Mice take calculated risks

Aaron Kheifets¹ and C. R. Gallistel¹

Department of Psychology, Rutgers University, Piscataway, NJ 08854

Contributed by C. R. Gallistel, April 18, 2012 (sent for review November 25, 2011)

Animals successfully navigate the world despite having only incomplete information about behaviorally important contingencies. It is an open question to what degree this behavior is driven by estimates of stochastic parameters (brain-constructed models of the experienced world) and to what degree it is directed by reinforcement-driven processes that optimize behavior in the limit without estimating stochastic parameters (model-free adaptation processes, such as associative learning). We find that mice adjust their behavior in response to a change in probability more quickly and abruptly than can be explained by differential reinforcement. Our results imply that mice represent probabilities and perform calculations over them to optimize their behavior, even when the optimization produces negligible material gain.

reinforcement learning | decision under uncertainty | model-based control | probability estimation | timing

A central question in cognitive neuroscience is the extent to which computed representations of the experienced environment direct behavior. Many contemporary models of adjustment to changing risks assume computationally simpler event-by-event value-updating processes (1–3). The advantage of these models is that a computationally simple process does in time adjust behavior more or less optimally to changes in, for example, the probability of reinforcement. The advantage of computationally more complex model-based decision making is the rapidity and accuracy with which behavior can adjust to changing probabilities (4, 5). Recent studies with humans (6) suggest that the rapidity of adjustment to changes in the payoff matrix in a choice task requires a representation of risk factors (probabilities and variabilities). We report an analogous result with mice when they adjust to a shift in relative risk. In our experiment, a change in the relative frequencies of short and long trials changes the optimal temporal decision criterion in an interval-timing task. We find that mice make the behavioral adjustment to the change in relative risk about as soon and abruptly as is in principle possible. Our results imply that neural mechanisms for estimating probabilities and calculating relative risk are phylogenetically ancient. This finding means that they may be investigated by genetic and other invasive procedures in genetically manipulated mice and other laboratory animals.

The hopper-switch task (7) has been used to show approximately optimal risk adjustment in mice (8). The task requires the mouse to decide on each trial whether to leave a short-latency feeding location (hopper) to go to a long-latency location (hopper) based on its estimate of how much time has elapsed since the trial began (Fig. 1). If the computer has chosen a short trial, the mouse gets a pellet in the short-latency hopper in response to the first poke there at or after 3 s. Failure to obtain a pellet there after 3 s have elapsed implies that the computer has chosen a long-latency trial. On a long-latency trial, a pellet is delivered into the other hopper in response to the first poke there at or after 9 s. The computer chooses short or long trials at random, with the probability of choosing a short trial denoted p_s . At the beginning of a trial, the mouse does not know which type of trial it will be.

The rational strategy given these contingencies is to poke in the short hopper at the start of each trial, then switch to the long hopper if and when 3 s or more elapse without a pellet being delivered in response to a poke into the short hopper. The mice

learn to do this action within a few daily sessions. When the mice have learned the action, they get a pellet on almost every trial. However, the mice sometimes switch too soon (after fewer than 3 s) or too late (after more than 9 s), and they sometimes go directly to the long hopper. Whether these errors cost them a pellet or not depends on which kind of trial the computer has chosen. Thus, the risk of one kind of error or the other depends on p_s , the probability that the current trial is a short one, and on the subject's timing variability, how precisely it can judge when 3 s have elapsed.

When p_s changes, altering the relative risks from the two kinds of error, mice shift the distribution of their switch latencies by an approximately optimal amount (8) (Fig. 2). The question we pose is, what kind of process underlies this behavioral shift in response to a change in the relative frequencies of the short and long trials?

On the reinforcement-learning hypothesis, the shift in the distribution of switch latencies is driven by a change in the relative frequency with which different switch latencies are reinforced. When the short-latency trials are frequent, the probability of a long-latency switch being reinforced increases and the probability of a short-latency switch decreases, and vice versa when long-latency trials are frequent. With this hypothesis, a slow trial-and-error hill-climbing process leads in time to a new stable distribution that maximizes the reinforcement probability. In this reinforcement-learning hypothesis, the mouse makes no estimate of the relative frequency of the short and long trials, nor of its own timing variability, and it does not calculate an optimal target. The mouse simply discovers by trial and error the target switch latency that maximizes the rate of reinforcement.

