( \tau \). As uncertainty decreases with sampling and expected value increases with learning, value-related information will begin to dominate over uncertainty. Positive values of \( \tau \) promote uncertainty-directed exploration, whereas negative values yield uncertainty aversion.

\[ Q(i) = V(i) + \tau U(i) \]  

(12)

To ensure that our model comparison results were robust to the specific implementation of the uncertainty-sensitive choice rule and dynamic learning rate, we tested a number of alternatives, detailed in the Supplementary Material, all of which proved to be inferior to the variants described above.
2.5. Information content of learned values.

This was estimated by Shannon’s entropy of basis function weights. The weights were normalized to have an area under the curve of unity (cf. Hausser & Strimmer, 2008), although we note that other normalization methods will yield similar results.

\[ H(w) = - \sum_{b=1}^{B} w_b \log_{10}(w_b) \]  

(13)


To rule out the possibility that the selective maintenance model fit well because it better represented sticky choices, we extended SCEPTIC models with two choice autocorrelation functions (ACF): a simple first-order autoregressive (AR[1]) ACF and an ACF extended over multiple trials (Schönberg, Daw, Joel, & O’Doherty, 2007). The AR(1) model modulated the probability of choosing a given time point \( t \) in a trial by \( \chi \cdot \pi^{\lfloor t - rt(i) \rfloor} \) (we omit the trial index \( i \) for simplicity in this paragraph), where \( \chi \) is the autocorrelation parameter and \( \pi \) is the temporal generalization parameter, followed by divisive normalization. The extended ACF maintained for each time point \( t \) an index \( c_t \) of how recently it was chosen. When \( t \) was chosen, \( c_t \) was set to 1; otherwise it decayed by a factor \( \lambda t \) being chosen in the softmax was a function of its value and the additive term \( \chi \cdot c_t \). We also tested a version of extended ACF with temporal generalization implemented using Gaussian smoothing with a kernel representing a temporal generalization parameter, but this resulting model had inferior fits (data available upon request).

2.7. Temporal Difference (Q-learning) Model

Compared to SCEPTIC variants, the complete serial compound (CSC) variant of the Q-learning model (Sutton & Barto, 1998) of instrumental learning represents the value of potential actions at each
discrete time step in the trial. We chose this as a robust benchmark for SCEPTIC models because TD models have been established in simulation and electrophysiological studies of time-based value representation (Fiorillo et al., 2008; Ludvig et al., 2008; Schultz, Dayan, & Montague, 1997). Moreover, by discretizing the four-second trial interval into 40 timesteps (100ms), Q-learning and SCEPTIC model fit could be compared on the same scale because both models produce a multinomial prediction about response time. For application to the clock task, we amended the standard Q-learning model in two ways. First, to overcome the problem of erroneous value back-propagation, where earlier responses become over-valued (Moustafa et al., 2008), we set up the state space such that each time step \( t \) has a pair of actions, \( A = \{ \text{wait}, \text{respond} \} \). For wait actions, we ensured that value was appropriately back-propagated only to previous waits:

\[
Q(s_t, a_t | a_t = \text{wait}) \leftarrow Q(s_t, a_t) + \alpha \left[ \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]
\]

where \( \alpha \) is the learning rate and \( \gamma \) is the discount parameter. We use the conventional \( Q(s_t, a_t) \) notation here, but since time steps \( t \) always map onto the same states (potential response times), \( s_t \) is redundant and \( Q(a_t) \) would suffice. Conversely, because respond actions led to the absorbing terminal state, ending the trial, they were directly updated by actual rewards and not by back-propagation:

\[
Q(s_t, a_t | a_t = \text{respond}) \leftarrow Q(s_t, a_t) + \alpha [e \cdot r_{t+1} - Q(s_t, a_t)]
\]

where \( e \) is the eligibility trace or credit for a reward assigned to a given action based on its temporal proximity. We assumed \( e \) to decay exponentially over time.

Second, we employed a modified \( \epsilon \)-greedy choice rule to ensure that the agent explored the later part of the interval:
\[
\pi(s_{t+1}) = \begin{cases} 
\text{wait} & \text{if } R \leq 1 - \frac{1}{T-t} \text{ if } \xi < \varepsilon \text{ (explore)} \\
\text{respond} & \text{otherwise} \\
\arg\max_a Q(s_t, a) & \text{otherwise (exploit)}
\end{cases}
\]  

(16)

where \( \varepsilon \) is the exploration/exploitation parameter, \( t \) is the current time step, \( T \) is the total number of time steps, and \( \xi \) and \( R \) are \([0,1]\) uniform random numbers drawn at each time step. While a simple \( \varepsilon \)-greedy agent does not effectively explore later time steps because of the fixed exploration probability, the \( 1 - \frac{1}{T-t} \) term produces more exploratory \textit{waits} early in the trial and more \textit{responses} late in the trial. A SARSA model was also tested, but is not included here because it was inferior to Q-learning in almost every respect (data available upon request).

**2.8. Frank Time-Clock (TC) Model**

To date, the TC model of Frank and colleagues (2009) has been applied in behavioral, genetic, and neuroimaging studies of uncertainty-directed exploration and dopaminergic influences on response times (Badre et al., 2012; Moustafa et al., 2008). Crucially, studies using the TC model have provided initial evidence of uncertainty-directed exploration on the clock task (Badre et al., 2012), making this an ideal comparator model for the present study. The TC model represents response times on each trial \( i \) as a linear combination of several potentially neurobiological processes:

\[
\hat{RT}(i) = K + \lambda RT(i - 1) + \nu[R_{\text{best}} - R_{\text{avg}}] - Go(i) + \text{NoGo}(i)
\]

\[
+ \rho[\mu_{\text{slow}}(i) - \mu_{\text{fast}}(i)] + \varepsilon[\sigma_{\text{slow}}(i) - \sigma_{\text{fast}}(i)]
\]

(17)

The details of each parameter and the underlying representation are provided in the Supplementary Methods. With respect to value-based decisions, the TC model separately updates the probability of a positive prediction error (PPE) for RTs that are slower or faster than the subject’s
average ($\mu_{\text{slow}}$ and $\mu_{\text{fast}}$, respectively). With learning, the model predicts that subjects shift toward faster or slower RTs that are associated with a greater expectation of a PPE according to a free parameter, $\rho$.

2.9. Data Analysis

2.9.1. Tests of model optimality.

In order to test and compare the efficacy of each model in learning temporal contingencies, we identified parameter sets that maximized the quality of choices in simulations of the clock task. More specifically, for each model, we fit parameters using a genetic algorithm ($ga$ function in MATLAB) that returned the greatest summed expected value across 60 runs consisting of 50 trials each. The temporal contingency was based on two sinusoidal functions for expected value and probability (see Figure 4a and Supplementary Methods for details). Next, we validated the robustness of parameter sets to contingencies that were not part of the parameter optimization. For this step, we simulated rewards earned by each model at the five best parameter sets from optimization. Each model was exposed to 100 instances of a randomly phase-shifted variant of the sinusoidal contingency and made choices according to its parameters. We then analyzed the distribution of returns for each model using multilevel regression models ($lmer$ function in R 3.4.0) where contingency instances were modeled by a random intercept (since contingencies were drawn from the population of possible contingencies) and model was treated as a fixed factor. Because models differed considerably in the variability of earnings across replications, we allowed for heteroscedastic residual variance by model.

Because the objective functions for most models were non-convex and prone to local minima, optimization using a genetic algorithm ($ga$ function in MATLAB) was repeated 100 times using different random starting values spanning the parameter space for the initial population (see Supplementary Methods for details). This resulted in a distribution of returns on policy, as well as multivariate distributions of parameters for each optimization. These data provided information about the ability of each model to learn the temporal contingency when parameters were tuned for that environment.
2.9.2. Identifiability.

In addition to comparing the ability of each model to solve temporal instrumental problems in simulated environments, we tested their recovery of each parameter in simulated data. As in conventional simulation studies of estimators (for a useful treatment in a neuroscience context, see Kass, Eden, & Brown, 2014), we focused specifically on variance (i.e., dispersion of estimated values around the population value, reported as $R^2$ between original and recovered parameters) and parameter bias (i.e., systematic deviations between population and estimated values). For each model, we simulated the behavior of 100 agents by drawing all parameters from the uniform distribution and fitting all parameters at once. For models using a softmax choice rule, the temperature parameter was not recoverable. Since this parameter tends to absorb misfit and account for exogenous factors when fitting behavior, we fixed it here at 0.1, a value corresponding to high exploitation that highlighted more substantive parts of the model. To avoid bias, we took a similar approach when fitting TD. We estimated best-fitting parameters using the $ga$ in MATLAB similarly to our optimality tests.

2.9.3. Subject behavior fitting.

We fit computational models to participant behavior in a deterministic state-space framework using a variational Bayes approach (VBA) implemented in MATLAB (Daunizeau, Adam, & Rigoux, 2014). An important advantage of this approach is that the relative evidence for different models can be compared using random effects Bayesian model comparison (BMC; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). This BMC treats both subjects and models as random variables, handling heterogeneity in model fits within a sample and allowing one to compute the probability that one model better characterizes the sample than all others (i.e., the exceedance probability). Here, we report protected exceedance probabilities (pEP), which examine if one model is more frequent than the others after accounting for chance. This notion builds on the Bayesian omnibus risk (BOR), which is measure of
statistical risk in group model comparisons quantifying whether chance is likely to explain differences in estimated model frequencies (Rigoux, Stephan, Friston, & Daunizeau, 2014).

In SCEPTIC and TD models, we discretized each subject’s response times into 40 time bins (100ms each) such that every trial was represented by a 40-element vector with a value of 1 at the chosen time bin and zeros elsewhere. This 400 x 40 matrix of trials by time bins was the basis for subject behavior fitting by VBA. The output of the model was a multinomial choice matrix of the same dimension containing model-predicted probabilities for every time bin on every trial. Furthermore, the global fit of the model to the subject’s behavior was quantified by the log model evidence based on the negative free energy from the VBA estimation algorithm (for details, see Stephan et al., 2009).

VBA parameterizes the choice history in a state-space framework consisting of dependent variables (i.e., time series to be predicted by the model), hidden states (i.e., latent quantities to be tracked over trials), and evolution and observation functions that define the dynamics of hidden state transitions and the model-predicted output, respectively (for details, see Daunizeau et al., 2014). For SCEPTIC models, the hidden state vector was composed of basis function weights representing expected value, which were initialized from a random uniform prior distribution spanning the range of values on the underlying contingencies. For variants that also tracked uncertainty, posterior uncertainty estimates were also tracked for each basis function and initialized as the variance of values across all timesteps in the distribution. We used weakly informative Gaussian priors ($M = 0, SD = 10$) for all other free parameters. The learning rate and selective maintenance parameters ($\alpha$ and $\gamma$, respectively) were bounded between zero and one. To accommodate this while still allowing the joint posterior to follow a multivariate Gaussian distribution under the mean field approximation (Daunizeau et al., 2014), Gaussian-distributed parameters for learning rate and selective maintenance were transformed via a logistic function prior to use in the learning rule. Because the temperature parameter, $\beta$, must be non-
negative, we exponentiated the Gaussian-distributed parameter estimate prior to its use in the choice rule.

For Q-learning, the hidden state vector consisted of estimated Q values for respond and wait actions at each time step (80 total hidden states). Similar to SCEPTIC, we used weakly informative Gaussian priors ($M = 0, SD = 10$) for both free parameters, transforming learning rate, $\alpha$, and exploration, $\varepsilon$, to the interval $[0,1]$ via a logistic function.

For TC, the hidden states tracked by the model were a) the response time associated with the largest reward experienced in a block, b) the $\alpha$ and $\beta$ hyperparameters for two beta distributions tracking the value and uncertainty of slow and fast responses, c) the expected value of each choice tracked according to a delta-rule model with learning rate of 0.1, and d) the value of Go and NoGo decision signals, e) the locally averaged response time, $RT_{locavg}$. Model parameters were initialized with broad weakly informative priors, although because of differences in scaling and parameterization, the distributional form of parameters varied (see Supplementary Table 1). Initial values for parameters were chosen based on Frank and colleagues (2009) and code generously provided by Michael Frank. Parameter distributions were chosen based on optimization bounds in the previous TC implementation, as well as observations about TC parameters in this and previous datasets. As noted above, the TC model provides a scalar estimate of the predicted response time on each trial, rather than estimating the probability of a response at any moment within the trial. Thus, the fit of the model was based on the correspondence between the observed and predicted RTs, rather than a multinomial choice distribution as in the case of SCEPTIC and Q-learning.

