MICE TAKE *CALCULATED* RISKS

By

AARON THRALL KHEIFETS

A thesis submitted to the

Graduate School-NewBrunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of Master of Science

Graduate Program in Psychology

written under the direction of

C. Randy Gallistel

and approved by

__________________________________________

__________________________________________

__________________________________________

__________________________________________

New Brunswick, New Jersey

October 2011
ABSTRACT OF THE THESIS

Mice Take *Calculated* Risks

by AARON THRALL KHEIFETS

Thesis Director:

C. Randy Gallistel

Balci et al. 2009 describe a simple timing task in which reward appears after a short latency at one location, with probability p(S), and at another location after a longer latency, with the complementary probability. Human and mouse subjects began each trial waiting at the short location. As time elapsed in a trial, they switched to the long location. In response to changes in p(S), they made approximately optimal changes in the distribution of their switch latencies. We have replicated the Balci et al. finding with mouse subjects, while measuring the latency and abruptness of the adjustments caused by changes in p(S). The latencies are short, implying that mice rapidly detect a change in probability and rapidly estimate the new probability. The changes from the old to the new distributions are also abrupt, making them indistinguishable from step changes. This suggests the explicit detection of the change in p(S), followed by the computation of a new decision criterion (a new target switch time), which requires an enduring representation of the subject's temporal uncertainty together with a new estimate of p(S). The abruptness of the adjustments does not appear to be consistent with the gradual attainment of a new dynamic equilibrium through "hill-climbing," as in simple reinforcement-learning models. To achieve the
observed degree of abruptness, the learning rate parameter must be set very high, but then a reinforcement-learning model would track the stochastic noise in the sequence of short and long trials, which the mice do not do.
Table of Contents

1. Introduction .................................................................................................................. 1

2. Method ........................................................................................................................ 6

3. Results.......................................................................................................................... 8

4. Discussion..................................................................................................................... 22
   4.1 Shifts occur shortly after changes ................................................................. 23
   4.2 Shifts were abrupt............................................................................................... 27
   4.3 Shifts Not Based on Differential Feedback ................................................... 28
   4.4 Shifts Are Based on Detecting Changes......................................................... 32
   4.5 Summary................................................................................................................. 35

5. Bibliography ............................................................................................................... 37
1. Introduction

Adapting to changing probabilities in one’s environment is a ubiquitous problem for all animals. This is an especially difficult task in comparison to, say, adapting to a change in temperature because probabilities are not directly observable. It is only through repeated observation of similar events (several flips of the same coin) that we can estimate hidden variables such as the probability of a heads. Estimating probability, however, is just the beginning of the problem. Factors influencing the probability of an event are constantly changing, so that one must also update one’s probability estimate. So, whenever an animal observes a string of events deemed unlikely by its current model of the distribution of said events, it must either change its behavior in accordance with a new model that better fits these events or continue its current pattern of behavior, treating the unexpected streak as stochastic variability. Doubtless, animals successfully navigate the world with its many, changing, hidden variables. We argue that their performance in detecting and responding to changes in the values of a hidden variable cannot be explained by reinforcement models which gradually tune behavior to maximize reward. Rather, they seem to detect changes and abruptly change their behavior to align with a new estimate of the hidden variable, similar to the results in the matching domain of Gallistel et al. (2001).

In the task we use, optimal performance requires estimating the probability of each trial type and noticing when that probability changes. In fact, each response from an optimal agent can be seen as an estimate of what that optimal agent takes
the probability of a given trial type to be. Without signaling subjects, we change
the probabilities of the trial types and then quantify the response to this change.
One can see from a scatter of the raw data (figure 2 below), that subjects
responded to changes in the probability of trial type shortly after the trial at which
the change occurred. The bulk of our analyses consisted in quantifying the latency
and abruptness of the changes in the distributions of switch latencies that are so
evident in figure 2 below.

Our study confronts subjects with a timing task that is influenced by a hidden
bernoulli process: each trial is a “short” or “long” trial as determined by an
unsigned flip of a coin of weight p. On short trials, subjects are rewarded for not
acting too early and on long trials, they are rewarded for not acting too late. This
means that the weight of the coin has an impact on the ideal temporal decision
criterion for an ideal subject: the higher the probability of a short trial, the later an
ideal subject with limited timing precision should set its temporal decision
criterion. If short trials are more likely, it is important to minimize the risk of
acting too early even at the expense of increasing the probability of acting too
late, since acting too late is not penalized on short trials. This calculation can be
understood in the framework of signal detection theory.

