Mice infer probabilistic models for timing

Yi Li and Joshua Tate Dudman[1]

Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, VA 20147

Edited by Randy Gallistel, Rutgers University, Piscataway, NJ, and approved September 4, 2013 (received for review June 6, 2013)

Animals learn both whether and when a reward will occur. Neural models of timing posit that animals learn the mean time until reward perturbed by a fixed relative uncertainty. Nonetheless, animals can learn to perform actions for reward even in highly variable natural environments. Optimal inference in the presence of variable information requires probabilistic models, yet it is unclear whether animals can infer such models for reward timing. Here, we develop a behavioral paradigm in which optimal performance required knowledge of the distribution from which reward delays were chosen. We found that mice were able to accurately adjust their behavior to the SD of the reward delay distribution. Importantly, mice were able to flexibly adjust the amount of prior information used for inference according to the moment-by-moment demands of the task. The ability to infer probabilistic models for timing may allow mice to adapt to complex and dynamic natural environments.

statistical inference | reinforcement-learning | mouse behavior

Animals learn the delay until a reward will be delivered following either an animal's own action or the presentation of a conditioned stimulus (1). The ability of animals to correctly infer reward delays is thought to be critical for a range of adaptive behaviors (2, 3) from operant and classical conditioning (4) to optimal foraging (5). Rodents asked to reproduce a particular time interval do so with a variance that scales in proportion to the mean (4, 6). Based largely upon this observation, neural models of timing propose that rodents learn the time interval between an action and its outcome as a mean interval perturbed by a constant coefficient of variation (~0.15) (4, 7–10). The constant variability with which the mean time is known ("scalar timing") is conceived of as the uncertainty with which a rodent knows the expected reward delay interval (11, 12). In the case of a constant reward delay such models will suffice to estimate an expected reward delay for a future action.

By contrast to the reliable timing of most operant conditioning paradigms, in a natural environment the timing with which events occur can be arbitrarily large (exceeding the variance of scalar timing) and dynamic. Consider, for example, the timing of responses from conspecifics in a social setting. Or consider a foraging animal that must decide how long to persist searching a patch for food (13). In the presence of variable information, optimal decisions require an agent to infer probability distributions that incorporate uncertainty learned through repeated experience (14, 15). Financial decision theory posits that knowledge of both the mean and variance of expected returns is necessary to select a portfolio optimally. Consistent with these predictions recent studies have shown that human subjects are capable of tracking uncertainty (16, 17). Moreover, neural correlates of uncertainty about future rewards have been observed in midbrain dopamine neurons of nonhuman primates (18) and in dopamine-recipient brain regions in human subjects (17). Although financial decision theory has largely considered uncertainty in the magnitude of returns after a fixed period, an agent may also be subject to uncertainty about the time until a positive return is realized as we described above. Optimal decisions about the amount of time one should persist in waiting for a positive return likewise require information about the average delay and the uncertainty (19).

Thus, action in the presence of uncertainty requires probabilistic information and optimal performance often requires knowledge of detailed probability distributions or their parameters. This raises the question of whether agents can infer the necessary probabilistic models. Several lines of evidence suggest that primates can infer probabilistic information about reward timing and that these inferred distributions are used to guide behavior. Human subjects asked to reproduce precise time intervals showed sensitivity to the distribution from which individual intervals were selected (20, 21). Human subjects given the option to wait for delayed rewards adjust their behavior optimally as a function of the probability distribution from which reward delays were drawn (19). Moreover, nonhuman primates allocate attention according to an arbitrary variability in timing (22, 23). Rodents can learn several discrete reward delays (1, 7, 24); however, it has been less clear whether rodents can adapt optimally to changes in uncertainty. A recent behavioral study demonstrated that mice can learn to switch between two expected reward delays rapidly (25), consistent with an inferred, probabilistic model of the task structure. Nonetheless, it remains unclear whether mice can infer probabilistic models of reward timing. Moreover, the dynamics by which a probabilistic model is constructed from recent experience remains poorly understood.

Here, we develop a switching interval variance (SIV) operant conditioning task for mice. Optimal performance of the SIV task required mice to adapt their behavior to the mean and the SD of reward delays. We find that mice adjust their behavior to the SD of reward delays across an order of magnitude change in variability. Quantitative analysis of the behavior was consistent with a process of statistical inference but not with switches among a small number of well-learned strategies. Our data were well fit by a model in which mice inferred a probabilistic model of reward delays from many tens of previous trials. Thus, our data suggest that the ability to infer probabilistic models for timing is not the privilege of primates, but rather arose much earlier in evolution.

Significance

To make optimal decisions in the presence of uncertainty requires the inference of probabilistic models. For example, financial decision theory posits that selection of optimal portfolios requires information about both the mean and variance of expected returns. In the timing literature it is often suggested that rodents learn the delay until reward delivery with a fixed relative uncertainty. However, in a natural environment the timing of events may have an arbitrary uncertainty. Thus, we asked whether mice could infer probabilistic models of timing in highly dynamic environments. Here we have developed a behavioral paradigm to show that mice can accumulate reward timing information over many tens of trials to infer accurate probabilistic models and make optimal decisions.

Author contributions: Y.L. and J.T.D. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: dudmanj@janelia.hhmi.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1310666110/-/DCSupplemental.