2.9.4. Mixed-effects analyses of computational experiments and human behavior

Most of the outcome measures from the computational and human experiments had a clustered structure such that there were many observations per simulation or person. For example, computational models of behavior yielded estimates of the expected value of the chosen response time on every trial,
and human participants completed 400 trials. Aggregating these data into summary statistics (e.g., mean response time) would limit the ability to resolve trial-level effects of interest (e.g., high entropy of the value distribution on one trial predicting longer responses on the next) and would also sacrifice precision and power (Aarts, Verhage, Veenvliet, Dolan, & van der Sluis, 2014; Singer & Willett, 2003). Thus, when data had a clustered structure (e.g., trials nested within subjects), we used multilevel regression models to estimate the effects of interest. Multilevel models were estimated using restricted maximum likelihood in the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R 3.4.0 (R Core Team, 2017). Estimated p-values for predictors in the model were computed using Wald chi-square tests.

To compare the optimality of different models in solving temporal instrumental contingencies, we estimated a multilevel model in which performance was regressed on model and run length (40, 60, or 110 trials); parameter set (best 5 sets for each model) and replication (100) were treated as random effects. Simple effects tests of model for each run length were estimated and adjusted p-values were computed according to the multivariate distribution of coefficients to maintain a familywise error rate of .05 for each run length (Hothorn, Bretz, & Westfall, 2008). We varied run lengths between 20 and 200 in increments of 5, but chose to report a subset of contingencies that were illustrative model performance changes as a function of run length.

For analyses of human behavior, most multilevel regressions were run on trial-level data in order to capture the temporal dynamics of learning and performance. To test temporal precedence in trial-level data (e.g., previous reward predicting a change in current RT swing), relevant predictors were lagged by one trial. For analyses focused on how early versus entropy controls overall success in a given run, behavioral indices were aggregated to the run level (esp. total points earned in a run). For trial-level analyses, subject and run were treated as random effects. In the case of run-level data, subject was treated as random.

2.9.5. Mixed-effects survival analyses of behavior
We also performed survival analyses predicting the temporal occurrence of response. These analyses employing mixed-effects Cox models (R `coxme` package; Therneau, 2018) were intended to ascertain that effects of model-predicted expected value and uncertainty on response times were robust to assumptions and potential confounds. First, survival analysis foregoes the assumption that the subject pre-commits to a given response time, instead modeling the within-trial response hazard function in real, continuous time (Singer & Willett, 2003). Second, this framework accounts for censoring of later within-trial time points by early responses. Most importantly, it assumes a completely general baseline hazard function, allowed to vary randomly across subjects and runs within-subject in our model. This feature avoids assumptions about the statistical distribution of response times and accounts for trial-invariant influences on response times. These influences may include motivation, fatigue in later runs, urgency, and opportunity cost. We also modeled only the 500 – 3500 ms interval, excluding response times that are too fast to be possible and the end of the interval which one may avoid in order to not miss responding on a trial. We included learned value from the selective maintenance model and uncertainty from the KF U + V model as time-varying covariates, sampled every 100 ms. Random effects included subject and run nested within subject.

3. Overview of Alternative Models

3.1. SCEPTIC architecture (Table 1).

Unlike tasks with a few actions whose values can be learned discretely, the clock task has a rather large action space in which the expected value of responding at any particular moment may be unique (Figure 1a-b). Choosing how to represent this time-varying contingency involves a tradeoff between generality and parsimony. A general temporal value representation is exemplified by the complete serial compound (CSC) in TD, which tracks a potentially unique value at every time step. This solution comes at the at the considerable cognitive and computational cost of maintaining many value estimates, which are updated after each time step. On the other hand, parsimonious parametric models such as Frank’s TC
model of the clock task are representationally lean, but make assumptions about the environment that may not hold on other tasks. This loss of generality poses a risk of failure when boundary conditions are encountered, such as new environments or contingencies. A middle ground between these extremes is potentially provided by the function approximation approach to learning using basis functions (e.g., Tsitsiklis & Van Roy, 1997). In the context of TD models of Pavlovian conditioning, Ludvig and colleagues (2008) have approximated the evolution of value over time with a mixture of Gaussian temporal basis functions (TBFs).

Extending their work, SCEPTIC uses Gaussian TBFs to approximate the time-varying instrumental contingency. Each function has a temporal receptive field with a mean and variance defining its point of maximal sensitivity and the range of times to which it is sensitive. The weights of each TBF are updated according to a delta learning rule (Figure 2; detailed in Material and Methods). Further, whereas learning and choice in TD occur on a moment-to-moment basis, humans are often more strategic, considering the entire interval at once, an observation reflected in Frank’s TC model (2009). Building on this insight, SCEPTIC considers updates and choices at interval level, reducing the number of computations relative to TD by a factor of $T$, where $T$ is the number of time steps in TD. All SCEPTIC models (Table 1) shared this general architecture.
Figure 2. The SCEPTIC model represents the clock paradigm using a set of temporal basis functions (TBFs) spaced evenly over the time interval (middle row). These TBFs approximate a continuous time-varying expected value function (top row). The effect of prediction errors on value estimates are spread symmetrically in time according to a temporal generalization function centered on the chosen response time (bottom row). The red dot represents the chosen response time and the adjacent text indicates the reward outcome. Data from a randomly selected subject. a.u. = arbitrary units.
Table 1. Summary of key models tested

<table>
<thead>
<tr>
<th>Model</th>
<th>Temporal basis function</th>
<th>Uncertainty-directed exploration</th>
<th>Value learning rule</th>
<th>Interval-level (vs. time step-level) policy</th>
<th>Response</th>
<th>Number of free parameters</th>
<th>Number of fixed structural parameters*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed learning rate, value (LR V)</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Fixed LR uncertainty + value (U+V)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Fixed LR selective maintenance</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Kalman filter (KF), softmax</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>KF U+V</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Time-clock (TC)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>Continuous</td>
<td>7</td>
</tr>
<tr>
<td>Temporal difference (TD), Q-learning</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Multinomial</td>
<td>2</td>
</tr>
</tbody>
</table>

* Fixed structural parameters (in SCEPTIC, number of basis functions, basis function width, width of temporal generalization function, measurement noise of the Kalman filter) were not optimized in studies of optimality or human behavior. Their values are given in Supplementary Table 2. In TC, they include the number of hidden states (fast, slow), average response time, and the fixed learning rates for updating value and tracking the average of recent response times. In Q-learning, they include the number of time steps and the discount parameter.
3.1.1. **SCEPTIC: selective maintenance.**

To test the hypothesis of selective maintenance of value representations under cognitive constraints, we developed a version of the SCEPTIC model, Fixed LR V selective maintenance, that allowed for the forgetting of value traces that were not selected on a given trial (illustrated in Figure 3). Its reward values reverted toward zero in inverse proportion to a temporal generalization function (Figure 2). Below, we refer to it as the “selective maintenance model.” By erasing the reinforcement history in seldom-visited parts of the interval, such selective maintenance tends to decrease the information content of the action value representation later in learning (Fig 3b). We quantified the amount of information contained in the value representation as Shannon’s entropy of the normalized vector of basis function weights, \( w \).
Figure 3. Representation of expected value in SCEPTIC models with and without selective maintenance (data from a representative subject). Top: evolution of entropy over trials; bottom left: action values at trial 3; bottom right: action values at trial 50. Notable differences emerge late in learning (right). Whereas the full-maintenance model (Fixed LR V; green line) contains a more detailed representation of the contingency, the selective maintenance model (orange line) tends to track a single value bump corresponding to a hypothesis about the best response time. This corresponds to a lower information content (entropy) of value representations in the selective maintenance model compared to the full model (see Figure 7 for more detail). The dots in the figure denote the weights for each basis function, which are multiplied by the Gaussian basis to form the integrated value representation, $V(i)$. Although the agent selects actions based on $V(i)$, we depict basis weights here because they are used to compute entropy, as shown in the top panel.

The maximum of each representation is rescaled to the same value (1.0) to facilitate comparison. Absolute values will be depressed by selective maintenance, but only relative values impact choice. a.u. = arbitrary units.
3.1.2. **SCEPTIC: the impact of uncertainty on exploration.**

To embody the alternative hypothesis that exploration is modulated by uncertainty, we developed SCEPTIC variants where choice was influenced by both uncertainty (U), estimated by Bayesian filtering, and reward value (V). In U+V models, choice was controlled by a weighted sum of uncertainty and value according to a parameter, \( \tau \), that could assume positive values reflecting uncertainty-directed exploration, or negative values reflecting uncertainty aversion. Since uncertainty may impact not only exploration but also the learning rate (Behrens, Woolrich, Walton, & Rushworth, 2007; Mackintosh, 1975; Mathys, Daunizeau, Friston, & Stephan, 2011; Pearce & Hall, 1980), we examined the impact of uncertainty in both fixed learning rate (LR) models and in models where the learning rate was controlled by a Kalman filter (KF). Further, to ascertain that our model comparison results are not limited to this specific implementation of uncertainty-directed exploration and fixed or dynamic learning rate, we tested a number of alternative models, described in the Supplementary Materials.

3.2. **Alternative account of uncertainty-directed exploration: the time-clock (TC) model.**

Developed by Frank and colleagues for the clock task (Frank et al., 2009), TC represents response times as a function of seven decision signals, including three free parameters of no interest reflecting subject’s mean response time, choice autocorrelation (Lau & Glimcher, 2005), and modulation toward the best outcome experienced thus far. Two parameters represent speeding or slowing of response times due to prediction errors, inspired by a computational model of the basal ganglia (Frank, 2006). Two final parameters represent the influence of expected value and outcome uncertainty. One noteworthy aspect of the TC model is that instead of separating the learning rule from the choice rule, decision signals contribute additively to the predicted response time.

3.3. **Benchmark: modified temporal difference (TD) model.**

TD explains a wide range of behavioral and neural dynamics on learning tasks, but has previously been shown to fail on the clock task (Moustafa et al., 2008) due to an erroneous back-propagation of
value. To obtain a robust benchmark, we adapted a TD Q-learning model with a complete serial compound stimulus representation. As detailed in Material and Methods, we modified the TD model to remediate flaws noted by Moustafa and colleagues.

4. Results

4.1. Computational experiment: Model validation in simulated environments

In order to examine whether model parameters could be estimated reliably from behavioral data, we conducted a series of parameter recovery simulations. In a variety of environments, the parameters of key SCEPTIC models (Table 1) were identifiable (all $R^2 > .93$), whereas the TD and TC models had problems with parameter indeterminacy (for details, see Supplementary Figure 1 and Supplementary Results). As described below, Bayesian model comparison provided strong evidence that the selective maintenance model characterized human behavior better than alternatives. In additional simulations, we corroborated that the selective maintenance model was reliably recovered when it generated the behavioral data (protected exceedance probability [pEP] = 1.0), whereas it did not characterize data simulated from other models (pEPs < .02; additional details in Supplementary Results).

4.2. Computational experiment: Does uncertainty-directed exploration improve foraging success in simulations?

We tested models’ optimality — or foraging success — in multiple novel environments, simulated with a set of complex non-monotonic temporal contingencies with local minima (Figure 4a). Using the five best parameter sets for each model from an initial search for optimal parameters, we simulated the proportion of possible points earned across 100 novel contingencies. We varied run lengths (40, 60, or 110 trials; detailed in Material and Methods) to gauge models’ relative performance early and late in learning. The average proportions earned as a function of model and run length are depicted in Figure 4b.
Figure 4. a) Expected value (EV), probability, and magnitude of rewards for sinusoidal temporal contingency used for model optimality comparisons. b) Mean proportion of possible points earned in simulated learning of sinusoidal time-varying contingencies as a function of computational model and run length. Outcomes were drawn from 100 phase-shifted variants of a sinusoidal contingency (details in Material and Methods). Dots represent the mean, whereas the intersecting line represents the bootstrapped 95% confidence interval around the mean. The model naming scheme is detailed in Table 1.