Two separate levels of signal detection.

1. One type of signal detection occurs at the trial-level. If a perfectly rational
agent were performing this task, each trial would begin with the agent taking the
probability of a short trial to be the recently observed proportion of short trials. As
time elapsed and the agent waited at the short-side hopper, the agent’s confidence would climb that three seconds had elapsed since the beginning of the trial. Since the agent would know that three seconds elapsing without being fed and the current trial being a short one are mutually exclusive, this would influence their decision of what moment to switch at as follows:

\[
\text{Switch iff } \text{baserate}_{\text{long}} \ast N(x;9,\text{cv}^9) > \text{baserate}_{\text{short}} \ast N(x;3,\text{cv}^3)
\]

where \( N \) stands for the Gaussian probability density function and \( \text{cv} \) stands for the coefficient of variability for the timer of the agent.

In this way, we can think of the agent as engaged in signal detection: attempting to tell if its own sense that three seconds have elapsed is due to the noise inherent in its timing mechanism or if it is veridical.

2. A second signal detection problem occurs on the inter-trial level: is the recent sequence of trials drawn from the same distribution that has been generating the trial sequences or has the distribution changed? By the end of each trial, the subject knows whether that trial was a long or a short one: either by being fed or by timing out on the long or short side. Thus, for every trial there is a question of what Bernoulli distribution it was drawn from. If the subject has seen a lot of long trials and then sees a streak of short trials, this could either be an unlikely set of draws from the same distribution or draws from an entirely new distribution. How
convincing a streak must be in order to be flagged as a change in distribution\(^1\)?

The question we pose is how rapidly can mice detect a substantial change in the relative frequency of long and short trials (a change in the exogneous stochastic parameter) and how abruptly do they adjust to this change.

We found that the distribution of switch latencies (the number of seconds from the beginning of a trial to the time when they left poking on the short side for poking on the long side) depended strongly on the probability of a long versus a short trial, as found by Balci et al, and as the signal detection framework would suggest: An increase in the probability of a short trial translated the distribution of switch latencies in the direction of longer latencies, and vice versa (Figure 1).

\(^1\) We assume subjects place equal utility on false positive and false negative.
Figure 1. The conditional probability distributions of the subjective times of reinforcement ($t^*$) on short and long trials, $p(t^*|S)$ and $p(t^*|L)$, scaled by the prior probabilities, $p(S)$ and $p(L) = 1 - p(S)$. The optimal subjective time at which to switch is when the posterior odds are equal (heavy short vertical lines at the points where the distributions intersect). Increasing the prior odds of a short trial (lower panel) shifts this optimum to the right (toward later times). In accord with the well-established scalar variability in subjective elapsed times, the standard deviations of the distributions are proportional to their means.

The distributions of switch latencies from each session appear to be basically normal (with a small extra component we will discuss below). This offers a
natural interpretation of the observed behavior: the mean of the normal
distribution is thought of as the mouse’s target switch time (the time at which the
mouse, on any given trial, reaches its decision criterion to switch sides) and the
standard deviation is thought of as a measure of the noise that comes from the
mouse’s systems of measuring time and enacting their decision to switch once the
decision is made.

2. Method
The subject’s cage contains two feeding hoppers, designated as “short-side” and
“long-side” hoppers, respectively and a control hopper. Subjects were first trained
on a matching task, in which the feeders connected to the two feeding hoppers
become armed at random intervals and poking in the hoppers will then release a
pellet. Next, they underwent an autoshaping procedure, in which one of two
feeding hoppers would light up and a reinforcer would be delivered after a set
delay. The delay for the left (short) side was three second and the delay for the
right (long) side was nine seconds. This prepares subjects for the primary task: an
interval-timing task. In this task, a trial is initiated when the subject pokes its head
into the control hopper. Once a trial begins, an unsignaled draw from a Bernoulli
distribution determines whether the trial is a “short trial” or a “long trial.” The
uncertainty about whether the trial will be short or long is the exogenous
stochastic variable, which we manipulate. On a short trial, reward is determined
by the first poke after three seconds have passed: if this first poke is on the short-
side hopper, the subject is reinforced; otherwise the trial ends. On a long trial,
reward is determined by the first poke after nine seconds: if this first poke is on
the long-side hopper, the subject is reinforced; otherwise the trial ends. Subjects have no way of knowing which trial type they are in until it has ended, but they can (and do) get reinforced on nearly every trial by poking at the short-side hopper until more than three seconds (but less than nine seconds) have elapsed and then switching to poking at the long-side hopper. The uncertainty about how much time has elapsed since the start of the trial—upon which the decision to switch from the short to the long hopper depends—is the endogenous stochastic variable.