Results

Behavioral Task. The SIV operant conditioning task is schematically illustrated in Fig. 1. Mice ($n = 9$) were placed in a behavioral chamber with two thin metal levers protruding from the walls of the box and a recessed central port ("reward port") at which water could be delivered (Fig. 1*A*). For any given trial only one lever (the "baited" lever) would lead to the delivery of a delayed water reward on ∼85% of trials (Fig. 1*B*) and the side of the baited lever was switched in a blockwise fashion (180–200 trials/block) to ensure goal-directed and thus potentially optimal behavior (26). This design produced three distinct trial types: a correct choice of the baited lever followed by water delivery ("rewarded"), a correct choice of baited lever with no water delivered ("probe"), and an incorrect choice of the unbaited lever ("error"). In all cases, completion of a trial required approaching the reward port, and in the case of rewarded trials, the water reward had to be collected. The time delay between lever press and delivery of the water reward was randomly drawn from a Gaussian probability density function (pdf). The SD of the reward delay pdf (σ_{delay}) in each block was selected from a set of three possible values: 50, 750, and 2,000 ms. All blocks had a mean reward delay (μ_{delay}) of 3 s.

Probe trials were not cued, and thus for a given block, we assume that mice choose a strategy that was consistent across rewarded and probe trials. An approximately uniform spacing of probe trials allowed us to estimate the dynamics of the reward delay expectation both within and between blocks. We used the amount of time that the mouse waited at the reward port ($t_{waiting}$) as a measure of the expected reward delay (Fig. 1*C*). In rewarded trials, animals' waiting time was determined by the reward delay distribution (Fig. 1*C*, gray dots). However, in probe and error trials, animals have to decide when to leave the water port independent of any timing cues and based only upon their expectation of reward timing. So the waiting time in probe and error trials reflects animals' estimate of the parameters of the reward delay distribution. Even a qualitative description of "optimal" performance of the SIV task makes a number of verifiable predictions. If mice are to maximize the rate at which reward is collected, then (*i*) they must adapt their waiting time to the variability of the reward delay distribution such that the port is not abandoned until the mice are confident that the current trial is a probe trial and no reward will be delivered; (*ii*) for symmetric distributions like those used here, mice should attempt to arrive at the reward port more rapidly as the variability of reward delays increases to collect early rewards as quickly as possible; and (*iii*) on trials in which mice cannot have knowledge of the specific parameters of the reward delay distribution, they should adapt their waiting time to the feature common to all blocks—namely, the mean reward delay. Quantitative analysis of the SIV task can be used to refine the first prediction. Under the constraints of the SIV task, mice should adapt their waiting times to be linearly proportional to the SD of the reward delay distribution (see *SI Materials and Methods* for details).

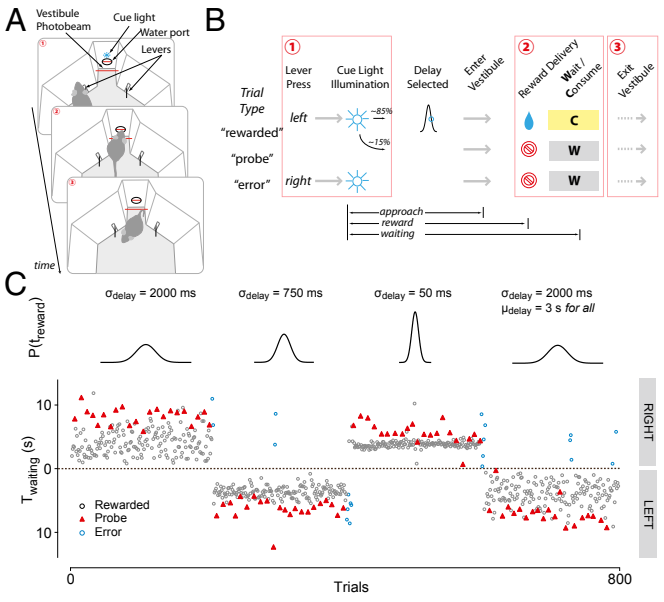

Fig. 1. Design and performance of the SIV task. (*A*) Schematic representation of select moments (red numbers 1–3) from the video of a single trial of the SIV task. (*B*) Timeline of three example trials taken from an example block in the SIV task. Blue sun indicates illumination of the cue light upon a successful lever press. Trial types are defined at left and referred to in the text. Probe trials were spaced by 6–10 trials (uniform random distribution, ∼15% of all trials in a block). Black Gaussian distributions represent the σ_{delay} for the block, and the cyan circle represents the delay chosen for the current trial. The presence or absence of water reward is indicated by a blue water drop or red cross, respectively. The duration of reward consumption and waiting periods are indicated by yellow and gray shaded areas, respectively. Exit from the vestibule was used as an indication of the attempt to initiate a new trial. Red numbers indicate the corresponding moments between the timeline (*B*) and performance schematic (*A*). Interval metrics used in the text and subsequent figures are schematically shown as intervals corresponding to the "approach," "reward," and "waiting" times for each trial. (*C*) Example data from four blocks of a single daily session in which a trained mouse performed the SIV task. For each block of trials the reward delay interval distribution is indicated in the upper row. Individual waiting times ($t_{waiting}$) on rewarded (black), probe (red), and error (cyan) trials are plotted as a function of trial.

Performance of Mice in the SIV Task Indicating They Infer the Reward Delay Distribution. We sought to test these three behavioral predictions by examining the behavior of mice trained to perform the SIV task. We first asked whether mice could adapt their waiting time in proportion to the σ_{delay} of the current block of trials. As predicted we found that $\Delta t_{waiting}$ ($t_{waiting}$ in each block relative to the $t_{waiting}$ in $\sigma_{delay} = 50$ block) was proportional to σ_{delay} of the distribution from which reward delays were drawn. Across mice $\Delta t_{waiting}$ was a linear function of σ_{delay} and well fit by a line with a slope of 1.04 ($P < 0.001$; Fig. 2*A*).