We found significant main effects of model and run length on the proportion of points earned (ps < .0001) that were qualified by a model x run length interaction, $\chi^2(14) = 228.13$, $p < .0001$. Regardless of run length, models that explicitly represented uncertainty (KF U + V and Fixed LR U + V) did not perform significantly better than simpler fixed learning rate models (Fixed LR V Selective Maintenance and Fixed LR V), adj. ps > .49. The KF V model performed significantly worse than other SCEPTIC variants at all run lengths, adj. ps < .01. TD performed worse than the top four SCEPTIC models at run lengths of 40 and 60 trials (adj. ps < .05), but not 110 (adj. p > .10). Likewise, TD was not significantly better than TC for 40- and 60-trial runs, but outperformed TC for 110-trial runs, adj. $p < .001$. Finally, TC did not significantly exceed the random exploration null model at any run length, adj. $p > .10$, whereas all other models did,
adj. ps < .05. To control for familywise error, p-values for pairwise model comparisons within each run length were adjusted by computing a single general linear hypothesis test using the multivariate distribution of the coefficients of interest from a multilevel regression that included run length and model as fixed effects of interest (Hothorn et al., 2008).

Contrary to our expectation, SCEPTIC variants that leveraged uncertainty to guide choice did not outperform fixed learning rate models guided by value alone, despite having access to information that could be used to sample the action space more systematically. Notably, however, the correlation between the model’s expected value and the underlying contingency was significantly higher early in learning (especially in the first 10 trials) for models that represented uncertainty compared to models that did not, p < .0001 (Supplementary Figure 2). Thus, despite yielding a higher fidelity representation of the contingency, uncertainty representation did not enhance overall model performance in simulations.

The selective maintenance model did not perform worse than its full maintenance analog (Fixed LR V; adj. p > .10), suggesting that selective maintenance of the value representation did not hamper foraging. Crucially, in a model variant where value traces decayed randomly across the interval, rather than as a function of choice history, performance was impaired (40 trials adj. p = .15, 60 trials adj. p < 10^-5; 110 trials adj. p = .002; optimized parameter set with γ = 0.24). This demonstrates that selective, but not random, decay promotes adaptive exploitation by maintaining the value of preferred actions (potentially stabilizing a value bump). Finally, to identify boundary conditions where entropy-driven exploration would not suffice, we tested our models in sparse, discontinuous environments with competing value maxima (Supplementary Figure 3). As detailed in the Supplementary Results, SCEPTIC variants with an explicit uncertainty representation gained a modest advantage (Supplementary Figure 4). Nevertheless, the selective maintenance model was superior to its full-maintenance equivalent, suggesting that maintaining a subset of valuable actions is efficient even in sparse environments, perhaps in order to avoid the inferior parts of the action space.
4.3. Human experiment

4.3.1. The clock paradigm and model-free overview of human behavior

The clock task is depicted in Figure 1a-b. Subjects were asked to find the “best” response time during a 4 s interval. Outcomes were controlled by one of the four probabilistic contingencies with varying probability/magnitude tradeoffs (1b), two of them learnable with value maxima in the beginning (decreasing expected value, DEV) or the end (increasing expected value, IEV) of the interval, and two unlearnable (constant expected value, CEV and constant expected value-reversed, CEVR). This design results in a high level of uncertainty and encourages trial-by-trial learning, making it difficult to find an optimal strategy.

As in previous studies (Badre et al., 2012; Moustafa et al., 2008), with learning, subjects’ response times (RT) rapidly shifted toward value maxima: late in the interval in IEV and very early for DEV (Figure 1c). As expected, these shifts were more prominent in better-performing subjects and were not apparent in unlearnable contingencies. The rate of exploration, as measured by trial-wise change in response times (i.e., ‘RT swings’), declined with learning and was higher in unlearnable contingencies (CEV, CEVR; Figure 1d) and also in poorly performing subjects (Fig. 1d, right vs. left panel), highlighting a stochastic underlying process. Interestingly, RT swings declined in both learnable and unlearnable contingencies. Even more remarkable was the fact that the switch from exploration to exploitation in both unlearnable contingencies was more pronounced in better-performing subjects, indicating that they tended to settle into a perceived value maximum even where objectively there was none. This suggests that successful learners rely on a mechanism that accelerates the transition from exploration to exploitation. At the same time, these results cast doubt on the strategic uncertainty-directed nature of RT swings, at least beyond the first few trials, for two reasons. First, uncertainty-directed exploration should improve performance by uncovering a value maximum, while we see the exact opposite: persistent RT swings reflect stochastic responding and indicate ignorance of the value maximum in learnable contingencies.
Second, uncertainty-directed exploration cannot explain higher RT swings in unlearnable contingencies, since constant expected value leads to more uniform sampling (Figure 1c), which diminishes uncertainty gradients. In summary, stochastic exploration appears to underlie RT swings without an obvious need to invoke uncertainty-seeking.

### 4.3.2. Representation: TBF is superior to TD.

Random-effects Bayesian model comparison indicated that SCEPTIC models using TBFs afforded a better fit to behavior than TD, as determined by protected exceedance probability (pEP) of 1 vs. 0 (Figure 5a; Rigoux et al., 2014). The representational power of the temporal basis was not simply due to a high number of hidden states ($b=24$, cf. 80 actions tracked by TD). SCEPTIC fits were qualitatively unchanged regardless of the number of basis function elements (data available upon request).

### 4.3.3. Selective maintenance of value representations.

In a Bayesian model comparison, the selective maintenance model dominated (Figure 5a; pEP = 1, Bayesian omnibus risk [BOR] < $10^{-51}$), indicating that subjects preferred recently visited segments of the interval much more than would be predicted by their long-term expected value. The advantage of the selective maintenance model — measured by the log ratio of free energies vs. each of the other models — was greater in better-performing subjects, all rs(74) > .40, ps < .001, except in comparison with TD: $r(74) = .06$, $p = .48$, suggesting that the selective maintenance model captured an adaptive strategy. Moreover, the selective maintenance parameter, $\gamma$, correlated significantly with total points earned on the task, $r(74) = .37$, $p < .001$. Posterior predictive checks on the fit of the model to subjects’ behavior suggested that the selective maintenance model captured trial-by-trial variation in behavior, as well as individual differences in response tendencies (Figure 5b).
**Figure 5.** a) Random-effects Bayesian model comparison of SCEPTIC variants and TD. EP = exceedance probability. Dots represent the estimated model frequency, and the intersecting line represents the standard error of the estimate. BOR = Bayesian omnibus risk. Model variants are detailed in Table 1. For a group model comparison including additional SCEPTIC variants, see Supplementary Figure 5. b) Trial-by-trial fits of the selective maintenance model in three randomly selected subjects (rows) at minimally informative prior values of the parameters (‘priors’, left) and at posterior values (right). Subjects’ responses are indicated in red; the model’s posterior predictive density is depicted in blue. Model predictions at the priors appear to capture subjects’ response tendencies rather accurately, while predictions at the posterior values seem more blurred. This change occurs because exploitative choices are predicted precisely, whereas exploratory large RT swings are difficult for the model to predict, degrading overall model evidence at the priors. This misprediction drives an increase in the softmax temperature (stochasticity) parameter, blurring predictions and improving model evidence at the posterior values, particularly for the more poorly performing subjects in the middle and bottom rows.
In addition, using subject parameter estimates from the selective maintenance (SM) model, we found a significant correlation between γ and nonverbal intelligence, \( r(74) = 0.39, p = .0005 \) (Figure 6) and, more weakly, with verbal intelligence, \( r(74) = 0.23, p = .045 \). Learning rate from the SM model and performance intelligence were also moderately correlated, \( r(74) = 0.25, p = .03 \). However, when γ, learning rate, and the learning rate x γ interaction were entered into a multiple regression model, they did not predict incremental variance in nonverbal intelligence beyond γ alone, \( F(2, 83) = 1.73, p = .19 \). The relationship between nonverbal intelligence and γ was not moderated by age (\( p = .19 \)) or sex (\( p = .97 \)).

4.3.4. **Entropy of the expected value distribution tunes the explore/exploit tradeoff and is shaped by selective maintenance.**

As illustrated above (Figure 3), selective maintenance of value traces should result in a compressed representation later in learning, as the value of unchosen actions decays toward zero, and entropy decreases. Supporting this prediction, for the selective maintenance model, entropy (information content) was high when participants entered a new contingency, then declined with learning (depicted in
Conversely, for the Fixed LR V model, entropy was much higher on average and remained relatively stable with learning. In a trial-wise multilevel regression of entropy on model, entropy was significantly higher for Fixed LR V than selective maintenance, $t = 246.03, p < 10^{-16}$. In other words, late in learning, selective maintenance reduced memory requirements by 1/3 in poorly performing subjects and by almost 1/2 in well-performing subjects (Figure 7b).

Figure 7. Evolution of value entropy starting from random uniform prior estimates on value. Lines represent the mean entropy across trials, averaging across subjects. Trial-wise entropy was derived from the estimated value distributions of subjects at their best-fitting parameters. Shaded ribbons represent the bootstrapped 95% confidence interval of the mean at each trial. In panel a, dark vertical lines depict boundaries between different
contingencies (50 trials each), explicitly signaled to participants. Panel b depicts the average change in entropy, averaging over subjects and runs (excluding run 1); this represents the typical increase in entropy during initial exploration followed by its decline as high-value actions are discovered and exploited. Better-performing subjects (right panel) exhibit proportionately greater entropy increases early in learning under the selective maintenance model, whereas poorer subjects (left panel) have higher mean entropy. Value traces were carried forward from one block to the next, an implementation that resulted in better fits for both models compared to resetting values in each block (data available upon request). Apart from differences in the first few trials of the experiment, the essential dynamics of entropy under the Fixed LR V and Selective Maintenance models are unchanged if the model is initialized with zero prior estimates on value (see Supplementary Figure 6).

Because entropy tunes the exploit/explore tradeoff (Achbany et al., 2008), we hypothesized that early increases in entropy facilitate exploration and the discovery of valuable actions, whereas entropy declines late in learning enable a shift to exploitation, contributing to foraging success. To test this hypothesis, we estimated the effect of entropy early and late in learning (averages in trials 2–10 and 41–50, respectively) on the total number of points earned in each run in a multilevel model (runs nested within subject; subject treated as random). As predicted, for the selective maintenance model, higher entropy early in learning predicted greater earnings over the run, $t = 5.77, p < 10^{-5}$, whereas entropy late in learning was associated with poorer earnings, $t = -3.85, p < .001$. Furthermore, we anticipated that greater early:late entropy ratios reflect an adaptive transition from exploration to exploitation and would be associated with better performance. We found strong support for this hypothesis: across subjects, higher early:late entropy ratios were associated with greater total earnings on the task, $r(74) = .56, p < .0001$. Crucially, the early:late entropy ratio for the Fixed LR V model (lacking a selective maintenance process), was uncorrelated with performance, $r(74) = -.16, p = .18$ (see Figure 8).
Figure 8. Relationship between early:late entropy ratio and total earnings during the clock task. The early:late entropy ratio was calculated as the quotient of entropy early in learning (trials 2-10) and late in learning (trials 41-50). To account for between-subjects variability in average entropy, for each run, early and late entropy were normalized by the subject’s mean entropy. One extreme high value of the early:late ratio and one participant with low total earnings were Winsorized for plotting and calculations based on regression diagnostics. The correlation between entropy ratio and performance was qualitatively unchanged using the original data: selective maintenance model $r(74) = 0.53$, $p < .0001$. The blue lines denote the least-squares regression line.

The entropy-performance relationship was specific to the selective maintenance model, suggesting that adaptive entropy dynamics are shaped by selective maintenance of value traces. To test this idea, we examined whether the positive association between selective maintenance, $\gamma$, and earnings was mediated by the early:late entropy ratio. Corroborating this account, in a path analysis, the indirect effect of $\gamma$ on earnings via the early:late entropy ratio was significant, $\beta = .33, z = 3.41, p < .001$ ($p$-values based on bootstrapped standard errors; MacKinnon, Fairchild, & Fritz, 2007). Importantly, the direct effect of selective maintenance on performance was nonsignificant ($p = .80$) after accounting for early:late entropy ratio, and the total effect was largely explained by the mediated path, indirect/total effect $= .91$. 

\[ r = -0.16 \]
\[ r = 0.56 \]
Conversely, an alternative model in which individual differences in selective maintenance mediated the relationship between early:late entropy ratios and performance was non-significant, \( z = .26, p = .79 \).