Subjects were C57/BL6 mice, run in the 24-hours-per-day, 6-days-per-week automated testing system and were analyzed by the Time Stamp analysis system both described in Gallistel et al. (2010). This means that subjects lived in the testing environment 24 hours per day, subsisted entirely off food earned in the experiment and had their actions cataloged and time-stamped as they occurred. Analyses were performed in quasi-real time (twice per day) to keep careful track of their progress in learning the task and adapting to new conditions.

During the active periods (9pm-11pm and 4am-8am) subjects would first experience a timeout period whose length was determined by a draw from an exponential distribution with expectation of sixty seconds. After the timeout period, the middle of the three hoppers (the control hopper) would become illuminated, signaling that the subject could initiate a trial by poking there. Once a poke at the control hopper was detected, the timer started, the lights in the left and right hoppers came on (while the light in the control hopper went out) and a trial
began. With no signal to the subject, the trial type was determined by a draw from a Bernoulli distribution whose p-value was experimentally manipulated (and ten percent of the time was an autotrial). If the trial was a short trial, the subject was reinforced if and only if the subject’s first poke after three seconds had elapsed was on the left side. If the trial was a long trial, the subject was reinforced if and only if the first poke after nine seconds was on the right side. In practice, subjects began poking at the left hopper immediately after the start of the trial and switched to the right hopper if the trial did not end. The end of a trial was signaled by the two hopper lights going out. The end of a trial was followed by the inter-trial timeout period described above and the next trial was armed.

3. Results

Depending on p, the optimal switch point for a mouse changed: if long trials were more likely, not switching too late became more important for reward maximization than the danger of switching too early, so the optimal switch time occurred earlier, which can be seen in figure 2: the scatter plot of the subjects’ responses (black circles) responds to the change in the probability of trial type (density of red crosses above vs below) shortly after the trial at which the change occurs (black dashed line). It is worth noting now that the majority of the analyses in this paper will be aimed at making more rigorous what is visible to the naked eye in this plot of the raw data: the behavioral shift in response to the change in the stimulus comes shortly after the change and is abrupt.
Figure 2. Scatter plot of switch latencies. Each black dot represents the latency of a switch on a long trial and the blue lines represent the median of these latencies in a given session. The red crosses that appear in lines above and below the black dots mark long and short trials, respectively. The vertical lines mark session boundaries at which the relative frequencies of the short and long trials changed. Note the change in density of the lines of red crosses that occur at these boundaries. Note further that when the density increases on the top line (and decreases on the bottom), the distribution of black dots shifts downward, away from the denser red crosses. Similarly, when the density of red crosses increases in the bottom line and decreases in the top line, the distribution of black dots shifts upward, again, away from the denser line of red crosses. The seemingly unprovoked changes marked by red dots were caused by unilateral feeder malfunctions.
Figure 3. The cumulative distribution of switch times from a single subject under two different probabilities of a long trial (.9 on the left and .1 on the right). All switches between the “temporal goalposts” (3s and 9s) are reinforced, but we see a strong shift toward shorter wait times when the probability of needing to switch (probability of a long trial) is high.

We found that mice shifted their target switch time in accordance with the change in the optimal switch time (we see a distribution shift to the left when the probability of a long trial increases, as in figure 3) and that this change was both abrupt and followed shortly after the change in \( p \). First, we replicated the findings of Balci et al., in that the subjects showed statistically significant differences in the means of the distributions of their switch times when the probability of a short trial was manipulated.
Also in line with the findings of Balci et al., we found that subjects missed very few of their possible reinforcers. Each plot in figure 4 shows two functions of a subject’s decision criterion and their timing precision. First, it shows the pair of values that best describe the subject’s behavior in that session (these are the roughly concentric circles). Second, it shows the proportion of total rewards possible that would be received for each pair of values from a hypothetical subject that went through the same sequence of trials (these are the upside-down u curves). In order to get above 99% of the possible rewards, a subject must have a mean-sd pair that falls below the lowest red curve. Note that as a hypothetical subject’s timing becomes more precise (points closer to the x-axis) the subject can freely vary its decision criterion without greatly changing its expected reward. In the degenerate case, when precision is perfect, subjects could put their decision criterion anywhere in the three to nine second range and still get all of the rewards. Conversely, the worse a subject’s timing precision, the narrower the range of mean values that will produce a given level of proficiency.