Mice could not discern rewarded trials with a particularly long delay from probe trials before the average waiting time had elapsed. Given the choice of waiting time as proportional to σ_{delay}, we thus predicted that mice would leave early on a subset of rewarded trials. To confirm the validity of this assumption, we asked whether the waiting times measured in probe trials could predict the probability with which mice abandoned the port on rewarded trials with long delays. We used the mean and SD of $\Delta t_{waiting}$ in the $\sigma_{delay} = 2{,}000$ ms blocks to predict the fraction of rewarded trials in which the mouse would leave before reward delivery [$P_{leave}(\Delta t_{waiting})$]. We found that the $P_{leave}(\Delta t_{waiting})$ calculated using probe trial waiting times could account for the observed probability with which mice abandoned the reward port before reward delivery ($\rho = 0.98$; $P < 0.001$, Pearson correlation) (Fig. 2*B*).

Given the symmetric distributions used in higher σ_{delay} blocks, there will be equal trials in which reward arrives with a short delay as a long delay. If mice indeed know the distribution from which delays are drawn, then they should also approach the reward vestibule faster in the high σ_{delay} blocks. To test this possibility we calculated the interval from lever press to reward vestibule entry ("approach time," $t_{approach}$). We found that the average approach time decreased with increasing σ_{delay} ($t_{approach}$ in $\sigma_{delay} = 2{,}000$ ms blocks is significantly shorter than it is in 50 and 750 blocks; $P < 0.001$, ranksum test; Fig. 2*C*), nearing the minimal delay achievable (∼500 ms; *Materials and Methods*).

An alternative explanation for the correlation between $\Delta t_{waiting}$ and σ_{delay} is that the large variation in reward delays used was disrupting performance. Due to the presence of two levers in our task, we could test this possibility by calculating the error rates across different blocks. We found that the error rates are very low across different blocks (0.085 ± 0.0109) and slightly

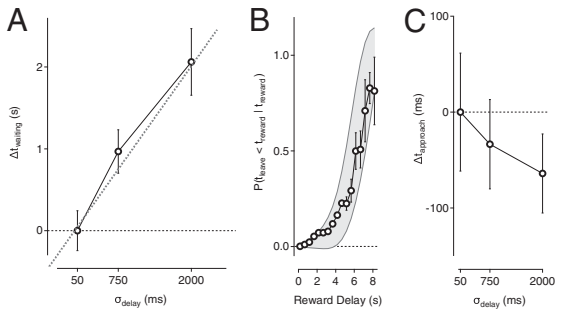

Fig. 2. Mice adjust their behavior to the SD of the action–outcome interval. (*A*) The average $\Delta t_{waiting}$ on probe trials from nine individual mice as a function of σ_{delay} of the distribution from which the reward delay was drawn. A unity line is shown. (*B*) $\Delta t_{waiting}$ tuned to the σ_{delay} predicts missed rewards on rewarded trials. For each mouse the distribution of $t_{waiting}$ on probe trials was used to predict the probability of leaving the port before reward delivery minus the probability that the leaving time (t_{leave}) is less than reward delivery time (t_{reward}). $P(t_{leave} < t_{reward})$ is plotted as a function of the actual delay on rewarded trials (gray shaded area indicates the prediction ± 1 SD from the probe trial waiting times). The observed probability binned and averaged for all mice is shown in black lines and symbols. (*C*) The time to approach ($t_{approach}$) the water port is plotted as a function of σ_{delay}. Error bars reflect the SEM of all mice.

decreased as a function of σ_{delay} (error rates in σ_{delay} = 2,000 ms blocks is significantly smaller than it is in 50 and 750 blocks; $P < 0.05$, ranksum test), suggesting that mice, if anything, were less confused during blocks with more variable reward delays. We also found no significant difference between error rates in blocks in which the left (0.0945 ± 0.0103) or right (0.0763 ± 0.0117) lever was baited. Thus, the high level of performance and adoption of both a waiting time and approach strategy correlated with σ_{delay} suggested that mice build a model of the expected reward delay distributions.

Waiting Times Are Proportional to the Mean During Exploration. We next asked whether mice could adjust behavior to the mean reward delay in the absence of specific knowledge of the reward delay distribution. In tasks with multiple operant responses, for example the two levers in our task, animals are thought to switch between a state of exploitation in which an action with a known outcome is pursued to obtain reward and a state of exploration in which an alternative option with unknown outcomes is evaluated for potential reward (27). Even after substantial training in the SIV task, mice still chose the unbaited lever in some trials during a block (Fig. 1*C*), suggesting a tendency to explore the alternative action. Consistent with a shift to exploratory behavior, we found that error trials occurred in short bursts of trials (Fig. 3*A*, *Inset*). The marginal value theorem from foraging theory predicts that during exploration one should only evaluate whether the unknown option is better than the best available option. In the context of the SIV task, the intertrial interval (ITI) is minimal when the reward delays are equal to the μ_{delay} (i.e., $\sigma_{delay} \sim 0$). Thus, during exploration mice would need to determine whether the reward delay distribution on the alternate lever is less than that of the σ_{delay} = 50 ms block. Consistent with this prediction, we found that during exploration trials mice waited for a time equal to the σ_{delay} = 50 ms block ($\Delta t_{waiting} \sim 0$), but independent of the σ_{delay} of the currently baited lever (Fig. 3*A*). A similar logic applies to the choice of waiting time after an uncued block change—that is, as confidence that the current lever is baited erodes. We found that in interblock error trials mice likewise adopted a $\Delta t_{waiting}$ that rapidly relaxed to $\Delta t_{waiting} \cong 0$ independent of the σ_{delay} of the previous block (Fig. 3*B*). Thus, waiting times in both exploratory trials and interblock error trials suggest that mice can adjust their waiting time to μ_{delay}. Importantly, this occurs independent of the waiting time on neighboring, rewarded trials.