### 4.3.5. Entropy-driven softmax exploration explains RT swings.

Our findings are inconsistent with the idea that RT swings (defined as the absolute change in response time on the current trial compared to the previous trial) reflect shifts toward uncertain options. Drawing on the observation that successful learners rely on early entropy to learn the contingency, however, an alternative hypothesis is that RT swings result from high entropy of the expected value representation. Consistent with this account, greater entropy (computed from the selective maintenance model) predicted larger RT swings in a trial-wise multilevel regression, \( B = 174.00, t = 16.15, p < 10^{-16} \) (subject and run were modeled as random effects). Given that the entropy dynamics are substantially different in the first run (Figure 7), when subjects are still learning the structure of the task, these analyses included runs 2-8, though the pattern of effects is qualitatively unchanged if the first run is included (details available from authors upon request).

Importantly, the positive relationship between entropy and RT swings could not be accounted for by many other factors — distance from the point of maximum estimated value, value of the chosen RT relative to the global maximum, distance from the edge of the time interval, trial, the magnitude of the previous RT swing, and whether the previous action resulted in a reward or omission. Even after controlling for these variables and their interactions, the magnitude of the entropy effect remained essentially unchanged, \( B = 144.04, t = 14.19, p < 10^{-16} \) (depicted in Supplementary Figure 7; see Supplementary Table 3 for all parameter estimates). To rule out the possibility that the entropy-exploration relationship reflected high average levels of entropy (e.g., reflecting consistently random responses), rather than trial-level effects, we decomposed entropy into between-run versus within-run variability and entered both as predictors of RT swings in a trial-wise multilevel regression. Although RT
swings were larger in runs with high average entropy ($t = 16.75, p < 10^{-16}$), high entropy on a given trial (relative to the run mean) predicted larger RT swings on the subsequent trial, and this relationship became stronger as learning unfolded (main effect $t = 3.26, p = .003$; trial x entropy interaction $t = 5.04, p < 10^{-7}$), suggesting a greater role of entropy in continual rather than initial exploration. Interestingly, higher entropy was associated with longer response times, $t = 10.34, p < 10^{-16}$, consistent with the idea that tracking more value information increases cognitive load, slowing response times or promoting indecision.

4.3.6. Uncertainty and exploration.

Although the selective maintenance model was uncertainty-insensitive, the next-best model, Fixed U + V, recovered a negative $\tau$ parameter, indicating uncertainty aversion, for 71 of 76 subjects, $t(75) = -6.9, p < 10^{-8}$. Within the KF family, the KF U + V model dominated (EP = 1, BOR < $10^{-22}$) and also recovered negative $\tau$ parameter values for 67 of 76 subjects, $t(75) = -5.8, p < 10^{-6}$. Exploring the large action space of the clock task, participants can shift their response times substantially from trial to trial, particularly early in learning. RT swings were first described by Frank and colleagues (Badre et al., 2012; Frank et al., 2009), who viewed them as a form of uncertainty-directed exploration in some individuals. An advantage of the SCEPTIC model is that the tendency of subjects to shift toward or away from the moment of maximal uncertainty (i.e., the RT about which the least is known) can be estimated. To test whether response times were related to uncertainty seeking or aversion, in a multilevel model we regressed trial-wise RT on the previous RT, whether the prior response was rewarded, the RT of maximal value, and the RT of maximal uncertainty ($RT_{uncertain}$). Trial-wise value and uncertainty estimates were obtained from the Fixed U + V model using fitted subject parameters. As expected, there was a strong positive relationship between the highest value option and the chosen response time, $t = 38.58, p < .0001$. We also observed a negative association between RTs and the most uncertain option, $t = -3.09, p = .002,$
indicating that subjects were uncertainty averse. Importantly, the effect of uncertainty was moderated by trial, such that subjects were increasingly averse to the most uncertain option later in learning, $RT_{\text{uncertain}} \times \text{trial} = -8.11, p < .0001$. Finally, consistent with the idea that the U + V model captures individual differences in uncertainty aversion, subjects with more negative $\tau$ parameters tended to avoid the most uncertain option to a greater extent, $RT_{\text{uncertain}} \times \tau \ t = 3.19, p = .001$.

### 4.3.7. Within-trial time-varying effects of value and uncertainty.

We employed multilevel Cox models to ascertain that key findings from the selective maintenance model hold even after relaxing certain assumptions and controlling for confounds that arise due to modeling time-dependent responses (see 2.9.5 Mixed-effects survival analyses of behavior). Indeed, within-trial time-varying expected value from the selective maintenance model predicted the timing of responses above and beyond the baseline response hazard function ($z = 46.45, p < .0001$), while time-varying uncertainty from the U + V model was associated with a lower response probability ($z = -58.75, p < .0001$; see Supplementary Table 4 for full model statistics), indicating uncertainty-aversion. In additional multilevel Cox models, we ruled out a number of alternative explanations for the tendency to over-exploit high-value temporal regions predicted by the selective maintenance model. It was not due to the inability to respond early in the interval (first 500 ms) or the avoidance of the last 500 ms out of potential concern of missing a trial (value: $z = 39.79, p < .0001$, U + V uncertainty: $z = -44.69, p < .0001$). Nor were they explained by the magnitude/probability tradeoff in which subjects may prefer a smaller reward that is more frequent, or vice versa. For three of the four contingencies tests (CEV, IEV, and DEV), reward magnitude was high late in the interval, whereas reward probability was low. As detailed above, the later parts of the interval were least sampled by human subjects and therefore, most uncertain on average. In the CEVR condition, however, the magnitude/probability tradeoff was reversed (high probability, low magnitude for longer response times) and yet later time points were still relatively under-sampled (Figure 1c). Crucially, in a multilevel Cox model of analysis of CEVR alone, we observed
the same pattern as the other contingencies: subjects were more likely to respond during high value timepoints (SM value: $z = 20.74, p < .0001$), but were averse to responding when uncertainty was high (U + V uncertainty: $z = -18.70, p < .0001$).

4.3.8. Are selective maintenance of action values and uncertainty aversion proxies for sticky choice?

Using various implementations of choice autocorrelation (Lau & Glimcher, 2005; Schönberg et al., 2007), we ascertained that sticky choice did not account for either selective value maintenance or uncertainty aversion (see Supplementary Results).

4.3.9. Time-Clock (TC) model.

One cannot directly compare fits of the TC model to SCEPTIC and TD because their outputs are of different dimensionality: TC yields a single predicted response time per trial whereas SCEPTIC and TD provide a multinomial choice distribution. We did, however, assess the explanatory power of each parameter in the TC model by fitting model variants with an increasing number of parameters, following the order in the model equation (i.e., varying from one to seven parameters; see Material and Methods). In so doing, we tested whether models that included only descriptive parameters fit substantially worse than models that included RT modulation due to prediction errors, value-based learning, or uncertainty.

Surprisingly, the substantively interesting parameters of TC — expected value of fast versus slow responses ($\rho$), uncertainty-directed exploration ($\varepsilon$), go ($\alpha_c$), and no-go ($\alpha_N$) terms — did not contribute substantially to fits. Rather, a comparison among models revealed a strong preference for a model containing three parameters of no interest: mean RT ($K$), choice autocorrelation ($\lambda$), and RT of maximum reward ($v$); pEP = 1.0, BOR < $10^{-35}$ (Figure 9). In generative simulations using the TC model (based on code from Frank et al., 2009), and using published parameters, we replicated published findings of model
sensitivity to different contingencies, but observed poor performance in longer learning episodes and under moderate variations in parameters (see Supplementary Figure 8).

Figure 9. Random-effects Bayesian model comparison of time-clock (TC) model variants incorporating free parameters incrementally (top to bottom). Each tick on the vertical axis represents the addition of that parameter into a TC model variant containing all parameters above it. Thus, models varied from one to seven parameters. EP = exceedance probability. BOR = Bayesian omnibus risk, a measure of statistical risk in group model comparisons quantifying whether chance is likely to explain differences in estimated model frequencies.

5. Discussion

Reinforcement-based timing involves exploration of a large continuous action space. We aimed to understand (1) how value information may be represented and maintained in this context, given realistic cognitive constraints, and (2) how information maintenance might shape exploration. Humans’ sampling trajectories reflected selective maintenance: actions with high perceived value were updated by sampling and prediction error, whereas infrequently sampled action values decayed. Selective maintenance dynamically shaped the information content (entropy) of the model-estimated value representation.
Upon entering a new environment, entropy was high during initial exploration and declined later in learning. These dynamics were associated with successful performance and intelligence. This was not the case under full maintenance of value traces, where entropy was high throughout learning and did not predict performance. Consistent with softmax exploration based on a Gibbs/Boltzmann distribution (Achbany et al., 2008), entropy controlled exploration, as indexed by response time swings. The idea that information compression by selective maintenance facilitates the transition from exploration to exploitation is new to the study of human decision-making.

By contrast, the notion that humans preferentially explore more uncertain actions in this context was not supported. In simulations, uncertainty-directed exploration yielded a more precise representation of the environment early in learning but conferred no appreciable foraging advantage over softmax exploration, with the possible exception of extremely difficult, discontinuous environments where high-value regions were sparse. Extending prior work on TD models of Pavlovian learning (Ludvig et al., 2008, 2012) we found that a neurobiologically plausible temporal basis function representation accounted well for instrumental reinforcement-based timing. Finally, we did not find evidence of computational mechanisms that might control a dynamic learning rate on the clock task, and models with a fixed learning rate afforded the best fit to behavior.

5.1. Information dynamics: selective maintenance of value traces and its effects on entropy

Under the best-fitting selective maintenance model, the information content (entropy) of value traces is high during the initial exploration of a new environment and declines later in learning, facilitating the shift toward exploitation (Figure 7). These information dynamics are more pronounced in successful subjects, indicating that they describe an adaptive approach. Although the idea that information is lost during adaptive learning may at first seem paradoxical, it fits the intuition that cognitive load is highest when one encounters a new (or altered) contingency and declines when one learns to exploit it.
Consistent with this account, high entropy was associated with longer response times, which parallel the detrimental effects of working memory load on reinforcement-based timing (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007). More generally, the concept of a maximum-entropy information source that emits exploratory actions, as is the case of Boltzmann's softmax with a uniform input, is not new. It directly relates to Borel's so-called infinite monkey theorem (Borel, 1913) — proposed as a comment on Boltzmann’s work — where a million monkeys typing randomly eventually produce volumes that will include “books of any nature” (also see the Library of Babel; Borges, 1998).

In simulations, a selective maintenance strategy was generally as successful as comparators that maintained learned values for all sampled actions. In other words, it typically suffices to track a subset of valuable alternatives without maintaining a detailed representation of the rest of the environment. In many individuals, later in learning, the decision function for the selective maintenance model had a peaked unimodal distribution around the perceived value maximum. Such dynamics suggest that individuals test a hypothesis about the location of the maximum value (i.e., when in time am I most likely to obtain the best outcome?) rather than tracking the expected value of all options (cf. Geana & Niv, 2014). In this framework, uncertainty about the temporal occurrence of maximum reward is encoded implicitly by the entropy of the value distribution (Pouget, Beck, Ma, & Latham, 2013).

Moreover, our results suggest one possible computational mechanism of selective maintenance of value representations. Short-term memory traces decay with time, but can be maintained by a refreshing process, such as rehearsal or memory search (Baddeley & Logie, 1999; Barrouillet et al., 2007; Engle, Tuholski, Laughlin, & A, 1999; Henson, 1998; Page & Norris, 1998). Extending these observations to reinforcement-based timing, we tested a model where (1) regardless of their valence and magnitude, prediction errors enhance the maintenance of action values and that (2) this enhancement is relatively precise in time following a temporal generalization gradient. Thus, prediction error updates, requiring a retrieval of value traces, can serve as a refreshing process similar to explicit rehearsal or memory search.
Similar to the pruning of branching action sequences described by Huys and colleagues (2012), TBFs and selective maintenance exemplify heuristic computational mechanisms for reducing the dimensionality of the environment to make learning tractable. Our findings with respect to cognitive constraints on learning also echo Collins and colleagues (2014), who found that working memory capacity limits learning in a large action space, accounting for reward learning deficits in schizophrenia. Similarly, Otto and colleagues found that the limits of working memory constrained learning in an environment with a sequential structure (Otto et al., 2013).