On what we call the model-based hypothesis, the mouse brain estimates p_s , it estimates its timing variability (σ) or its uncertainty about how much time has elapsed, and it calculates from its estimates of these stochastic parameters the optimal target switch time (Fig. S1). With this hypothesis, there will be a step-change in the distribution of switch latencies as soon as the mouse detects the change in the relative frequency of the short and long trials, estimates the new relative frequency, and calculates the new optimum. In this step hypothesis, a step-change in the distribution of switch latencies can occur before the change in the relative frequency of the trial types has produced any change in the rate of reinforcement. In the reinforcement-learning hypothesis, this cannot happen because the change in behavior is driven by a change in the relative rates at which short and long switches are reinforced.

Results

As may be seen from the raw data on switch latencies plotted in Fig. 2, when we changed the relative frequency of the short and long trials, the mice adjusted the distribution of their switch latencies so quickly and so abruptly that the adjustments appear step-like.

Author contributions: A.K. and C.R.G. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

¹To whom correspondence may be addressed. E-mail: kheifets@eden.rutgers.edu or galliste@ruccs.rutgers.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1205131109/-DCSupplemental.

Fig. 3. (A) A transition function, with parameters, S (start) and D (duration), describes the trial-by-trial change in the parameter vector, $\theta = \langle e, \mu, \sigma, p \rangle$, of the ExpGauss mixture. Trial 0 is the last trial before the change in p_s . (B) Cumulative distribution of S estimates. (C) Cumulative distribution of Bayes factors favoring $D = 1$ (a step switch) vs. $1 \leq D \leq 100$ (a more gradual switch).

from 3 to 105. However, in all 14 transitions, the likelihood at $D = 1$ was higher than the mean likelihood over the range $1 \leq D \leq 100$. Thus, in every case, the transition data favored the step hypothesis ($D = 1$) over the alternative hypothesis that the transition lasted somewhere between 1 and 100 trials (see Fig. S2 for the marginal-likelihood functions). The cumulative distribution of the Bayes Factors (the odds in favor of $D = 1$) is in Fig. 3C. The combined odds in favor of a step transition (the product of the Bayes factors) are enormous.

Our key findings are thus: (i) It takes mice about 10 trials to detect a substantial change in the relative frequency of the short and long

trials and to compute an estimate of the new p_s , based on the data since the estimated change; (ii) The change in target switch latency—hence the distribution of switch latencies—is accomplished in a single trial. The step change in the distribution implies that the prechange estimate of p_s is replaced by a postchange estimate, based on postchange data that do not overlap the prechange data.

To constrain the space of possible calculations that could underlie this behavior, we considered two basic kinds of algorithms: reinforcement-driven algorithms that optimize behavior without modeling the experienced world, and model-driven algorithms that estimate stochastic parameters of the experienced world (probabilities and variabilities) and calculate an optimal change in decision criterion (target switch time) given those estimates. Reinforcement-learning algorithms are often used in machine learning when computing a model of the environment in which the machine is to act is thought to be too complex. The assumption that reinforcement shapes behavior without the brain's constructing a representation of the experienced world has a long history in psychology and behavioral and cognitive neuroscience (9–13). In reinforcement learning, behavioral change is driven by changes in the rate of reinforcement that a pattern of behavior (“strategy”) produces. In the present case, this translates to changes in the average number of pellets obtained per trial for a given target switch latency. The target switch latency is the strategy. Varying this strategy will vary the rate of reinforcement.

In line with the findings of ref. 8, we found that subjects missed very few reinforcements. Each plot in Fig. 4 plots two functions in the plane of the parameters of the Gaussian distribution (μ and σ). The gain contours (red curves) delimit the expected percentage of reinforced trials for combinations of values for these parameters. When the μ -and- σ estimates for the Gaussian component of a subject's switch-latency distribution are below the 99% gain contour, the subject gets pellets on more than 99% of the trials. When above a given contour, the subject gets pellets on a smaller percentage of the trials.