The Dynamics of the Waiting Behavior Are Consistent with Inference of a Probabilistic Model. If it were the case that mice were using past trials to infer the distribution from which reward delays were being drawn, several predictions would follow. First, if mice adapt their behavior to a property of the inferred distribution we would predict that both the reward delay in the previous trial (Fig. S1*A*, *Left*) and/or mean reward delay for the 10 previous trials (Fig. S1*A*, *Right*) would only weakly correlate with the waiting time on a given probe trial. Indeed, we found that the correlations were weak for all blocks ($R < 0.1$). Finally, when we extracted the linear kernel (28–30) that related the waiting time in a probe trial to the reward times of the most recent trials, we found no significant structure (Fig. S1*B*; see *SI Materials and Methods* for details). The form of the linear kernel was consistent with an approximately uniform accumulation of timing information over at least 10 trials.

Second, we would predict that the transition from a low σ_{delay} block to a high σ_{delay} block should occur faster than the converse. This follows from a simple statistical argument: a high σ_{delay} block contains experienced intervals that could not be observed given a low σ_{delay} (Fig. 4*A*); however, any individual observation in the low σ_{delay} block is also consistent with a high σ_{delay} (Fig. 4*A*). To test this possibility we aligned probe trials on the block switches between the highest → lowest (H > L) and lowest → highest (L > H) σ_{delay} blocks. We found that $\Delta t_{waiting}$ on H > L transitions were adjusted more slowly (a lag of two probe trials corresponded to ∼15 rewarded trials; Fig. 4*B*) than L > H transitions. We found no significant difference in the number of error trials between transitions (Fig. 4*B*, *Inset*), and thus the slow transition at the H > L block switch could not be explained as a mere performance deficit.

A gradual change in the average $\Delta t_{waiting}$ on H > L transitions is consistent with two, very different, transitions on individual block switches (Fig. S2*A*). A smooth average transition could reflect the average of multiple step-like transitions with varying delays, or could reflect the average of smooth transitions around each individual block switch. Previous work from Kheifets and Gallistel (2012) (25) argued that mice represent probabilities and perform a step-like change in behavioral strategy following the detection of a change in reward timing. However, if mice were performing inference from a history of reward delays, we would expect a gradual transition as new timing information was incorporated following the block switch. Consistent with the latter model, we observed clear gradual transitions around many block switches (an example is shown in Fig. S2*B*). To examine

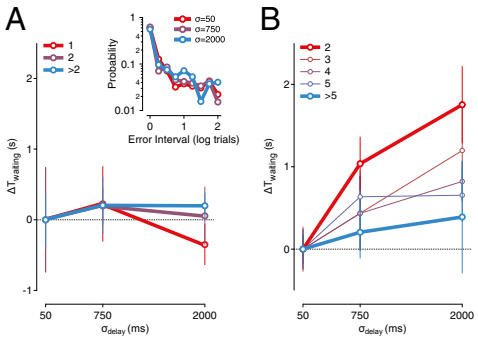

Fig. 3. Mice adapt waiting times to the mean of reward delay during exploration. (*A*) $\Delta t_{waiting}$ time on the first, second, or subsequent (colors as indicated in legend) exploration trials (error trials that occur in the middle of a block) as a function of σ_{delay} in the current block. (*A*, *Inset*) Log–log plot of the probability density for the number of trials between errors ("intererror interval," IEI) as a function of σ_{delay}. Peak IEI is at 1 (10^0), indicating a tendency to commit errors in bursts. (*B*) $\Delta t_{waiting}$ on error trials following a block switch (persistent choice of the previously baited lever). $\Delta t_{waiting}$ on each error trial plotted as a function of the σ_{delay} of the previous block for the first 2–5 trials following the block switch (colors as indicated in the legend).