Despite the fact that many subjects exhibit temporal precision that would allow fairly large variability while still staying within the range of values that would garner 95% of all possible rewards, subjects tend to have decision criteria close to the mode of the expected value function at their level of precision (the x-value for the top of the closest upside-down u function). This suggests that they are able to perform at near-optimal levels, even when doing so has little impact on their net reward – a calculation that requires sensitivity to one’s own uncertainty/variability which has also been seen in humans (Trommershäuser et al. 2008, 2003, Behrens
et al. 2007). Finally, there was virtually no overlap between 95% confidence
intervals under different short-trial probabilities, which underscores the fact that
subjects were shifting their temporal decision criteria based on the frequency of
short trials.

Figure 4. Each subject’s behavior (black concentric circles) falls near the expected
reward contour (upside-down u-shaped curves) that yields 99% of the total
rewards possible in the task. That is to say, the subjects were reinforced 99% of
the times that they were given the opportunity to be. Such ceiling-level
performance was typical on all the transitions. Assuming that subjects cannot
influence their timing variability, the y-position of their behavior is fixed and they merely select the x-component.

Also similar to Balci et al., we typically found between 5-10% premature switches in our data, possibly due to better quantification. These are trials in which subjects switched sides even before the 3s short latency had expired. This was a significantly higher rate than could be explained by inaccuracies in a subject’s internal timer: for a small portion of trials, subjects were simply not doing the task, but rather switching sides immediately\(^2\). Following the analysis of Balci et al., this led us to model our distribution of switch times as a mixture distribution; a normal component for the timing behavior aimed at the target point, and a small component to account for the mode we saw below the 3s mark. Our analysis departed from Balci et al. in that we modeled the distribution of switch times by a mixed exponential-Gaussian distribution rather than a mixed Weibull-Gaussian distribution. This form of distribution accounts for the data as well as the Weibull-Gaussian mixture, but has one fewer parameter. The fact that these early switches

\(^2\) Like the trials in which subjects failed to do the timing task by simply switching immediately, there were also a few sessions with slightly more timeout trials than could be accounted for by stochastic variability. These were taken to be trials at which the mouse was “asleep at the wheel” and these were excluded. These were too few in number, however, to materially affect the analyses discussed here.
are well-modeled by an exponential distribution may suggest that this impulsive behavior is caused by some constant-rate neural process and certainly warrants further study. For the purposes of this study, however, we were primarily interested in how quickly the Gaussian component of a mouse’s distribution adapted in response to a change in short-trial probability.

The question of "how quickly do the mice react to the change in p?" splits up into two more specific questions: 1. how many trials does the mouse need to detect the change in the p value? and 2. Once the change begins, how abrupt or gradual is it? The answer to the first question speaks to how good the mice are at detecting changes in probability; the second speaks to the method by which they change their decision criterion: gradually trial-by-trial or determining a new target value and abruptly switching to it.

In order to answer these questions, we considered two-parameter (corresponding, roughly, to location and abruptness) models of the change. We considered several forms for the model of the transition (Gaussian, Weibull and linear) but they were all essentially similar: we know the animal's behavior at the beginning and we know it at the end, the question is just how the behavior transitions from one pattern to the other. The different forms of functions yielded similar results so we
will discuss only the linear class of model for simplicity's sake.

Figure 5. In the range of values plausible for our analyses, the form of the transition function had negligible impact in the range of parameter values that were plausible. The transitions were too arupt to discriminate between the forms we tested:

<table>
<thead>
<tr>
<th>Model</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>( L(x) = \min(\max((\text{start}-x)/\text{duration},0),1) )</td>
</tr>
<tr>
<td>cumulative Gaussian</td>
<td>( G(x) = \Phi(x; \text{start}+\text{duration}/2, \text{duration}^2) )</td>
</tr>
<tr>
<td>cumulative Weibull</td>
<td>( W(x) = \text{start} + W(x; \text{start}, \text{duration}) )</td>
</tr>
</tbody>
</table>

The linear model states that, given trials 1 to \( M \) (in which p-value has changed somewhere) the animal's behavior is drawn from exponential-Gaussian distribution1 for trials 1 through \( k_1 \) and the animal's behavior is drawn from
exponential-Gaussian distribution2 for trials k2 through M. The behavior between k1 and k2 is drawn from an exponential-Gaussian distribution whose parameters change linearly from those of distribution1 to those of distribution2.