The joint confidence limits around our estimates for the values of the parameters of the Gaussian component of their switch-latency distribution are the small concentric circles in Fig. 4. We assume that subjects cannot control their timing variability; that

Fig. 4. Parameters of the Gaussian switch-latency distributions (black concentric circles) in relation to gain contours (red upside down U curves). A gain contour traces the upper limit on the combinations of variability (σ) and mean (μ) consistent with an expected percentage of reinforced trials. The small concentric circles are the confidence limits on our estimates of the mean and SD of the Gaussian component of the subject's switch-latency distribution. The circles consistently fall near the mode of the 99% gain contour. Note the change in gain contours with changes in p_s . Note also the shallowness of the gain hill. The shallower the gain hill, the more slowly a hill-climbing algorithm can find its summit (that is, converge on the optimal strategy). Note finally that the confidence limits under different p_s conditions do not overlap. Thus, the changes in p_s produce statistically significant changes in the mean switch latency, as is also evident in the raw data (Fig. 2).

is, the location of the concentric circles along the σ axis. Subjects determine the location along the μ axis when they calculate or otherwise arrive at a target switch latency. As may be seen in Fig. 4, the observed locations are generally close to the mode of the 99% gain contour (close to the optimal location), even though substantial deviations from that mode would have little impact on gain (rate of reinforcement). The gain contours shift when p_S changes. The optimal target latency depends on p_S and on a subject's timing variability. Sensitivity to one's own variability has also been demonstrated in humans (4, 14, 15).

We examined whether the rate of reinforcement could drive these behavioral shifts by calculating the number of pellets missed before the change in behavior went to completion. In estimating when the change went to completion, we used the expected duration of a transition, rather than the maximally likely duration, which, as already noted, was always 1. Over 30% of the behavioral shifts went to completion before a single pellet was missed (because of switching too soon or too late). Moreover, the pattern and number of reinforcements missed after the behavioral shift was indistinguishable from the pattern and number missed over the same span of trials before the shift. Thus, our data rule out reinforcement-learning as an explanation for the changes in the distribution of switch latencies. The change in the distribution of switch latencies must be driven by a change in the estimate of p_S , based on the subject's experience over several trials.

Finally, we asked whether the adjustments in the estimates of p_S were a consequence of trial-by-trial probability tracking (leaving later after experiencing one or two short trials and sooner after experiencing one or two long trials). Given the abruptness of the changes, the resulting estimates of p_S would have to be highly sensitive to the most recent two or three trials. However, the distributions of switch latencies conditioned on the number of the most recent trials (for $n = 1, 3$, and 5) did not differ. That is, the distribution of switch latencies following three short trials did not differ from the distribution following three long trials. This finding implies that the abrupt behavioral shifts resulted from a change-detecting mechanism that parsed the experienced sequence into trials before a change and trials after a change, with the new estimate of p_S based only on the relatively few trials between the estimated change point and the trial on which the change was perceived. In arguing for a model-based adjustment algorithm, we mean only that the algorithm operates on estimates of probabilities and change points. Whether it is necessary to assume that brains—even human brains—make such estimates has long been a question of moment in psychology, cognitive science, and neuroscience.

In a subsequent experiment with six new mice, we made the change approximately 40 trials into a session rather than at the beginning of a new session. We observed similarly abrupt changes. The marginal-likelihood functions from this experiment are in Fig. S3.

We conclude that mice can accurately represent sequential Bernoulli probability (p_S), quickly perceive a substantial change in this probability, locate the point in the past at which the change occurred, estimate the new probability on the basis of the postchange trials, represent the uncertainty in their endogenous estimate of time elapsed in a trial, and correctly compute a new approximately optimal target switch latency.

This simple paradigm may be used with genetically manipulated mice, opening up the possibility of using genetic and molecular biological methods to discover the neurobiological mechanisms that make these evolutionarily ancient, life-sustaining computations possible.