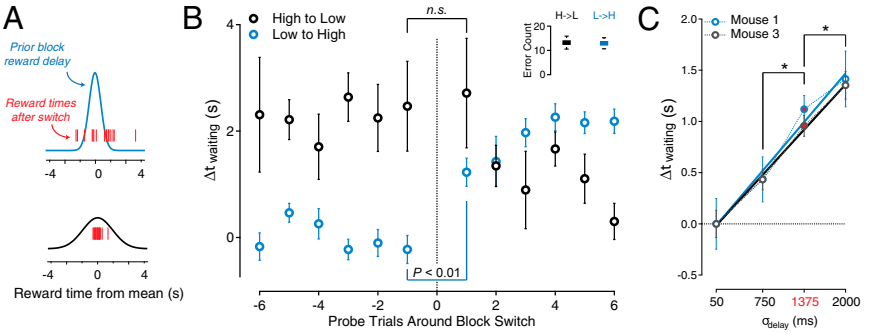

Fig. 4. Mice use recent rewarded trials to infer a probabilistic model of reward delay. (*A*) Schematic showing the observed reward delays (red ticks) and the probability density of reward delays for the previous block for an example low to high (L > H; *Upper*) and high to low (H > L; *Lower*) block transition. In an L > H transition, reward delay times after block switch fall outside of prior distribution. However, in an H > L transition, reward delays after block switch are all consistent with the prior distribution. (*B*) The $\Delta t_{waiting}$ on probe trials was calculated for transitions from the lowest (σ_{delay} = 50) to highest (σ_{delay} = 2,000) (cyan) and from the highest to lowest (black) σ_{delay} blocks and aligned to the block switch. Inset plots the number of errors following each block switch. The L > H transition is characterized by a significant shift in waiting time at the first postswitch probe trial (paired *t* test). Note that there were 6–10 rewarded trials between each probe trial. (*C*) For two mice that had learned the task with σ_{delay} = [50, 750, 2,000] blocks, an intermediate σ_{delay} = 1,350 block was introduced (red points). $\Delta t_{waiting}$ averaged across the first 10 sessions after the new block was introduced for all delays are shown. *P < 0.01; *t* test.

transitions across the entire dataset, we calculated the derivative of $\Delta t_{waiting}$ in probe trials immediately before and after a block switch. To facilitate comparison of the time course of transitions around block switches, we normalized the range of $\Delta t_{waiting}$ to the minimum and maximum mean waiting times for each mouse (absolute waiting times were idiosyncratic). If transitions were step-like, we would predict that the distribution of the derivative of $\Delta t_{waiting}$ would have a mode around –1 (Fig. S2*C*). By contrast, if mice adjusted their behavior incrementally we would predict that the distribution of the derivative of $\Delta t_{waiting}$ would have a mode that was slightly less than 0. We indeed observed that a distribution of the derivative of $\Delta t_{waiting}$ around block switches was very similar to the derivative for all probe trials, but slightly shifted below 0 with a mode of –0.22 (Fig. S2*C*). Thus, these data are consistent with a model in which mice infer the SD of reward delays.

We further evaluated whether behavior was consistent with a process of statistical inference in a subset (n = 2) of mice. If mice are attempting to transition between a small number of strategies, we would predict that by introducing a novel σ_{delay} (1,375 ms) intermediate between two previously trained σ_{delay}, mice would adopt a waiting time of one of the previously trained σ_{delay} blocks. We introduced the novel σ_{delay} into sessions containing the σ_{delay} blocks trained previously. We found that in these sessions mice adapted their $\Delta t_{waiting}$ accurately to the new σ_{delay} = 1,375 ms block ($\Delta t_{waiting}$ was significantly greater than 750 blocks and significantly shorter than 2,000 blocks; $P < 0.001$, ranksum test) (Fig. 4*C*). Thus, the absence of local correlation structure, gradual and assymetric transitions around switches in $\sigma_{delay,}$ and training with a novel σ_{delay} all support the idea that mice choose a waiting time in the current trial based upon an inferred property of the history of reward delays.

Online Computation of the SD Is Sufficient to Explain Waiting Times. If mice are computing the σ_{delay}, then how much prior timing information do mice use for this computation? To produce a quantitative estimate number of past rewarded trials (subsequently referred to as the "memory length" or "N") used by mice to infer the σ_{delay}, we assumed a parsimonious model in which an iterative algorithm uses weighted averaging to calculate the mean and SD of an array of numbers (17, 31) (*SI Materials and Methods*). This model produces an estimate of the σ_{delay} at every trial with minimal requirements for storage of past information. To fit the behavioral data, a $\Delta t_{waiting}$ on each probe trials was calculated from the predicted σ_{delay} and μ_{delay} and converted to a predicted $\Delta t_{waiting}$ using the observed dependence of waiting time upon σ_{delay} and μ_{delay} for each mouse (Fig. 2*A*). The only free parameter is thus the number of previous trials (N) used to update the predicted μ_{delay} and σ_{delay}. We fit the model by calculating the mean-squared error (MSE) between model predictions and observed $\Delta t_{waiting}$ over a range of memory lengths from 1 to 200 trials for each behavioral session.

A comparison of the $\Delta t_{waiting}$ predicted from the model and the observed $\Delta t_{waiting}$ is shown for a single session in Fig. 5*A*. The frequently changing σ_{delay} in the block structure of the SIV task is useful for revealing the contribution of previous reward delay interval information. We considered three distinct task epochs to evaluate the memory length being used by the mice: a L > H transition, a H > L transition, and the final five probe trials of any given block ("steady state"). In all cases we found that the MSE as a function of N was concave. As discussed above, there is an expected asymmetry between the L > H and H > L σ_{delay} transitions. Thus, in L > H transitions, mice should discard past timing information and bias their computation of the reward delay distribution toward more informative (recent) trials. Indeed, we found that mice had a lower memory length in the L > H transition (Fig. 5*B*) than the H > L transition (Fig. 5*C*). By contrast to transitions, in the steady state it is optimal to use as much prior information as possible. We found that the data were best explained by assuming that mice were using ~50 trials of prior reward timing information—significantly more than either transition condition (Fig. 5*D*). Finally, we note that the model provides a prediction of the rate at which $\Delta t_{waiting}$ changes after a block switch. We found that the mode of the distribution of derivatives from the model (–0.19) and the behavioral data (–0.22; *t* test, P = 0.2418) were not significantly different (Fig. 5*E*).

The behavior of mice during steady state performance was poorly explained by either the previous trial or the local mean of reward delays (Fig. S1). Estimation of the optimal linear kernel also suggested that reward timing information was weighted evenly for at least the prior 10 trials. However, around transitions our data are most consistent with a model in which mice use many fewer trials of reward timing information. Thus, we next asked whether a model in which the SD was computed was still a better predictor of behavior than the previous trial around block transitions. We found that at transitions, as is the case in steady state, an inference model was a better predictor of the change in waiting time (Fig. 5*F*).