One may note a homology between our algorithmic selective maintenance model and neural network models with attractor dynamics. Basis function elements can be thought of as tuning curves of excitatory neurons, whereas selective maintenance is homologous to inputs from inhibitory neurons. Thus, inhibitory inputs could stabilize the attractor bump in order to maintain a hypothesis about the global value maximum, a testable hypothesis. Reinforcement-based timing is not entirely abolished by any specific brain lesion (Aparicio, Diedrichsen, & Ivry, 2005; Coslett, Wiener, & Chatterjee, 2010; Meck, Church, & Matell, 2013; Meck, Church, & Olton, 1984; Yin & Troger, 2011), and it is likely that TBF dynamics are found in multiple networks implicated in both timing and reward learning such as the basal ganglia/dopaminergic midbrain, the cerebellum, and the premotor cortex (Klein-Flugge, Hunt, Bach, Dolan, & Behrens, 2011; Merchant, Harrington, & Meck, 2013; Wiener, Turkeltaub, & Coslett, 2010). It is also likely that selective maintenance is an active process mediated by circuits involved in time-based resource allocation, such as the frontoparietal and cingulo-opercular networks (see 5...). (Dosenbach et al., 2007; Smith et al., 2009)

5.2. Uncertainty-directed or entropy-driven exploration?

Intuitively, when choosing among a multitude of actions with uncertain reward values, an agent should benefit from uncertainty-directed exploration, at least early in learning (Badre et al., 2012; Meuleau & Bourgine, 1999). Yet the resulting need to appraise uncertainty further complicates the
intractable exploration/exploitation dilemma (Dayan & Sejnowski, 1996), giving rise to a number of specific problems and objections. Our simulations showed that optimizing a single tradeoff between directed, uncertainty-directed exploration and exploitation, despite yielding a more precise map of the environment early in learning, did not confer an advantage over models with undirected, stochastic exploration. These results likely generalize beyond temporal contingencies to other action spaces with correlated returns along a continuous dimension and may help to explain why humans do not engage in uncertainty-directed exploration in other tasks (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011).

Importantly, cognitive constraints likely limit tracking of uncertainty. If we assume that uncertainty representation decays like other memory traces, representations in the infrequently visited (and most uncertain) regions of the action space would be most subject to decay, causing the uncertainty estimate to revert toward a prior (cf. Payzan-LeNestour & Bossaerts, 2011) and hindering uncertainty-directed exploration. The total loss of information about unsampled actions given a fixed representational capacity and a decay of uncertainty estimates also scales with the size of the action space.

Although uncertainty-directed exploration can be efficient initially in complex environments (e.g., large spatial landscapes or discontinuous environments with sparse high-value regions; Makarenko et al., 2002), the tasks used in studies of human value-based decision-making typically emphasize continual exploration, where one exploits the perceived contingency and periodically explores to adjust the policy to the environment. Our results suggest that an implicit representation of global uncertainty expressed in the entropy of action probability distribution suffices for continual exploration. That said, things change in sparse, discontinuous environments in which only a small fraction of possible actions are reinforced. In additional simulations (Supplementary Figures 3 and 4) of such environments, models incorporating uncertainty-guided exploration gained a small advantage over value-guided choice alone. At the same time, these very challenging environments could not be learned by TD probably because of their temporal discontinuity, which defeats value back-propagation. Considering that and the human
performance on the much simpler monotonic contingencies (Figure 1), we doubt that an average person can learn sparse, discontinuous contingencies reliably even given hundreds of trials.

5.3. Basis Function Representation of a Temporal Contingency

The representation of value over time involves a tradeoff between the generality of representation on one hand and the number of free parameters or values stored on the other. A completely general temporal value representation is exemplified by complete serial compound (CSC) in TD. On the other end we find parsimonious parametric models such as Frank’s TC. Albeit elegant, parametric solutions often turn out to explain a narrow range of phenomena; they break down more easily at boundary conditions. On the clock task, such misspecification occurs, for example, when the TC model is confronted with a non-monotonic contingency (Figure 4). Gaussian temporal basis function representation, in our opinion, finds the middle ground between these two extremes: it reduces the memory and computational load compared to TD, while maintaining generality of representation, which enables it to learn virtually any contingency in one continuous dimension.

The SCEPTIC model represented expected value using several radial basis functions with contiguous and overlapping temporal receptive fields. We note that discrete Gaussian elements were employed here for computational convenience, and we make no claims about the superiority of this solution over alternatives such as cosine basis (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012). Extending earlier findings in Pavlovian conditioning (Ludvig et al., 2008, 2012), we found that TBF representations afforded a better fit to human instrumental behavior on the clock task than TD with CSC. TD has been previously reported to fail on the clock task due to erroneous back-propagation of value from later to earlier points in the interval (Moustafa et al., 2008). Importantly, our TD model with a task-specific state space partition successfully overcame this problem, as indicated by its foraging success. We can thus be confident that SCEPTIC’s superiority is specifically due to the combination of TBF value
representation and the strategic consideration of the entire interval in the choice rule. Like Ludvig, Sutton, and Kehoe (2008, 2012), we did not attempt to estimate the functional form, number, and placement of TBF elements, fixing them at reasonable priors. New experiments are needed to constrain the parameterization of TBF with behavioral and physiological data. One testable hypothesis — articulated but not tested by Ludvig and colleagues — is that the density of elements decreases progressively with the passage of time within the interval, which would reflect Weber’s law (Fiorillo et al., 2008). This question relates to the broader unresolved problem of interval adaptation (Fiorillo et al., 2008; Merchant et al., 2013), or in computational terms, how the agent learns to optimally place the TBF elements based on time intervals experienced in a particular context.

5.4. Selective maintenance and entropy-driven exploration as boundedly optimal strategies

Information compression under selective maintenance and the resulting entropy-driven exploration policy harness limited computational and memory resources to solve a difficult problem. The value of computation (VOC) theory applied to human cognition suggests that we allocate cognitive resources to achieve the best tradeoff between the expected utility of the solution and the opportunity cost of time required for computation (Lieder et al., 2014; Shenhav et al., 2017). This opportunity cost grows under time pressure (e.g. in DEV), whereas we demonstrate that decision times increase under a higher information load. Furthermore, our computational studies suggest that the expected utility gains achieved by the more resource-intensive strategies (TD, full maintenance, uncertainty-directed exploration) are often trivial, favoring the SCEPTIC selective maintenance model. Yet, the tradeoff may be different in small action spaces demanding lower computation times, such as two-armed bandits (Gershman, 2018; Wilson et al., 2014), favoring memory-intensive uncertainty-directed exploration.

An intuitive prediction of VOC is that information compression (as in selective maintenance or disregard of local uncertainty) becomes increasingly advantageous as the computation time scales up
with the number of potentially valuable options, captured by entropy. Consequently, the entropy of value representation could serve as a neural signal triggering compression. One putative neural mechanism for such control would involve entropy tracking by a region sensitive to cognitive load (i.e. lateral prefrontal cortex, lPFC; (McGuire & Botvinick, 2010; Wang, Trongnetrpunya, Samuel, Ding, & Kluger, 2016)) and relaying this signal to a region that performs control actions implementing compression (i.e. dorsal anterior cingulate cortex, dACC; (Heilbronner & Hayden, 2016; Shenhav et al., 2017)).

5.5. Limitations

While our data support entropy-driven continual exploration on the clock task, they do not rule out the possibility that initial exploration is driven by uncertainty seeking. That said, we tested a model variant (KF U → V) designed to switch from initial exploration to later exploitation, which was inferior in model comparisons (details in Supplementary Materials). The time course of exploration (Figure 1) and information dynamics (Figure 7) suggest that the period of initial exploration on the clock task may be as short as 5-8 trials. One possibility is that humans track uncertainty explicitly during the first few trials in order to seed potentially valuable actions in the softmax function, essentially defining a subset of eligible actions for refinement in continual exploration. Further, in environments where participants are explicitly told that they will repeatedly sample a stable contingency (i.e., there is a longer horizon for learning), they are more likely to use directed exploration to identify the best action than when they have only a single choice (Wilson et al., 2014). More speculatively, the cognitive load of an explicit uncertainty representation combined with a high-entropy value representation may be overwhelming early in learning. On the other hand, the shift toward continual exploration using a Boltzmann strategy may ease working memory demands. We note that time is a unique action space (Merchant et al., 2013), and we cannot know to what extent our findings generalize to other continuous, multidimensional, or discrete spaces. That said, the results of our survival analyses are reassuring in that selective maintenance
dynamics do not appear to be explained by trial-invariant determinants of response times, psychomotor speed constraints, or the pressure to respond before a trial ends. Our null findings with respect to dynamic learning rates should be taken with caution since there are a number of other approaches to volatility-modulated learning rates that could be potentially adapted to SCEPTIC (Behrens et al., 2007; Iglesias et al., 2013; Payzan-LeNestour & Bossaerts, 2011). Finally, our analyses did not address the issues of opportunity cost and the intrinsic cost of waiting (Drugowitsch et al., 2012), important independent influences on behavior, which will need to be examined in the context of the clock task.

5.6. Conclusions

In contrast to previous proposals for resolving the exploration/exploitation dilemma at policy level, we show that this can be also accomplished at the level of updating and maintaining the value representation. In this study of reinforcement-based timing a simple selective value maintenance strategy reduced information load and facilitated the transition to exploitation, further work is needed to evaluate alternative information maintenance strategies and their putative neural implementations. At the same time, our findings broadly align with emerging evidence that the implementation of optimal inference is limited by representational capacity (Collins & Frank, 2012) and that heuristic approaches to reinforcement learning are more likely to be effective in complex environments (Geana & Niv, 2014; Niv et al., 2015).
Note. The codes and data for all analyses reported in this paper are publicly available at:
https://github.com/DecisionNeurosciencePsychopathology/temporal_instrumental_agent (Hallquist, Dombrovski, & Wilson, 2018)

Acknowledgements. The authors thank Michael J. Frank for helpful comments on validation and comparison of reinforcement learning models, as well as codes for the experimental paradigm and time-clock model. We thank Jonathan Wilson for help with figure preparation and model implementation. We are grateful to Peter Molenaar for useful suggestions regarding adaptive filtering and autoregressive processes. We appreciate Yael Niv’s comment about the role of basis representation in our learning rule. We also thank Michael Woodford for his suggestion of examining information theory measures in behavioral analyses. We thank Beatriz Luna for help with data collection and quality assurance.

This research was performed, in part, using resources and the computing assistance of the Pennsylvania State University Institute for CyberScience Advanced CyberInfrastructure (ICS-ACI).
References


Borel, É. (1913). La mécanique statique et l’irréversibilité. *Journal de Physique Théorique et Appliquée,*


Supplementary Material for

Selective Maintenance of Value Information Helps Resolve the Exploration/Exploitation Dilemma

Hallquist, M. N. & Dombrovski, A. Y.
Supplementary Methods

Numerical estimation of parameter in computational experiments

We chose to use the ga (genetic algorithm) function in MATLAB for optimization because the cost function for reinforcement learning models is sometimes discontinuous due to the discrete nature of choices in output (observation) function. That is, small changes in parameters such as a learning rate can lead to the same set of predicted discrete choices (e.g., 50 trials in a run) and, thus, the same overall performance of the model. This complicates parameter estimation because standard algorithms based on the parameter gradients can obtain false convergence due to local minima in the cost function. The genetic algorithm performs better in such circumstances because it spawns optimizations in many parts of the parameter space and use survival heuristics to keep parameter sets that have lower costs (iterating until convergence). This approach tends to avoid the local minimum/discontinuous cost problem in reinforcement learning models.