The majority of p-value changes show the same result: the animal's behavior seems completely shifted to distribution2 by the time we have just a few measurements from that animal. Looking at a contour plot of the likelihoods for various pairs of location and gradualness, we can see that both values are low, meaning that the change is centered shortly after the change in stimulus and is completed rapidly.

Indeed, across subjects and sessions, we see variation in how wide our confidence intervals are (this is often tighter when we have a high percentage of long trials because we can only take measurements at long trials, which means more data with which to constrain our model in the low range of values) but we typically see the peak of the marginal likelihood functions close to zero (in the case of the start of the transition, it is often too low to be taken at face-value for a priori reasons). This means that the most likely model for the data observed is one in which subjects begin the change in their behavior shortly after the actual change and complete the change in their behavior in a small number of trials.
Figure 6. The marginal distributions on the location likelihood (number of trials after actual change in stimulus probability at which linear transition begins) and duration of the behavioral shift (span of trials from start to end of linear transition function) for a sample of eight transitions.

Looking at the marginal likelihood of various transition durations, we see that the log-likelihood at zero is higher than typical likelihood for other transition durations, showing that a simpler, one-parameter, model actually does a better job of explaining this data: the transition is a single step and the only free parameter is where that step occurs.

All these analyses were also conducted using a 10-parameter model: one in which the four parameters for the expgauss distribution before the transition and the four
for distribution after the transition are also allowed to vary. We used Metropolis-Hastings MCMC to sample from the likelihood function of this 10-parameter model and found that the maximally likely values for the start and duration of the transition were the same in the 10-parameter model as in the 2-parameter model – integrating out the other 8 parameters which represented the distributions before and after the change, did not change the basic shape of the marginal likelihood functions for the two parameters of interest\(^3\). The likelihood functions derived from the 2-parameter model were essentially the same as those obtained by applying the 10-parameter model and averaging out the other 8 parameters.

\(^3\) We also confirmed that the maximally likely exponential-Gaussian distributions obtained by fitting the 10-parameter to the switch behavior before and after the change were the same as the exponential-Gaussian distributions we obtained by fitting each data set individually.
Figure 7. Cumulative distribution of locations (number of trials after actual change in probability of a long trial) and durations of the behavioral shifts subjects exhibited in experiment 1. The CDFs are shown using various criteria for the cut-off point: Maximally likely start/duration, the lowest start/duration with at least ½ the maximum likelihoods, 1/5 maximum, 1/20 maximum and the expectation of the likelihood distribution. The ½ max and expectation criteria seem to best summarize the data for our purposes, but none of our analyses hinge on choosing one criterion over the others.

In addition, we ran a version of the experiment in which the change in probability occurred within a session rather than at the start of a session. This manipulation
decreased the strength of the effects discussed above but the pattern of results was the same, as can be seen again from a scatter of the raw data.

Figure 8. Raw scatter of the data from Experiment 2. The plot is similar to that of experiment 1 but green lines now indicate the points at which the probability of a long trial changed, as these no longer coincided with session boundaries. Blue circles mark the subjects’ switch behavior in seconds, red crosses above and below the plots show the history of long and short trials.

We see similar results in our optimality analysis, though less tightly aligned with the center of the optimality curve.
Figure 9. The alignment between the subject’s behavior and the center of the optimal area given their timing precision was not as tight as in experiment 1.

And we see that the marginal likelihood functions again favor transition functions with a short duration that start soon after the actual change in p-value, which look similar to the distributions seen in experiment 1.
Figure 10. The cumulative distributions of the start and duration of the behavioral shift are in line with the results from experiment 1. The marginal likelihood distributions are heavily skewed toward instantaneous changes that occur shortly after the actual change in the probability of a long trial.

Thus we have seen that, regardless of whether the objective probability changes are at session boundaries or within-session, subjects altered their switch times shortly after the change, abruptly and accurately to the maximally rewarding value given their level of timing precision.