Discussion

Here we have used a unique operant paradigm to provide evidence that mice can adapt their behavior to an inferred distribution from which reward delays are drawn. Several lines of evidence suggest that mice use reward delay times from previous trials to develop a probabilistic model of the distribution of reward delays: (*i*) Local correlations between recent reward

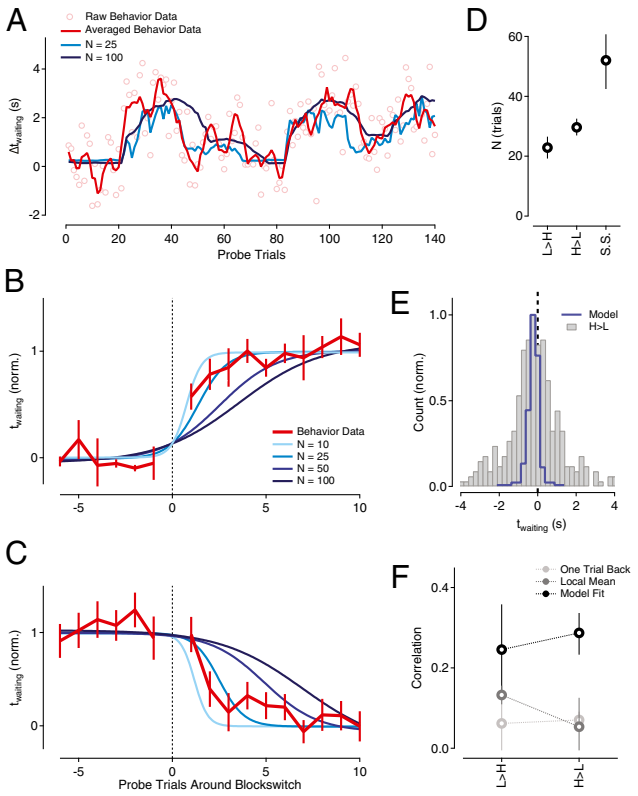

Fig. 5. Mice can use many trials of prior information to infer the SD. (*A*) $\Delta t_{waiting}$ on probe trials is shown as both individual trials (red circles) and a running average (red line, Savitzy–Golay method, 20-trial window) for an example session; rewarded trials are not shown. Predicted $\Delta t_{waiting}$ from the graded model with different memory lengths are superimposed (n = 25, cyan; n = 100, dark cyan). (*B* and *C*) Normalized predicted $\Delta t_{waiting}$ from modeling centered on block transitions from low to high (*B*) and high to low (*C*) reward interval distributions. Value of N for the model as shown. Overlaid is the averaged and normalized $\Delta t_{waiting}$ from behavioral data for all mice (red lines). (*D*) Memory length (minimum value of the error function) estimated for low to high (L > H) transitions, high to low (H > L) transitions, and the last five probe trials of a block (steady state, S.S.). (*E*) Gray bars are a replotting of the data in Fig. 4*E*. The derivative of the probe trial waiting time was calculated for the model data calculated at the optimal memory length for each mouse (blue line). The mean of the two distributions were not significantly different. (*F*) Correlation between inferring model and behavior data (dark line) was much higher than the correlation from linear model and data (gray lines) at an L–H and an H–L block transitions.

delays and waiting times on subsequent probe trials were low; (*ii*) mice adapted their behavior around sudden changes in reward delay distributions in a manner consistent with the asymmetry of statistical inference; (*iii*) mice rapidly adapted their behavior to the SD of a new distribution that had not been previously trained; (*iv*) waiting times gradually adjusted after block switches, consistent with within block learning of the distribution; and (*v*) mice could also adapt behavior to the mean reward delay even during blocks where behavior was adapted to the SD on neighboring trials. These observations were well described by a quantitative model in which the sample SD and mean reward delay were flexibly computed from 20 to 60 previous trials or reward timing information.

The ability to accumulate past information over many tens of trials is surprising given previous model-based analyses of animal learning in variable environments (28). In behavioral tasks where optimal (or near optimal) performance depends upon rapid detection of transitions in reward probability, it has repeatedly been observed that primates use locally weighted kernels to adapt decisions (28, 29). By contrast, strong sensitivity to recent trials is suboptimal in tasks where there is substantial uncertainty about reward delivery (16). Likewise, in the SIV task described here, optimal performance requires mice to remain relatively insensitive to the substantial fluctuations in reward timing between trails. Taken together these results suggest that animals, like human subjects, are capable of adapting the local weighting of reward information according to the dynamics of the task. Here we provide evidence that mice are capable of adjusting the memory length over which reward delay information was accumulated when appropriate (around block switches) or could ignore the reward delay variation altogether (during exploration). This suggests that sensitivity to local reward history is a learning parameter that is flexibly controlled.