Extension of RBFs beyond the bounds of the time interval

In order to approximate a function over a finite interval, \( t \in [0, T] \), a set of RBFs spaced equally over the interval are commonly used. That said, function approximation over finite intervals using basis functions (of any sort) falls prey to interpolation errors near the bounds of the interval such that interpolated data diverge from the true underlying function and basis function weights take on implausible values. This is called the Runge phenomenon and has been well characterized in mathematics (Boyd, 2010a). One approach to mitigate this problem is to extend the centers of the RBFs, \( \mu_b \), slightly beyond the bounds of the interval of interest (Boyd, 2010b). Here, we spaced the RBF centers evenly over
the expanded interval $t \in [0 - .1T, T + .1T]$, a 10% expansion, but expansions between 2% and 20% yielded highly similar results.

$$
\mu_b := (b - 1) \frac{T(1 + 2D)}{B - 1} - DT, \quad b \in \{1, 2, ..., B\}, \quad D = 0.1
$$

**Truncated Gaussian basis**

In addition, to obtain accurate estimates of the value function near the edge of the time interval, basis weight updates applied according to the learning rule were computed using truncated Gaussian basis functions and a truncated temporal generalization function. Basis functions centered near the middle of the time interval tend to receive larger weight updates than those at the periphery because the temporal generalization function of both fast and slow response times often overlaps the receptive field. By contrast, consider a basis function centered at a time of zero seconds. Because response times must always be larger than zero, updates to this basis function will only occur for fast, positive responses, but not negative response times that would also overlap its receptive field. This asymmetry leads basis functions near the edge of the interval to systematically underestimate the true value of the function, which would bias choices predicted by the SCEPTIC model. In order to mitigate this problem, when applying weight updates due to prediction error (or reduction in uncertainty), we renormalized all RBFs to have an equal area under the curve (where a basis function in the center was used as the reference). This correction is identical to using truncated Gaussian functions, where the distribution is bounded on a finite interval $[a, b]$:

$$
\psi(\mu, \sigma, a, b; x) = \begin{cases} 
0 & \text{if } x \leq a \\
\frac{\phi(\mu, \sigma^2; x)}{\Phi(\mu, \sigma^2; b) - \Phi(\mu, \sigma^2; a)} & \text{if } a < x < b \\
0 & \text{if } b \leq x
\end{cases}
$$
where \( \Phi(\mu, \sigma^2) \) is the cumulative distribution function of the Gaussian distribution. In simulations of random foraging across numerous contingencies, updates using a truncated basis recovered a high fidelity representation of the environment: average correlation of true and estimated values = 0.97, \( SD = .01 \); average correlation of proximity to the interval edge with misestimation of value = .01, \( SD = .02 \), whereas the standard Gaussian basis underestimated values at the periphery (details available upon request).

**Alternative choice rule: \( \epsilon \)-greedy**

Although SCEPTIC variants primarily employed a softmax choice rule, we also considered an \( \epsilon \)-greedy model in which with probability \( \epsilon \) (a free parameter), the agent chose a random uniform response among potential options, and with probability \( 1 - \epsilon \), the agent selected the response associated with the highest expected value, \( \text{argmax}(V(i)) \).

**Alternative model of uncertainty-guided choice: KF U \( \rightarrow \) V**

In addition to the U + V policy that weights uncertainty (U) and value (V) in action selection, we considered an alternative policy, KF U \( \rightarrow \) V, that shifts from exploratory choices (U-driven) early in learning to exploitative choices (V-driven) as uncertainty about the environment declines. More specifically, the choice rule follows a two-parameter logistic function that estimates the probability of an exploratory choice on trial \( i \):

\[
p(\text{explore}(i)) = \frac{1}{1 + \exp \left[ -d(u(i) - (1 - \zeta)u_0) \right]}
\]
where \( u(i) \) is the estimated total uncertainty represented by the area under the curve of the \( U \) function and \( d \) is a discrimination parameter that controls the sharpness of the transition from exploratory to exploitative choices (i.e., the steepness of the sigmoid). The term \((1 - \zeta)u_0\) represents the proportion reduction in total uncertainty (relative to the prior) for which the probability of exploration is 0.5 (i.e., the indifference point between exploration and exploitation), where \( \zeta \) is a free parameter between zero and one and \( u_0 \) is the total prior uncertainty as defined above. Under this policy, the agent chooses the most uncertain option, \( \text{argmax}(U(i)) \), for exploratory choices and the most valued option, \( \text{argmax}(V(i)) \), for exploitative choices.

**Maintaining sensitivity to perceived nonstationary contingencies (unexpected uncertainty) and uncertainty about expected value (estimation uncertainty) under the KF.**

As well-described in the engineering and adaptive filtering literature (Goodwin & Sin, 1984), a noteworthy limitation of the basic KF is that as posterior estimates of uncertainty decline with sampling and the effective learning rate shifts toward zero, the filter becomes insensitive to new feedback. Thus, to maintain sensitivity to changing contingencies or perceived volatility, we tested two adaptations of the SCEPTIC model, KF Process Noise and KF Volatility, that build on previous work on dynamic enhancement of learning rates (Behrens, Woolrich, Walton, & Rushworth, 2007; Mathys, Daunizeau, Friston, & Stephan, 2011; Payzan-LeNestour & Bossaerts, 2011).

**SCEPTIC Process Noise KF.** The first variant modulates the process noise of the filter to represent changing dynamics of the underlying system (a variant of noise-adaptive filtering; (Stengel, 1994)). More specifically, process noise in KF models represents uncertainty about the representation of the
underlying system (here, the reinforcement contingency), which may reflect changing dynamics (e.g., contingency reversal) or nondeterministic observations (e.g., probabilistic reinforcement). In the context of reinforcement learning, prediction errors are the closest analogue of process noise because they signify an unexpected outcome that may signal the need to pay greater attention to subsequent outcomes (analogous to the Pearce-Hall model). Here, we adapt the SCEPTIC KF gain calculation to be influenced by preceding prediction errors:

\[ k_b(i) = \frac{\sigma_b(i)^2 + Q_b(i)}{\sigma_b(i)^2 + Q_b(i) + \sigma_{rew}^2} \]

where process noise is defined as the absolute value of the prediction error on the previous trial, \( \delta_b(i - 1) \), scaled by a sensitivity parameter, \( \omega \):

\[ Q_b(i) = \omega |\delta_b(i - 1)| \]

This parameterization enhances the gain (learning rate) on a trial in proportion to the magnitude of the preceding prediction error. Notably, increasing the gain results in larger updates to the estimates of expected value and decreases in posterior uncertainty estimates.

**SCEPTIC Volatility KF.** The second model with a dynamic learning rate injects noise into posterior variance (uncertainty) estimates via the variance update rule. More specifically, the volatility model tracked a zero-reverting AR(1) volatility process according to the history of prediction errors:

\[ z(i + 1) = \kappa z(i) + \phi |\delta(i)| \]
where \( \kappa \) varies between zero and one and represents an autoregressive process (akin to learning rate) that influences the degree to which recent versus remote prediction errors influence estimated volatility, and \( \phi \) scales the influence of the prediction error relative to the posterior variance estimate. Here, the scalar prediction error \( \delta(i) \) is summed across all basis functions to encode its magnitude irrespective of timing. Thus, an agent tracking volatility in this way would be sensitive to a series of large prediction errors according to the following modified variance update rule:

\[
\sigma_b(i + 1) = [1 - e_b(i)k_b(i)]\sigma_b(i) + z(i)
\]

Note that because volatility is tracked globally by the model, irrespective of basis function, the posterior update injects an equal amount of volatility-related noise into each basis, effectively increasing uncertainty about all actions. In turn, this dynamically enhances the learning rate when volatility is high.

**Sinusoidal Contingencies for Testing Model Performance in Simulations**

The master contingency used to test model performance in simulations was defined by two sinusoidal functions evaluated over 5-second trials (Figure 4a):

\[
EV(t) = 10 \sin \left( \frac{2\pi t}{T} \right) + \frac{5}{2} \sin \left( \frac{4\pi t}{T} \right) + 2 \cos \left( \frac{8\pi t}{T} \right)
\]

\[
P = 25 \cos \left( \frac{2\pi t}{T} \right) + 10 \cos \left( \frac{6\pi t}{T} \right) + 6 \sin \left( \frac{10\pi t}{T} \right)
\]
where \( t \) represents the time step within a trial and \( T=500 \) (10ms bins). We rescaled the probability function onto the interval \( p = 0.3–0.7 \) according to the following transformation:

\[
p(r|t) = \frac{0.4(P - \min(P))}{\max(P) - \min(P)} + 0.3
\]

and magnitude was defined as

\[
m(r|t) = \frac{EV(t)}{p(r|t)}
\]

In order to ensure that model parameters did not reflect a peculiar temporal feature of these functions, agents were exposed to 100 variants of the sinusoidal contingency that were shifted only in phase, but not amplitude. Phase shifting was employed in order to 1) test a set of distributions whose maximum value did not consistently fall in one part of the interval, and 2) to equate the total area under the curve of the value function in order to achieve equal weighting across runs in the parameter optimization. The phase-shift approach also led to some environments where optimal responses fell near the beginning or end of the interval and other environments where the best response fell in the middle.

**Additional Details about the Frank TC model**

As noted, the main text, in the TC model, the primary equation for predicting subjects’ response times (RTs) is:

\[
\hat{RT}(i) = K + \lambda RT(i - 1) + v[RT_{best} - RT_{avg}] - Go(i) + NoGo(i) +
\]
\[ \rho [\mu_{\text{slow}}(i) - \mu_{\text{fast}}(i)] + \varepsilon [\sigma_{\text{slow}}(i) - \sigma_{\text{fast}}(i)] \]

In this model, \( i \) represents trial, \( K \) represents the average response time when all other signals are zero (akin to an intercept in multiple regression), and \( \lambda \) represents the degree to which response times follow a first-order autoregressive process. The model also tracks the response time of the largest reward thus far in the run, \( \text{RT}_{\text{best}} \), and is provided with the average response time over all trials, \( \text{RT}_{\text{avg}} \). The extent to which individuals modulate their RTs toward the best response time (relative to their average) is scaled by \( \nu \). These three parameters are more descriptive and have not been linked to reinforcement learning or neurobiology in previous work.

By contrast, the Go and NoGo terms of the model are thought to reflect the effects of striatal dopamine subpopulations on response times after reward prediction errors. More specifically, the Go term captures decreases in response times following positive prediction errors that may be mediated by D1 receptors in the nigrostriatal pathway, whereas the NoGo term captures slower responses after negative PEs via D2 receptors in the striatopallidal pathway. The contribution of these signals to the response time is updated according to prediction errors:

\[ \text{Go}(i + 1) = \text{Go}(i) + \alpha_G \delta_+ (i) \]

\[ \text{NoGo}(i + 1) = \text{NoGo}(i) + \alpha_N \delta_- (i) \]

where \( \alpha_G \) and \( \alpha_N \) are free parameters between zero and five that scale the effect of positive prediction errors (PPEs) on response speeding and negative prediction errors (NPEs) on response
slowing, respectively. Prediction errors are defined in terms of whether the outcome — occurrence versus non-occurrence of reward — is better than the expected value, $V$, of completing a trial $i$:

$$V(i + 1) = V(i) + \kappa_{\text{rew}} \delta(i)$$

$$\delta(i) = r(i) - V(i)$$

where $r(i)$ is the trial reinforcement (0/1) and $\kappa_{\text{rew}}$ is a learning rate that controls the speed of value learning, fixed to 0.1 in our analyses and in previous papers using this model. Thus, prediction errors affect the Go and NoGo terms depending on their valence (positive versus negative):

$$\delta_+(i) = \begin{cases} 
\delta(i) & \text{if } \delta(i) > 0 \\
0 & \text{otherwise}
\end{cases}$$

$$\delta_-(i) = \begin{cases} 
\delta(i) & \text{if } \delta(i) < 0 \\
0 & \text{otherwise}
\end{cases}$$

One of the key insights of the TC model is that individuals may track the value statistics of different response options as a function of time within trial. The TC model makes a simplifying assumption that individuals track information about the probability of a PPE separately for fast versus slow responses within the 4-second trial window, which are represented using two beta distributions, the conjugate prior for a binomial process. Distributions for fast and slow responses track both the expected probability of a PPE (i.e., the mean of the distribution), as well as uncertainty about the outcome (i.e., the variance) for fast and slow responses. For example, if a PPE occurs for a fast response, the mean of the corresponding beta distribution will increase (i.e., a higher expectation) while its variance
decreases (i.e., a reduction in uncertainty). The shapes of these distributions are updated via their shape parameters, \( \alpha \) and \( \beta \), according to Bayes’ rule. For example, if a PPE occurs for a fast response on trial \( i \), \( \alpha_{\text{fast}} \) is incremented by 1: \( \alpha_{\text{fast}}(i) = \alpha_{\text{fast}}(i - 1) + 1 \), whereas if a negative prediction error occurs, \( \beta_{\text{fast}} \) is incremented by 1: \( \beta_{\text{fast}}(i) = \beta_{\text{fast}}(i - 1) + 1 \). This follows from the fact that the mean of a beta distribution is proportional to the ratio of alpha and beta:

\[
E[X] = \frac{1}{1 + \beta/\alpha}
\]