4. Discussion
One can imagine proficient performance on our task from two basic types of machines: machines that try various switch criteria, sticking with the ones that
work best and machines that calculate near-optimal switch criteria given the data available to them. Our data suggest that our subjects behaved more like the second type of machine than the first because of four basic empirical findings about the shifts in their switch latencies made in response to the changes in the stimulus probability:

1. Behavioral shifts occur shortly after the stimulus probability change (as soon as there is objective evidence that there is been a change, see Figure 11).

2. Behavioral shifts are completed in few trials; many shifts are completed so rapidly that the transition can't be distinguished from a step.

3. Behavioral shifts are accurate. They are often closer to the optimal value than could be expected to be found with the limited amount of feedback subjects received from missed pellets.

4. Behavioral shifts are based on detecting changes.

4.1 Shifts occur shortly after changes

Speaking to the first point, we have seen that the marginal likelihood distributions are skewed in favor of shifts that begin shortly after the objective changes in the probability of a long trial. In addition, we can examine, given the particular sequence of long and short trials a subject saw on a given transition, the objective odds that a change has occurred at the trial that subject began its shift in behavior (figure 11). In the plot of the cumulative distribution of these odds under various criteria for deciding where the behavioral shift actually began, we see that in both experiments, a large proportion (50% in Experiment 1, 30%-40% in experiment 2) of changes begin by the trial at which the objective odds reach 3:1. This
supports the assertion that behavioral shifts begin nearly as soon as there is substantial evidence that a change has occurred – that learning occurs quickly when trials are surprising as in Courville et al. (2006).

Figure 11. The cumulative distribution of the objective odds that a change had occurred as of the estimated start of a transition. The different colors are for different estimates of the start (maximum likelihood, expectation, and confidence limits of increasing stringency--see legend).

We see changes with poor information in favor of them due to the confluence of two factors: our model being poorly constrained in the low values where it is most likely and the abruptness with which odds rise after a change. These two together mean that estimating a change point as too early by even a few long trials (as is
most likely true in some of our cases) results in a behavioral shift estimated to have started before there was good evidence for a change in probability.

In some cases, our estimates of the start of the transitional shift are so low (and therefore unconstrained by the data in that range) that our best estimate for the start of the change are before substantial evidence has accrued that the change had occurred. Because of the rapid rate at which the odds grow, estimating the start of the transition even a small number of trials early can cause the odds at the start to be very low indeed. Likewise, estimating a few trials late can cause the odds to be very high, which is why we see the starts of transitions with sometimes well over $10^4$ odds in favor of there having been a change.
Figure 12. The likelihood function for the location of the start of the behavioral shift superimposed with the objective odds that a change in probability occurred (based on the sequence of trials leading up to the current one). Estimates for the start of the behavioral shift are in dashed lines for various decision criteria: \( \frac{1}{2} \) maximum likelihood (blue), \( \frac{3}{5} \) maximum likelihood (cyan), \( \frac{1}{20} \) maximum likelihood (red), and the expectation of the likelihood function (green). Note that the abruptness of the rise in the odds that an objective change has occurred means that we see high variability in the odds when a behavioral shift is estimated to start. This is especially true when the location of the start is in the neighborhood of the sharp increase in odds.
4.2 Shifts were abrupt

Once behavioral shifts begin, they are completed quickly. The marginal likelihood functions for the gradualness parameter in both experiments favor shifts so abrupt as to nearly be indistinguishable from a step. If we compare the hypothesis that subjects make step-like shifts to the hypothesis that they made gradual shifts (uniform prior over the range of plausible gradualnesses) we see that the Bayes factor favors a step change 10:1 for over 70% of changes in experiment 1 and for over 50% of changes in experiment 2.

Figure 13. Cumulative distribution of of the odds on the null (values larger than 1 favor a step change, rather than a gradual one). In Experiment 1, odds were in
favor of a step change for 75% of transitions, with over 50% over 100:1. In Experiment 2 and 85% of transitions favored the null with 50% over 10:1.

More to the primary aim of our experiment, both experiments support a model of subject behavior that explicitly represents the probability of a long vs. a short trial and a subject’s own timing precision. We see in figure 4 that approximately optimal selection of one’s decision criterion depends both on the frequency of long vs. short trials and on the subject’s timing precision. The fact that subjects’ observed temporal decision criterion was close to the optimal for their level of timing precision suggests that their decision criterion is set as a function of these variables. This obviously means that the variables must be represented by the subject.