Materials and Methods

Protocol. Subjects lived 24/7 in the environment shown in Fig. 1, obtaining all their food by performing the tasks. For a complete description of this behavior-testing environment and the associated software, see ref. 16. The five mice of the CB57BL/6j strain obtained from Jackson Laboratories were ~100 d old at the start of the experiment. In session 1 (1 d), the task was a concurrent VI matching protocol. In session 2 (1 d) it was a two-latency, two-hopper, autoshaped hopper-entry protocol: steps 1 and 2 were as in Fig. 1, except that at step 2 (trial onset), the light came on in one or the other of the two feeding hoppers and a pellet was delivered at the short or long latency, regardless of the mouse's behavior. This process taught the mice the feeding latency associated with each hopper. In the switch task, the interval from the end of a trial to the next illumination of the trial-initiation hopper was exponentially distributed with an expectation of 60 s. The pellets were 20 mg Purina grain-based from WF Fisher and Son, Inc. The sole condition on pellet release was that the first poke after the computer-chosen feeding latency be in the correct hopper (the short hopper on short trials and the long hopper on long trials). Any behavior pattern that met this criterion was reinforced (for example, going directly to the long hopper on a long trial or switching from the short hopper to the long and then back to the short within 3 s on a short trial). This protocol was approved by the Institutional Animal Care and Use Committee of Rutgers University.

Transition Analysis. We assume that all four parameters follow the same two-parameter linear transition trajectory (Fig. 2A). Thus, our statistical model for a transition has 10 parameters: 4 for the distribution before the transition (θ_b), 4 for after the transition (θ_a), and the transition parameters, S and D (Fig. 2A). We used Metropolis Hastings Markov chain Monte Carlo methods to sample from the posterior distribution to estimate values for these 10 parameters. The estimates for θ_b and θ_a were essentially the same as the maximum-likelihood estimates. To check the validity of the marginal-likelihood functions for S and D obtained from Markov chain Monte Carlo sampling, we computed the latter numerically after fixing θ_b and θ_a at their maximum-likelihood values.

ACKNOWLEDGMENTS. We thank Joseph Negron for help in running the experiment and Jacob Feldman for valuable comments on an early draft. This research was funded by Grant R01MH077027 from National Institute of Mental Health, entitled "Cognitive Phenotyping in the Mouse."

- Corrado GS, Sugrue LP, Seung HS, Newsome WT (2005) Linear-nonlinear-Poisson models of primate choice dynamics. *J Exp Anal Behav* 84:581–617.
- Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* 8:429–453.
- Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annu Rev Psychol* 57:87–115.
- Trommershäuser J, Maloney LT, Landy MS (2008) Decision making, movement planning and statistical decision theory. *Trends Cogn Sci* 12:291–297.
- Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704–1711.
- Stüttgen MC, Yildiz A, Güntürkün O (2011) Adaptive criterion setting in perceptual decision making. *J Exp Anal Behav* 96:155–176.
- Fetterman JG, Killeen PR (1995) Categorical scaling of time: Implications for clock-counter models. *J Exp Psychol Anim Behav Process* 21:43–63.

- Balci F, Freestone D, Gallistel CR (2009) Risk assessment in man and mouse. *Proc Natl Acad Sci USA* 106:2459–2463.
- Hull CL (1943) *Principles of Behavior* (Appleton-Century-Crofts, New York).
- Skinner BF (1938) *The Behavior of Organisms* (Appleton-Century-Crofts, New York).
- Skinner BF (1981) Selection by consequences. *Science* 213:501–504.
- Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
- Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci* 9:545–556.
- Trommershäuser J, Maloney LT, Landy MS (2003) Statistical decision theory and the selection of rapid, goal-directed movements. *J Opt Soc Am A Opt Image Sci Vis* 20:1419–1433.
- Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214–1221.
- Gallistel CR, et al. (2010) Screening for learning and memory mutations: A new approach. *Acta Psychologica Sinica* 42(1):138–158.