If reward delays shorter or longer than the current waiting time policy were updated with (properly chosen) asymmetric learning rates, then mice could adopt a waiting time that was proportional to the SD through trials and error. A number of observations in this study are inconsistent with a model in which the waiting time strategy was updated in proportion to the reward timing of the previous trial. For example, such a model cannot account for our observation that mice approach faster, wait longer, and explore for the same amount of time on neighboring trials. Moreover, even around block transitions we found that behavior was better described by a model in which the sample SD was used to select waiting times (Fig. 5*F*). As a final evaluation of this alternative model, we also examined steady state behavior. A model in which waiting times were updated on every trial should exhibit asymmetric fluctuations in waiting times (only a small fraction of rewarded trials are equal to the mean plus 1 SD). Thus, we compared the waiting time on probe trials (an approximation of the waiting time policy) to the relative reward delivery time on the prior trial. When we examined such trials (Fig. S3), we failed to find any evidence of asymmetric learning rates. If anything, the slope (–0.08) was the opposite sign of that predicted for learning rates that would produce a waiting time proportional to the SD. Thus, we believe that the observations presented here are most consistent with a model in which mice integrated timing information from many trials of prior experience to infer a probabilistic model of reward timing. Recent work (25) has similarly argued that the behavior of mice in reward timing tasks was better described by assuming mice inferred a probabilistic model of environmental dynamics than the assumption of incremental reinforcement learning.

We propose that mice use the differences between the actual reward time and the mean reward time to compute the sample SD in the SIV task. This is analogous to the computation proposed (32) to explain how human subjects could compute the uncertainty about the probability of reward delivery. The difference between the actual time of reward and the mean time of reward is a "prediction error" (33). The transmitter systems and circuits that underlie interval timing (6) are the same as those that signal prediction errors (34, 35), and thus a common mechanism is attractive. Reinforcement learning describes a process in which prediction errors are used to modify reward-seeking behavior (36). Learning is completed when prediction errors are eliminated and behavior stabilizes. Our data imply a different description of the role of prediction errors. In the case of substantial variation in reward timing, behavioral performance would stabilize despite persistent "errors" in prediction (or risk prediction errors) (16). Rather, prediction errors would be used to estimate uncertainty, but could reduce to zero in the case of deterministic reward timing. Behavioral responses, in our case waiting times, are then adjusted based upon properties of the distribution of errors rather than upon the current prediction error per se. Although the basic elements of such a model are consistent with the critical role of dopaminergic signaling in the striatum for timing (6), confirmation will require neural recordings from mice performing tasks with variable reward timing such as the one developed here.

The inference of a model of the environment is thought to be a critical feature of diverse behaviors in a number of organisms

(37–40). The associative learning literature has focused extensively on the capacity of organisms and neural circuits to learn through incremental updates of synaptic weights or association strengths. However, in a natural environment events are embedded in a constant stream of variable stimuli and actions, and the traditional associative learning model may be insufficient to track such dynamics (11). By contrast to the associative models often proposed for learning in rodents, inference in humans has been argued to proceed through the application of internal statistical models to experience (41, 42). Here we have shown that mice can generalize properties of reward timing across actions (learning across block switches that involve a change in the baited lever) and also use a distinct behavioral strategy for two closely related actions despite the same prior information (exploration of the unbaited lever). Moreover, mice learned to rapidly adapt their waiting time despite limited numbers of observations and then stabilized that choice despite substantial local fluctuations. Rapid shifts in behavior combined with stability in the presence of uncertainty and generalization across actions are the strengths of "model-based" learning. Thus, our data suggest that the inference of probabilistic models may be as critical to reasoning about uncertain environments in mice as it is in man (43).

Materials and Methods

Details of the SIV Task. We used nine adult (>30 g, >4-mo-old) male C57/Bl6 mice from the in-house breeding colony. In each block, levers were never retracted and remained freely available to the mice at all times. Suprathreshold vertical displacement of either lever during the ITI caused the immediate illumination of the cue light. In a given block, movement of correct lever (left or right) resulted in a delayed delivery of water in ~85% of trials (unrewarded trial spacing was selected from a uniform random distribution of 6–10 rewarded trials). The opposite lever (right or left) was never rewarded. Water was delivered to the reward port with a constant mean delay of 3 s but with trial-to-trial delays chosen according to distributions with three different SDs (50, 750, and 2,000 ms). Extreme values (occasionally chosen at random) were bounded by a minimal delay of 500 ms and a maximal delay of 9 s. To complete a trial, the mouse was required to enter and depart the reward vestibule at least once. Upon departure the mouse was free to choose either of the two levers to start another trial with only a minimal delay required for data transfer. The SD of the reward delay and lever were random with respect to the associated lever (left or right) and order of presentation within a daily session (generally 3–7 blocks).

Analysis of Behavioral Data. Analysis was performed using custom-written routines in Matlab R2011a (Mathworks) and Igor Pro (Wavemetrics). Trials were first classified into three types: a correct choice of the baited lever followed by water delivery (rewarded), a correct choice of baited lever with no water delivered (probe), and a choice on the unbaited lever (error). We further defined a set of distinct time intervals: the time between successful lever pressing and crossing the vestibule beam break (approach time), the time between a successful lever press and the delivery of water reward ("reward time"), and the time between a successful lever press and restoration of the vestibule beam break ("waiting time"). See also Fig. 1*B* and *SI Materials and Methods*.

ACKNOWLEDGMENTS. Joseph Paton and Parvez Ahammad provided critical input to the design of the project. We thank Winfried Denk, Roian Egnor, Brett Mensh, Albert Lee, Nelson Spruston, Dima Rinberg, Alla Karpova, and Arbora Resulaj for reading and commenting on a previous version of the manuscript. This work was presented at internal Janelia Farm Research Campus (JFRC) seminars; we are indebted to the constructive comments from numerous other colleagues and the anonymous reviewers. This work was supported in part through the JFRC Visiting Scientist Program. Y.L. is a postdoctoral associate and J.T.D. is a Group Leader at the JFRC of Howard Hughes Medical Institute.