4.3 Shifts Not Based on Differential Feedback

Fundamentally, reinforcement models base behavior on the subject’s level of reinforcement. Such models cannot account for the change in behavior we observe for two reasons. First, for many changes, subjects shifted their behavior with little difference in their level of reinforcement. That is to say, the rate of pellet loss shortly after the change in the probability of a long trial (solid lines in figure 14), when subjects are adapting to the change in their stimulus probability, is nearly indistinguishable from the rate of pellet loss leading up to that change (dashed lines in figure 14). Reinforcement learning models therefore have no difference in the level of reinforcement received by the subject to motivate the
observed shift in behavior.

Figure 14. The number of trials to the midpoint of the behavioral shift, n, is determined by the criterion used for the start and duration of the shift as in figure 7. One can then look at the n trials after the change in probability to find the number of reinforcers missed while the subject adjusts to the new probability of a long trial. We can then compare this number with the number of reinforcers missed in the n trials leading up to the change in probability (when the subject is already adjusted to the previous value) to see if that number is typically different from the number missed while the subject gathers evidence that a change has occurred and shifts its behavior accordingly. These two values are compared in figure 11 for varying criteria on the start and duration of the behavioral shift.

That is not to say that their behavior does not shift: it becomes closer to optimal given their level of timing precision, as we saw in figure 4. It is rather to say that this task, despite being dependent on monitoring a hidden variable as well as their own uncertainty, is so easy for subjects that even when they are getting nearly every reward possible, they are able to adjust their behavior to stay comfortably on top of the hill of expected rewards.

Second, as figure 7 shows, in 30% of cases the subjects were half way through their behavioral change before a single reinforcement was missed. Since these
shifts in behavior preceded even a one-pellet difference in the level of reward, the level of reward could not contain any information about the existence of changes that would then drive the shifts in behavior. Hill-climbing methods of learning cannot explain learning in these circumstances because the hill the subject needs to climb suddenly moves when the probability of a long trial changes and yet the subject remains on the peak of the reinforcement hill without losing a substantial number of reinforcers in figuring out where the new top of the hill is. In general, parameter tuning by trial-and-error cannot account for our data because subjects had so few trials that would inform their tuning before their behavior changed--keeping in mind that they do not switch at all on short trials and so get no feedback on the appropriateness of their target switch time.

On the basis of these data, a transition is typically half complete after 15 trials (fully complete after 20) in the first experiment (figure 7) and half complete after 19 trials (fully complete after 26) in the second (figure 10). In that number of trials, subjects are likely to have missed 2-3 pellets or 4-6 pellets respectively (figure 11).
Figure 15. Cumulative distribution of the number of pellets missed after the probability value changed but before the animal was half way through its behavioral shift, using various likelihood criteria for the start and duration of the behavioral shift. Note that in both Experiment 1 (left) and Experiment 2 (right), in the majority of instances, very few pellets were missed before the subject changed its behavior. Moreover, the number of pellets missed over this period (solid lines) is almost indistinguishable from the number missed during the (same number of) trials before the shift (dashed lines). This underscores how quickly behavioral shifts are made relative to the feedback the subject receives about the consequences of its behavior.
We see a similar pattern of results in the second experiment, in which session boundaries no longer function as a possible demarcation between probability values. In both cases, we can see that the cumulative distribution of pellets lost between the change in probability and the precipitant change in behavior is not distinguishable from the distribution of pellets lost in the same number of trials leading up to the change in probability. We also still see a large number of cases in which shifts begin before a single pellet is lost.

4.4 Shifts Are Based on Detecting Changes

Having shown that our data cannot be explained by a model that relies on the level of reinforcement to drive behavior, the question still remains as to whether the pattern of behavior we see is the result of what Dayan (2011) calls model-based control as opposed to model-free control. Dayan characterizes this difference as how much of behavior is driven by inference as opposed to experience. In our case, it is perhaps more useful to characterize the distinction as how much of the world must be represented by an agent. For example, we have shown that an agent cannot exhibit the behavior we have observed by merely doing calculations on changes in its rate of reward, it must represent some estimate of the probability of a long vs a short trial. Obviously, there are many aspects of the stimulus that an agent need not represent: changes in ambient lighting, temperature or noise, or (less obviously) the ninth-order relationships between times of long trials in one session with long trials three sessions later (this would be a very difficult calculation which would result in no predictive power in our task). We argue that the quick and stable behavioral shift we see is
the result of a model-based control process that represents changes in the probability of a long versus a short trial as opposed to simpler models that adjust to changes in probability quickly by using an estimate of long trial frequency that is only impacted by the last few trials.