In order to decide whether a given RT is classified as fast or slow, the TC model compares it against the average response time for recent trials, \( RT_{\text{locavg}} \), which is updated with each trial:

\[
RT_{\text{locavg}}(i) = RT_{\text{locavg}}(i) + \kappa_{RT}[RT(i) - RT_{\text{locavg}}(i)]
\]

where \( \kappa_{RT} \) is an update rate fixed to 0.1. This response boundary determines which beta distribution is updated by prediction error:

\[
\alpha_{\text{fast}}(i) = \alpha_{\text{fast}}(i - 1) + \begin{cases} 
1 & \text{if } RT(i) < RT_{\text{locavg}} \text{ and } \delta(i) > 0 \\
0 & \text{if } RT(i) \geq RT_{\text{locavg}} \text{ or } \delta(i) \leq 0
\end{cases}
\]

\[
\beta_{\text{fast}}(i) = \beta_{\text{fast}}(i - 1) + \begin{cases} 
1 & \text{if } RT(i) < RT_{\text{locavg}} \text{ and } \delta(i) < 0 \\
0 & \text{if } RT(i) \geq RT_{\text{locavg}} \text{ or } \delta(i) \geq 0
\end{cases}
\]

\[
\alpha_{\text{slow}}(i) = \alpha_{\text{slow}}(i - 1) + \begin{cases} 
1 & \text{if } RT(i) > RT_{\text{locavg}} \text{ and } \delta(i) > 0 \\
0 & \text{if } RT(i) \leq RT_{\text{locavg}} \text{ or } \delta(i) \leq 0
\end{cases}
\]
\[
\beta_{\text{slow}}(i) = \beta_{\text{slow}}(i - 1) + \begin{cases} 
1 & \text{if } RT(i) > RT_{\text{locavg}} \text{ and } \delta(i) < 0 \\
0 & \text{if } RT(i) \leq RT_{\text{locavg}} \text{ or } \delta(i) \geq 0
\end{cases}
\]

The final two signals of the TC model, \( \rho \) and \( \varepsilon \), use the means and variances of these beta distributions to capture response time variations linked to learned value and uncertainty, respectively. More specifically, the model predicts that individuals modulate their response times in proportion to differences in the mean expected values for fast versus slow responses, \( \mu_{\text{fast}} \) and \( \mu_{\text{slow}} \), respectively, scaled by \( \rho \). Furthermore, the model predicts that if individuals use relative uncertainty about responses to guide choice, then differences in the variances of the two beta distributions, \( \sigma_{\text{fast}} \) and \( \sigma_{\text{slow}} \), can be used to shift toward more uncertain options, scaled by \( \varepsilon \). Note that the contribution of uncertainty-driven exploration is nullified on a given trial if the agent explored in that direction on the previous trial. For full details about the TC model, see Frank and colleagues (10).
Supplementary Results

Parameter Identifiability in Simulated Environments

To have confidence that model parameters are reliably estimated and interpretable in empirical data, where the data generation process and parameters are unknown, it is crucial to test that known and estimated parameters converge in simulations. Model parameters were considered identifiable if they were reliably recovered ($R^2 > .70$). In simulations, SCEPTIC variants reliably recovered their parameters (all $R^2 > .93$; see Supplementary Figure 1). TD recovered its exploration parameter, $\xi$ ($R^2 = .88$), but not the learning rate, $\alpha$ ($R^2 = .14$). The TC model did not reliably recover any of its parameters (all $R^2 < .25$). TC parameter identifiability for prediction error-related response time (RT) modulation ($\alpha_c$ and $\alpha_N$), relative value ($\rho$), and relative uncertainty ($\varepsilon$) improved markedly (all $R^2 > .86$) by fixing the parameters of no interest ($K$, $\lambda$, and $\nu$), but as reported in the main text, these parameters were most influential in fitting subjects’ behavior. We included the unstable models in further tests for reference, but note that some of their parameters are not identified.

Discriminability of the Selective Maintenance Model from Alternatives

As detailed in the main text, Bayesian model comparison provided strong evidence that the selective maintenance model characterized human behavior better than alternatives. Yet it is nevertheless possible that (a) the selective maintenance model might simply be more flexible than other models, allowing it to fit data from any number of generative processes; or (b) that behavior generated from the selective maintenance model is not substantially different from other SCEPTIC variants and could be readily fit by them. Such questions are impossible to discern definitively in empirical data, where the generative processes are unknown, but can be tested in simulations.
To bolster our confidence in the discriminability of the selective maintenance model from alternatives, we simulated data from TD and SCEPTIC models using plausible parameters from human subject fits. Simulated datasets for 120 “subjects” consisted of 50 trials in sinusoidal value contingencies (Figure 4a). We then fit these data using TD and all SCEPTIC variants using the same procedure as our empirical data. In Bayesian model comparisons of data generated by alternative models, the selective maintenance model was not chosen as the best model in any case (estimated model frequencies ranging from .10 to .30, exceedance probabilities < .02), addressing concern (a). However, for datasets simulated from the selective maintenance model using values of \( \gamma \) ranging from .16–.48 (consistent with estimates from human behavior), the selective maintenance model was consistently selected above alternatives (estimated model frequency = .84; EP = 1.0; BOR < 10^{-38}), addressing concern (b). Altogether, these model recoverability analyses indicate that data generated by a selective maintenance model are discriminable from the other generative models tested and that the selective maintenance model is not prone to fitting data from alternative models.

**Potential Advantages of Uncertainty-Guided Exploration in Sparse Reward Contingencies**

We set out to find the boundary conditions where entropy-driven exploration would no longer suffice and also to ascertain that uncertainty-driven exploration can be effective in the SCEPTIC framework. To that point, both the sinusoidal contingencies used to test model optimality and the monotonic contingencies tested in human subjects vary smoothly in time and often had a single global maximum. With a sufficient number of trials, global maxima in such environments could be identified through local exploration with gradual shifts in response times. A sparser environment in which only a few actions are reinforced with little continuity among adjacent values might require the incorporation of
uncertainty to uncover highly valued options and not to get “stuck” in local maxima\(^1\). In order to test this possibility, we simulated environments containing two relatively narrow value peaks with maximum values of 50 and 100 points, whereas other action values varied randomly between approximately 5 and 15 points (see Supplementary Figure 3). On average, only about 7% of the interval contained values exceeding 50 points, which would be considered optimal. These contingencies would be difficult to test in human experiments, since the aforementioned challenges would necessitate a very large number of trials. Using the same methodology as our initial foraging simulations, we identified five optimal parameter sets for each model and simulated the proportion of possible points earned across 100 novel contingencies, randomly permuting the location of the value peaks.

In this environment, models incorporating uncertainty in the choice policy (i.e., Fixed U + V and KF U + V) outperformed models guided by value alone (Fixed LR V, Selective Maintenance, and Kalman V), \(adj. ps < .001\) (Supplementary Figure 4). Corroborating the account that uncertainty-guided exploration is instrumental in solving such sparse contingencies, the \(\tau\) parameter weighting the relative contribution of uncertainty in the decision function was positive, average Kalman U + V \(\tau = .19\) \((SD = .11)\), \(t = 17.32\); average Fixed U + V \(\tau = .09\) \((SD = .06)\), \(t = 17.00\). Interestingly, at trial lengths of 60 and 110, the Selective Maintenance model performed better than the Fixed LR V and KF V models \((adj. ps < .03)\), suggesting the possibility that maintaining a subset of valued actions may also be useful in sparse environments, perhaps to avoid inferior parts of the action space. The TC and TD models performed poorly in this environment, earning only 20-30% of possible points, though TC was consistently better than random exploration alone.

\(^1\) We thank an anonymous reviewer for pointing out this possibility.
Are selective maintenance of action values and uncertainty aversion proxies for sticky choice?

One might argue that selective maintenance is confounded with choice autocorrelation or sticky choice, a residual response tendency unaccounted for by the substantive parts of a reinforcement learning model. Sticky choice has been previously described on armed bandit tasks (Lau & Glimcher, 2005; Schönberg, Daw, Joel, & O’Doherty, 2007) and the clock task (Badre, Doll, Long, & Frank, 2012). The cardinal difference between selective maintenance and sticky choice, however, is that selective maintenance results from the interaction between choice history and reinforcement history, whereas choice autocorrelation does not (Lau & Glimcher, 2005). Thus, after accounting for sticky choice, the selective maintenance model would dominate over alternatives only if it truly reflected a decaying representation of the reinforcement history. Similarly, an argument can be made that what appears to be uncertainty aversion is simply a manifestation of sticky choice.

We tested a simple AR(1) choice auto-correlation function (ACF) as well as a multi-trial ACF where effects of each previous choice decay exponentially as a function of the lag (Schönberg et al., 2007). In a comparison of model families (AR[1] ACF vs. multi-trial ACF vs. no sticky choice, each implemented for all SCEPTIC models), AR(1) ACF dominated ($EP = 1.0, BOR < 10^{-42}$). In the same analysis, the selective maintenance model with AR(1) choice ACF dominated among the 21 models ($EP = .99, BOR < 10^{-42}$). With both ACFs, uncertainty-sensitive models (Fixed LR U+V and KF U+V) recovered a negative parameter $\tau$, indicating uncertainty aversion ($t_{75} < -6.6, p < 10^{-8}$). Given the observations of response time swings on the clock task (Badre et al., 2012), we considered the opposite possibility, that negative choice anti-correlation or response alternation (Lau & Glimcher, 2005) mimics as uncertainty-seeking. However, allowing for negative anti-correlation in the best-fitting model (AR[1] with temporal generalization) resulted in worse fits and did not change the estimate of uncertainty aversion substantially, $t_{75} = -7.5, p$
< 10^{-10}. Thus, sticky choice did not account for either value selective maintenance or uncertainty aversion.
Supplementary Tables

*Supplementary Table 1.* Prior distributions for Frank TC model implementation in VBA.

<table>
<thead>
<tr>
<th>TC parameter</th>
<th>Prior value</th>
<th>Parameter distribution</th>
<th>Lower bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>2000</td>
<td>Uniform</td>
<td>0</td>
<td>4000</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.5</td>
<td>Sigmoid</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\nu$</td>
<td>0.1</td>
<td>Sigmoid</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>$\alpha_G$</td>
<td>2.5</td>
<td>Sigmoid</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$\alpha_N$</td>
<td>2.5</td>
<td>Sigmoid</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$\rho$</td>
<td>1300</td>
<td>Gamma(2,2)*400</td>
<td>0</td>
<td>~10000</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>693</td>
<td>Exponential($\beta = 1000$)</td>
<td>0</td>
<td>~10000</td>
</tr>
</tbody>
</table>
**Supplementary Table 2. Value of all fixed structural model parameters used in computational experiments**

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed structural parameter</th>
<th>Value in computational experiments</th>
</tr>
</thead>
</table>
| **All SCEPTIC variants**  
(Fixed LR V, Fixed LR U+V,  
Fixed LR selective maintenance,  
Kalman filter softmax, Kalman  
filter U+V) | Number of basis functions ($b$) | 24 |
| | Variance of basis functions ($s_b^2$) | 17921.2 ms$^2$ |
| | Variance of temporal generalization function ($s^{\gamma}_{\theta}$) | 2376.6 ms$^2$ |
| **SCEPTIC variants with explicit uncertainty representation**  
(Kalman filter softmax,  
Kalman filter U+V, Fixed LR  
U+V) | Measurement noise of the Kalman filter ($\sigma_{\text{rew}}^2$) | DEV: 1445.2  
IEV: 1110.2  
CEV: 1101.6  
CEVR: 2830.1 |
| **Time-clock (TC)** | Number of beta distributions to map time | 2 (fast and slow) |
| | $RT_{\text{avg}}$ | Precomputed mean of subject’s RTs in the run |
| | $\kappa_{\text{rew}}$ | 0.1 |
| | $\kappa_{\text{RT}}$ | 0.1 |
| **Temporal difference (TD), Q-learning** | Number of timesteps in complete serial compound representation | 40 |
| | Discount factor ($\gamma$) | 0.99 |
### Supplementary Table 3

Change in response time (aka RT swing) as a function of entropy and alternative explanatory variables