1. Gallistel CR, Gibbon J (2000) Time, rate, and conditioning. *Psychol Rev* 107(2):289–344.
2. Staddon JE, Higa JJ (1999) Time and memory: Towards a pacemaker-free theory of interval timing. *J Exp Anal Behav* 71(2):215–251.
3. Meck WH (2003) *Functional and Neural Mechanisms of Interval Timing* (CRC, Boca Raton, FL), pp xli, 551 p, 554 p of plates.
4. Gibbon J (1977) Scalar expectancy theory and Weber's law in animal timing. *Psychol Rev* 84(3):279–325.
5. Krebs JR, Kacelnik A (1984) Time horizons of foraging animals. *Ann N Y Acad Sci* 423: 278–291.
6. Buhusi CV, Meck WH (2005) What makes us tick? Functional and neural mechanisms of interval timing. *Nat Rev Neurosci* 6:755–765.
7. Machado A, Malheiro MT, Erlhagen W (2009) Learning to time: A perspective. *J Exp Anal Behav* 92(3):423–458.
8. Gibbon JCR, Meck WH (1984) Scalar timing in memory. *Timing and Time Perception*, ed Gibbon J (New York Academy of Sciences, New York), Vol 423, pp 52–77.
9. Simen P, Balci F, de Souza L, Cohen JD, Holmes P (2011) A model of interval timing by neural integration. *J Neurosci* 31(25):9238–9253.
10. Kirkpatrick K, Church RM (2003) Tracking of the expected time to reinforcement in temporal conditioning procedures. *Learn Behav* 31(1):3–21.
11. Balsam PD, Drew MR, Gallistel CR (2010) Time and Associative Learning. *Comp Cogn Behav Rev* 5:1–22.
12. Balci F, et al. (2011) Optimal temporal risk assessment. *Frontiers in Integrative Neuroscience* 5:56.
13. Kacelnik A, Bateson M (1996) Risky theories—The effects of variance on foragining decisions. *Amer Zool* 36:402–434.
14. Knill DC, Pouget A (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci* 27(12):712–719.
15. Ma WJ, Navalpakkam V, Beck JM, Berg Rv, Pouget A (2011) Behavior and neural basis of near-optimal visual search. *Nat Neurosci* 14(6):783–790.
16. Preuschoff K, Bossaerts P (2007) Adding prediction risk to the theory of reward learning. *Ann N Y Acad Sci* 1104:135–146.
17. Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. *J Neurosci* 28(11):2745–2752.
18. Fiorillo CD, Tobler PN, Schultz W (2005) Evidence that the delay-period activity of dopamine neurons corresponds to reward uncertainty rather than backpropagating TD errors. *Behav Brain Funct* 1(1):7.
19. McGuire JT, Kable JW (2012) Decision makers calibrate behavioral persistence on the basis of time-interval experience. *Cognition* 124(2):216–226.
20. Jazayeri M, Shadlen MN (2010) Temporal context calibrates interval timing. *Nat Neurosci* 13(8):1020–1026.
21. Miyazaki M, Nozaki D, Nakajima Y (2005) Testing Bayesian models of human coincidence timing. *J Neurophysiol* 94(1):395–399.
22. Janssen P, Shadlen MN (2005) A representation of the hazard rate of elapsed time in macaque area LIP. *Nat Neurosci* 8(2):234–241.
23. Ghose GM, Maunsell JH (2002) Attentional modulation in visual cortex depends on task timing. *Nature* 419(6907):616–620.
24. Caetano MS, Guilhardi P, Church RM (2012) Stimulus control in multiple temporal discriminations. *Learn Behav* 40(4):520–529.
25. Kheifets A, Gallistel CR (2012) Mice take calculated risks. *Proc Natl Acad Sci USA* 109(22):8776–8779.
26. Balleine BW, Dickinson A (1998) Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37(4-5):407–419.
27. Stephens DW, Krebs JR (1986) *Foraging Theory* (Princeton Univ Press, Princeton, NJ), pp xiv, 247 pp.
28. Corrado G, Doya K (2007) Understanding neural coding through the model-based analysis of decision making. *J Neurosci* 27(31):8178–8180.
29. Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304(5678):1782–1787.
30. Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47(1):129–141.
31. Knuth DE (1998) *The Art of Computer Programming. Seminumerical Algorithms* (Addison–Wesley, Boston), 3rd Ed, p 232.
32. Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 51(3):381–390.
33. Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol Rev* 88(2):135–170.
34. Schultz W, Dickinson A (2000) Neuronal coding of prediction errors. *Annu Rev Neurosci* 23:473–500.
35. McCoy AN, Platt ML (2005) Expectations and outcomes: Decision-making in the primate brain. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 191(3):201–211.
36. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA), pp xviii, 322 pp.
37. Körding KP, et al. (2007) Causal inference in multisensory perception. *PLoS ONE* 2(9):e943.
38. Shadlen MN, Britten KH, Newsome WT, Movshon JA (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci* 16(4):1486–1510.
39. Dally JM, Emery NJ, Clayton NS (2006) Food-caching western scrub-jays keep track of who was watching when. *Science* 312(5780):1662–1665.
40. Zentall TR (2011) Maladaptive "gambling" by pigeons. *Behav Processes* 87(1):50–56.
41. Griffiths TL, Tenenbaum JB (2011) Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *J Exp Psychol Gen* 140(4):725–743.
42. Griffiths TL, Sobel DM, Tenenbaum JB, Gopnik A (2011) Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cogn Sci* 35(8):1407–1455.
43. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022):1279–1285.