We know that subjects must use an estimate of the probability of a long trial and that getting an estimate of probability from experience requires some sort of averaging. If this averaging places significant weight on trials in the distant past, the model will not be able to respond to a change in probability as quickly as our subjects did. So, we checked, for various small memory sizes m, whether our results could be simulated by a simple strategy that based switch time off of the frequency of long trials in the last m. These simple strategies are not reinforcement strategies because they are based on an estimate of the probability of a long trial but they do not explicitly represent a change in this probability. Rather, they adjust to a change quickly by not allowing anything more than a few trials back to influence behavior. Thus, the change in the stimulus causes the behavioral shift without the subject’s having to represent that a change occurred.

To test whether such a strategy could explain our data, we sorted all switches by the frequency of long trials preceding them and then compared, via Bayes factor, whether the switches seemed to be drawn from the same distribution or from different distributions. We did this for several sizes of memory (1, 3, 5) with little difference in our results. We found (as in figure 15) that the Bayes factor favored the null hypothesis (that both sets were drawn from the same distribution roughly 90% of the time in Experiment 1 and roughly 80% of the time in Experiment 2.
This shows that switch behavior cannot be driven solely by the frequency of long trials in the last few. In particular, it shows that the animal is not simply adjusting its switch point based on whether the last trial was long or short (this is the degenerate case, when memory size is 1). It is the amount that recent history diverges from more distant history that signals a change and is the key component to adjusting to changes quickly in the way that our subjects do.

Figure 16. Cumulative distribution of the weight of the evidence in favor of the alternative hypothesis: that in comparing two groups of switch times (one with \( k \) long trials preceding it and one with \( j \) long trials preceding it where \( k \neq j \)) the two groups were drawn from different distributions. The analysis is repeated with memory sizes 1, 3 and 5. The distributions clearly favor the null hypothesis.
If the switch behavior is independent of the recent frequency of long trials, this variable cannot be what is driving the shift in switch behavior after a change in probability. Of course we do not claim that something beyond the sequence of long and short trials comprising the stimulus that drives the subject’s behavior. Rather, the simple models are missing a necessary feature: the ability to detect a change in probability. This requires more than probability estimation; it requires storing a probability value and comparing that value to a recent probability estimate – determining the evidence that the current value is truly different from the historical value.

4.5 Summary

We replicated the findings of Balci et al. in that the mice were near-optimal at the task and showed highly significant changes in their behavior in response to the change of the hidden variable. Using a slightly more parsimonious model of the distribution of their switch behavior (exponential-Gaussian mixture rather than Weibull-Gaussian) we expanded their analysis to quantify how long the change in behavior took after the change in the hidden variable.

Subjects detected changes in the underlying probability of a long trial soon after they occurred, even when that change was in the middle of a session – which are what Dayan (2011) takes to be the hallmarks of model-based control. They then adjusted to that change rapidly, suggesting that subjects entertained distinct options for temporal decision criteria and moved to one or another when sufficient evidence had been accumulated, rather than gradually tuning their decision criterion to maximize their reward. Behavior shifted quickly and abruptly and
appeared to be relatively stable once the change was made, rather than exploratory. The changes were also made with very little negative feedback with which they might have judged one criterion against another. Their criteria were near optimal given the timing precision they displayed, despite the fact that the reward curves were often fairly flat at these precision levels. This means that they selected near-optimal criteria even when doing so did not have much effect on their earnings. Finally, observed performance could not be simulated by simple models which did not take into account the subject’s uncertainty or detect changes in probability. This pattern of results suggests a computational system that represents the relative probabilities of the two types of trials, evidence that a probability has changed and the system’s own uncertainty. The ability to represent risk and uncertainty (at least on some basic level available for calculation) appears to be present in creatures at least as evolutionarily distant from humans as the mouse. In a sense, this should come as no big surprise given that all animals have survived by navigating an uncertain and risky world.
5. Bibliography