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>692.51 (27.97)**</td>
<td>854.60 (18.23)**</td>
</tr>
<tr>
<td>entropy (i - 1)</td>
<td>174.00 (10.77)**</td>
<td>144.04 (10.15)**</td>
</tr>
<tr>
<td>Trial (i)</td>
<td>-1.42 (0.31)**</td>
<td></td>
</tr>
<tr>
<td>RT swing (i - 1)</td>
<td>14.00 (0.62)**</td>
<td></td>
</tr>
<tr>
<td>$V_{\text{max}} (i - 1) - V_{\text{chosen}} (i - 1)$</td>
<td>6.94 (0.67)**</td>
<td></td>
</tr>
<tr>
<td>Reward (i - 1)</td>
<td>-291.12 (8.94)**</td>
<td></td>
</tr>
<tr>
<td>$RT_{\text{chosen}} (i - 1) - RT_{\text{Vmax}} (i - 1)$</td>
<td>7.63 (0.84)**</td>
<td></td>
</tr>
<tr>
<td>dist. from edge (i - 1)</td>
<td>-16.26 (1.37)**</td>
<td></td>
</tr>
<tr>
<td>entropy (i - 1) x Trial</td>
<td>1.02 (0.40)**</td>
<td></td>
</tr>
<tr>
<td>$V_{\text{max}} - V_{\text{chosen}} x Reward (i - 1)$</td>
<td>0.44 (0.83)</td>
<td></td>
</tr>
<tr>
<td>$V_{\text{max}} - V_{\text{chosen}} x RT_{\text{chosen}} (i - 1) - RT_{\text{Vmax}} (i - 1)$</td>
<td>0.15 (0.06)**</td>
<td></td>
</tr>
<tr>
<td>Reward (i - 1) x $RT_{\text{chosen}} (i - 1) - RT_{\text{Vmax}} (i - 1)$</td>
<td>-9.41 (1.10)**</td>
<td></td>
</tr>
<tr>
<td>$V_{\text{max}} - V_{\text{chosen}} x \text{dist. from edge} (i - 1)$</td>
<td>-0.42 (0.13)**</td>
<td></td>
</tr>
<tr>
<td>Reward (i - 1) x dist. from edge (i - 1)</td>
<td>14.56 (1.75)**</td>
<td></td>
</tr>
<tr>
<td>$RT_{\text{chosen}} (i - 1) - RT_{\text{Vmax}} (i - 1) x \text{dist. from edge} (i - 1)$</td>
<td>-1.40 (0.15)**</td>
<td></td>
</tr>
<tr>
<td>$V_{\text{max}} - V_{\text{chosen}} x Reward (i - 1) x RT_{\text{chosen}} (i - 1) - RT_{\text{Vmax}} (i - 1)$</td>
<td>-0.04 (0.08)</td>
<td></td>
</tr>
<tr>
<td>$V_{\text{max}} - V_{\text{chosen}} x Reward (i - 1) x \text{dist. from edge} (i - 1)$</td>
<td>0.17 (0.17)</td>
<td></td>
</tr>
</tbody>
</table>
$$V_{\text{max}} - V_{\text{chosen}} \times RT_{\text{chosen}}(i - 1) - RT_{V_{\text{max}}}(i - 1) \times \text{dist. from edge (i - 1)}$$

$$0.03 \ (0.01)^{***}$$

$$\text{Reward} (i - 1) \times RT_{\text{chosen}}(i - 1) - RT_{V_{\text{max}}}(i - 1) \times \text{dist. from edge (i - 1)}$$

$$1.30 \ (0.20)^{***}$$

$$V_{\text{max}} - V_{\text{chosen}} \times \text{Reward (i - 1)} \times RT_{\text{chosen}}(i - 1) - RT_{V_{\text{max}}}(i - 1) \times \text{dist. from edge (i - 1)}$$

$$-0.05 \ (0.01)^{***}$$

| Akaike Inf. Crit. | 410,492.40 | 398,407.40 |
| Bayesian Inf. Crit. | 410,533.20 | 398,594.50 |

*Note:* The subscript i refers to trial. Trial-wise entropy and value statistics were based on estimates from the SCEPTIC selective maintenance model. Distance from the edge refers to the time between the chosen response and the nearest interval boundary (either 0s or 4s). $RT_{\text{chosen}}$ is the chosen response time, whereas $RT_{V_{\text{max}}}$ is the response time associated with the highest subjective expected value based on the reinforcement history. $V_{\text{chosen}}$ is the expected value of the chosen option. $V_{\text{max}}$ is the expected value of the best option based on the reinforcement history. Reward refers to whether the previous trial was reinforced (1) or not (0). *$p < .05$; **$p < .01$; ***$p < 0.001$. 
**Supplementary Table 4**

Within-trial response time as a function of within-trial expected value and uncertainty: mixed-effects Cox survival model.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
<th>Exp(coefficient)</th>
<th>Standard error (coeff.)</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagged response time</td>
<td>-0.00044</td>
<td>0.99956</td>
<td>0.00001</td>
<td>-54.77***</td>
</tr>
<tr>
<td>Trial</td>
<td>-0.02422</td>
<td>0.97607</td>
<td>0.00060</td>
<td>-40.15***</td>
</tr>
<tr>
<td>Value, selective maintenance model</td>
<td>0.02452</td>
<td>1.02482</td>
<td>0.00053</td>
<td>46.45***</td>
</tr>
<tr>
<td>Uncertainty, U + V model</td>
<td>-0.00098</td>
<td>0.99902</td>
<td>0.00002</td>
<td>-58.75***</td>
</tr>
</tbody>
</table>

*Dependent variable:* response time hazard

**Note:** Fixed effects coefficients from a mixed-effects Cox model. Value and uncertainty were within-trial time-varying predictors. Responses and time-varying predictors were sampled every 100 ms. *p < .05; **p < .01; ***p < 0.001.
Supplementary Figures
Supplementary Figure 1. Identifiability of model parameters in simulations. Original parameters used in simulations (x-axis) vs. recovered parameters (y-axis). Parameters for all models in the SCEPTIC fixed learning rate family (left column) were recovered with high precision and minimal bias. Among the SCEPTIC KF models (middle column), KF process noise and KF U+V parameters were recovered reliably. Only some parameters in the KF U->V and KF volatility model were identified. None of the TC parameters (right column, top) were identified. In TD, $\epsilon$ was identified while $\alpha$ was not.
Supplementary Figure 2. Median correlation between trial-wise model-estimated value and true value across 500 simulated contingencies as a function of SCEPTIC parameterization. Shaded ribbons represent the standard error of the median estimated using a LOESS smoother. Contrary to our prediction, models that allowed for uncertainty-driven exploration (especially KF U → V and KF U + V) did not have an advantage over similar fixed learning rate models in optimality tests. We expected that including uncertainty in the SCEPTIC choice rule would confer an advantage early in learning because the agent would recover a higher fidelity representation of expected value across the entire action space. Compared to a simple softmax choice rule over the expected value vector, V(i), uncertainty-driven sampling is more likely to sample the action space systematically and develop a better representation of the contingency. This advantage should be especially pronounced early in learning because uncertainty-driven sampling enhances the unique information gained by each action. More specifically, incremental information is maximized by choosing the most uncertain action each time (an extension of entropy reduction; (Cover & Thomas, 2006)).

To test the hypothesis that models that included uncertainty in the choice rule would more rapidly recover an approximation of the true reinforcement contingency, for each model we computed the Pearson correlation between the trial-wise estimate of expected value, V(i), and the true reinforcement schedule. This step was repeated for each of the 500 simulated replications/contingencies, generating a distribution of correlation estimates at each trial. The tradeoff between exploration and exploitation means that an agent that explored until it had very little uncertainty about each action would do quite poorly on the clock task (unless there were an inordinate number of trials) because it would miss the opportunity to exploit high-value regions. Thus, to obtain a positive control, we simulated an infinitely exploratory agent, a model that selected a random action (time step) on each trial with equal probability. The KF U → V model, which used uncertainty-driven exploration to choose actions early in learning, tended to outperform other models in the first few trials. In addition, models that incorporated uncertainty into choice tended to recover a better estimate of the contingency in the first 10–15 trials than those that chose based on value alone. Finally, whereas models that shifted toward value exploitation later in learning
did not improve their approximation of the value function, the pure exploration null agent further refined its estimate.
Supplementary Figure 3. Example of a sparse reinforcement schedule with two valuable temporal regions used in optimality tests of uncertainty-guided exploration. Each schedule was generated by centering two Gaussian distributions randomly over the 5-second interval, each having SD = .15 seconds. The value maxima for these distributions were 50 and 100. These distributions were mixed together with random exponential numbers with $M = 10$ (one tenth the magnitude of the optimal choice), which were smoothed with a 20-span LOESS smoother to avoid extreme discontinuities between adjacent timesteps. The centers of the two Gaussians were required to be separated by at least 5 standard deviations (0.75s) to enforce temporal nonoverlap between the local and global maxima. Finally, rewards were delivered probabilistically with a constant probability of 0.7. In optimality tests, echoing the approach articulated in Materials and Methods, 60 “double bump” contingencies were randomly simulated in order to identify optimal parameter sets. Performance of each model at optimal parameters was then calculated by simulating earnings in a separate set of 100 randomly generated “double bump” contingencies.
Supplementary Figure 4. Mean proportion of possible points earned in simulated learning of time-varying contingencies in an environment containing two value temporally narrow maxima with relatively sparse reinforcement elsewhere (see Supplementary Figure 4). Panels depict model performance at different run lengths. Outcomes were drawn from 100 environments in which the temporal position of value maxima was randomly permuted. The central line in each bar denotes the mean, whereas the left and right bounding lines denote the bootstrapped 95% confidence interval around the mean. Notably, uncertainty-sensitive SCEPTIC variants gain a small advantage here, followed by the selective maintenance model. The failure of TD is likely due to discontinuities in the contingency defeating back-propagation of value from later to earlier ‘wait’ actions.
Supplementary Figure 5. Random-effects Bayesian model comparison of TD and SCEPTIC variants including an alternative policy ($U \rightarrow V$) and models that increase uncertainty estimated in response to prediction errors (KF Process Noise and KF Volatility). This figure extends from Figure 5 in the main manuscript, but includes the supplementary models described above. EP = exceedance probability. BOR = Bayesian omnibus risk. † Denotes models with one or more unidentified parameters in simulations.
**Supplementary Figure 6.** Evolution of value entropy starting from zero prior estimates on value. Lines represent the mean entropy across trials, averaging across subjects (cf. Figure 7 with random uniform prior value estimates). Trial-wise entropy was derived from the estimated value distributions of subjects at their best-fitting parameters. Shaded ribbons represent the bootstrapped 95% confidence interval of the mean at each trial. In panel a, dark vertical lines depict boundaries between different contingencies (50 trials each), explicitly signaled to participants. Panel b depicts the average change in entropy, averaging over subjects and runs (excluding run 1); this represents the typical increase in entropy during initial exploration followed by its decline as high-value actions are discovered and exploited. Better-performing subjects (right panel) exhibit proportionately greater entropy increases early in learning under the selective maintenance model, whereas poorer subjects (left panel) have higher mean entropy. Value traces were carried forward from one block to the next, an implementation that resulted in better fits for both models compared to resetting values in each block (data available upon request).
Supplementary Figure 7. The effect of entropy on trial-wise absolute response time changes (RT swings), derived from a multilevel model of all subjects. Predictors of RT swings in the model were: 1) entropy of the value representation, 2) trial, 3) distance of the previously chosen action from the maximum estimated value, 4) value of the chosen action compared to the estimated global maximum value, 5) the magnitude of the previous RT swing, 6) whether the previous outcome was a reward or omission, and 7) distance of the prior RT from the edge of the time interval. Interactions among these predictors were also included in the model, and subject and run were included as random effects. The data depicted represent the model-predicted marginal effect of entropy on RT swings at the mean of all other predictors. Vertical bars adjoining the dots denote the standard error of the predicted value.
Supplementary Figure 8. Simulated performance of Frank TC model at different parameter values across contingencies. Response time data were simulated for 500 simulated participants (replications) who completed 500 trials in each of four contingencies: increasing expected value (IEV), decreasing expected value (DEV), constant expected value (CEV), and constant expected value-reversed (CEVR). The order of contingency blocks was randomly permuted across subjects. Response time data were smoothed using a LOESS smoother (span = 10) and averaged across subjects to emphasize general patterns in response times. Parameters in the baseline configuration (top left) were identical to Frank 2009 (Supplementary Material; (Frank, Doll, Oas-Terpstra, & Moreno, 2009)): \( K = 1500, \lambda = 0.2, \nu = 0.2, \alpha_C = 0.3, \alpha_N = 0.3, \rho = 1000, \epsilon = 3000 \). As in Frank 2009, -1000ms - 1000ms of random uniform noise was added to each response time. Panel a displays simulations for 50-trial runs; panel b contains simulations for 500 trial runs. In the top right subpanels, \( \rho = 10000 \), but other parameters are unchanged. In the lower left subpanels, \( \alpha_N = 1.0 \), but other parameters are at baseline. In the lower right subpanels, \( \lambda = 0.5 \), but other parameters are at baseline.
References:


