Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Games and Economic Behavior 52 (2005) 410–423
GAMES and Economic Behavior
www.elsevier.com/locate/geb

Deconstructing the law of effect

C.R. Gallistel

Department of Psychology and Rutgers Center for Cognitive Science, Rutgers University, 152 Frelinghuysen Rd., Piscataway, NJ 08854-8020, USA

Received 8 June 2004

Available online 30 November 2004

Abstract

Do the consequences of past behavior alter future policy, as the law of effect assumes? Or, are behavioral policies based on behaviorally produced information about the state of the world, but not themselves subject to change? In the first case, stable policies are equilibria discovered by trial and error, so adjustments to abrupt changes in the environment must proceed slowly. In the second, adjustments can be as abrupt as the environmental changes. Matching behavior is the robust tendency of subjects to match the relative time and effort they invest in different foraging options to the relative incomes derived from them. Measurement of the time course of adjustments to step changes in the reward-scheduling environment show that adjustments can be as abrupt as the changes that drive them, and can occur with the minimum possible latency. Broader implications for theories about the role of experience in behavior are discussed.

Economists and psychologists commonly assume that behavior is shaped by its consequences. Psychologists call this the law of effect, by which they understand that we and other animals try different behaviors, assess their effects, and do more of those with better effects and less with those with worse. On this view, the behaviorally important consequence of a behavior is the information it provides about behavioral outcomes. The effect of the information is to alter policy.

There is, however, a different way in which the consequences of behavior may shape future behavior. Some policies depend for their execution on information about the state

E-mail address: galliste@ruccs.rutgers.edu.

0899-8256/$ – see front matter
doi:10.1016/j.geb.2004.06.012

of the world. Changing the information fed to a policy changes the behavior it generates. Because behavior generates information about the state of the world, the effects of past behavior may alter future behavior by changing the model of the world that a fixed policy takes as input.

Attempts to distinguish between these possibilities are rare. Three collaborators and I (Gallistel et al., 2001) have recently distinguished between them in studying the genesis of matching behavior, which is the robust tendency of animals, human and otherwise, to prorate their behavioral investments in different foraging options so that the investment proportions match the income proportions (Herrnstein, 1961; Harper, 1982; Godin and Keenleyside, 1984; Davison and McCarthy, 1988 Herrnstein, 1991). If the subject gets $2/3$ of its income from foraging in one location and $1/3$ from foraging in another, then it spends $2/3$ of its foraging time in the first and $1/3$ in the other.

In the laboratory, matching behavior is most commonly studied using what is called free operant behavior with concurrent variable interval schedules of reinforcement. In this paradigm, subjects have two different reward-generating response options. Typically, if they are pigeons, the options are two different keys, either of which may be pecked; if they are rats, the options are two different levers, either of which may be pressed. The rewards for pecking or pressing are typically small amounts of food. The behaviors are called free operants because they operate on the environment to produce reward and because the subject's opportunity to engage in them is not constrained. Subjects can make either response whenever they like.[1] A schedule of reinforcement is the experimenter-determined function relating investment to reward. In a variable interval schedule, the next reward delivered by a response on one of the two options is set up at a randomly varied interval after the harvesting of the last reward from that option. Once set up, the reward remains available until it is harvested by the next response. The expected interval to the next set-up distinguishes one variable interval schedule from another. The schedules for the two response options are concurrent when they run in parallel, with the timers for both schedules running regardless of which option the subject is exercising at the moment.

The fact that the reward-scheduling function is called a schedule of *reinforcement* suggests the extent to which the law of effect is taken for granted by psychologists. It is assumed that rewards act to strengthen rewarded behaviors, that is, to make them relatively more likely to occur. In a purely descriptive sense, of course, they do: the shorter the expected set-up interval for one schedule is, relative to the other, the more likely it is that at any given moment the subject will be investing in that option.

It does not follow, however, that the schedule of reinforcement affects the subject's behavior by way of an effect on the subject's policy. Subjects may have a fixed policy for translating relative expected incomes into relative investments. In that case, what they get from responding is not policy guidance but rather an estimate of the income to be expected. The income from an option is the amount of reward it yields per unit of time *tout court—not* per unit of time invested. Provided that the subject samples an option at intervals that are on average shorter than the expected interval to the next reward, the income from a variable

[1] For matching to occur, a minimal amount of time (on the order of a second or two) must be lost in shifting between the options. Otherwise, subjects can in effect exercise both options at once (play both machines simultaneously), which is what they do.

interval schedule is only weakly affected by the size of the subject's investment. In short, to experience the income from an option a subject must spend some time exercising that option, but, within broad limits, the amount of time it spends has little effect on the income it yields. The subject's behavior reveals, so to speak, the income that may be obtained from a given option.

There is some reason to think that matching might be an innate policy, because both human and pigeon subjects pursue it even under circumstances where it is the worst policy, the policy that minimizes their overall return (Herrnstein, 1991). At the very least, this implies that there are limits to the ability of response consequences to shape policy.

1. Two contrasting accounts

One of the attractions of addressing the issue of the role of past behavioral consequences in the determination of future behavior by considering the matching phenomenon is that it permits a clear formulation of the alternative accounts. The first account involves what Herrnstein called melioration, which is "the process of comparing the rates of return and shifting toward the alternative that is currently yielding the better return" (Herrnstein and Prelec, 1991, p. 361). Some version of this idea has been the basis for most attempts to explain matching behavior, although none of these attempts has succeeded in specifying the details in such a way as to yield a model that captures the details of the behavior (Lea and Dow, 1984; Herrnstein and Prelec, 1991). A particularly vexing problem has been the specification of the interval over which subjects average when estimating their returns. The wider this averaging window, the more slowly subjects will approach the new stable equilibrium when the relative richness of the schedules changes. Melioration models have never been able to specify an empirically defensible averaging window (Lea and Dow, 1984).

What melioration models have in common is the assumption that matching is not itself the policy. The policy is whatever melioration leads to. Matching is what melioration leads to when there are variable interval schedules of reinforcement, because in that environment, matching equates returns. Because subjects sample both options at intervals shorter than the expected intervals between rewards, the income-limiting factor is the expected set-up interval of the schedule. Increasing or decreasing the expected duration of a visit—hence, the average investment in an option—has little effect on the income realized from it. Return is income divided by investment. Therefore, increasing the investment in the richer option and decreasing the investment in the poorer decreases the return from the richer and increases the return from the poorer. When the investment ratio matches the income ratio, the returns are equal. Matching is the dynamic equilibrium point, the point at which the consequences of behavior (the returns) do not favor a shift toward either option. Any drift away from this point, produces a countervailing inequality in returns, which drives the behavior back toward matching.

The alternative account assumes that matching *is* the policy and that it is an immutable policy (Gallistel and Gibbon, 2000; Gallistel et al., 2001). In accord with the experimental findings on the microstructure of matching behavior (Heyman, 1979; Gibbon, 1995), this model assumes that visits to the options are terminated by a Poisson (random rate) process.

When a visit has begun, subjects, in effect, repeatedly flip a biased coin to decide when to leave (that is, to temporarily stop exercising the option). When the coin comes up heads, they leave. This assumption has two consequences: first, the distribution of visit durations should be exponential, which it is (Heyman, 1979: Gibbon, 1995). This is odd if one believes that the function relating behavior to its consequences shapes behavior, because the reward for trying an option becomes more certain the longer a subject has neglected it. Thus, the probability of terminating a visit to try the other option should increase as the visit is prolonged; the longer the subject has been there, the more likely it should be to leave. This, however, is empirically false; the probability of terminating a visit does not change as the visit is prolonged, which is why visit durations are exponentially distributed.

Second, the expected duration of a visit is determined by the bias on the coin, the rate at which it comes up heads. This parameter of the subject's behavior is assumed to be determined by its estimates of the expected incomes from the available options in accord with the following two equations:

$$E(d_1)/E(d_2) = \widehat{H}_1/\widehat{H}_2 \quad \text{and} \tag{1}$$

$$\frac{1}{E(d_1)} + \frac{1}{E(d_2)} = a\left(\widehat{H}_1 + \widehat{H}_2\right) + b, \tag{2}$$

whence

$$\frac{1}{E(d_1)} = a\widehat{H}_2 + b\frac{\widehat{H}_2}{\widehat{H}_1 + \widehat{H}_2} \quad \text{and} \quad \frac{1}{E(d_2)} = a\widehat{H}_1 + b\frac{\widehat{H}_1}{\widehat{H}_1 + \widehat{H}_2}, \tag{3}$$

where $\widehat{H}_i$ is the subject's estimate of the income to be expected from option i; and $E(d_i)$ is the expected duration of a visit to option i. The policy consists in setting the bias of the coin for a given side so that the rate at which it comes up heads is given by (3). As indicated in (3), these leaving rates are the reciprocals of the expected visit durations.

Equation (1) sets the ratio of the expected stay durations for the two options equal to the ratio of the estimated incomes. It builds matching into the policy. Equation (2) makes the sum of the leaving rates on the two sides a linear function of the sum of the incomes. The greater is the combined income, the higher are both leaving rates, and the shorter are the expected stay durations. This is known to be empirically true (Gallistel et al., 2001). And, it makes functional sense, because it means that the rate at which a subject cycles between the options is adjusted in accord with the expected interval between rewards from the two options combined. In impoverished environments, where rewards come infrequently, the subject cycles slowly; in rich environments, it cycles rapidly. Its rate of cycling is adjusted to the expected interval between rewards so that the schedules remain the income-limiting factor.

In summary, on the first account, the subject's model of the world is maximally simple: It consists only of the labels that distinguish the response options. The subject's policy is defined by the probabilities of its engaging in those two options. Pursuing this policy yields over time estimates of the returns from the two options of experimental interest: the rewards received divided by the time invested in obtaining those rewards. The policy-adjustment rule (the learning rule) is melioration: increase the probability for the option with the greater return and decrease the probability for the option with the smaller return. In

an environment where the investment is not the income-limiting factor, meliorating eventually equates the returns, bringing the behavioral system to an equilibrium state. On this account, matching is not a strategy that is specified a priori; it is the outcome of a feedback process. Therefore, it cannot be attained abruptly. It cannot happen in less time than it takes to obtain reliable information about the returns to be expected from policies intermediate between the initial policy and the equilibrium policy, because the system must adopt those intermediate policies and evaluate their returns en route to the equilibrium state.

On the second account, the subject has a more complex model of the world: it has experience-derived estimates of the incomes currently to be expected from rapidly sampling those options. The critical experiential variable associated with an option is expected income, not return, that is, rewards per unit of foraging time, not rewards per unit of time invested in that option. Note that in the computation of income, the behavioral investment that generated it plays no role, whereas in the computation of return, the behavioral investment is the denominator.

When this system learns, what changes are its estimates of the expected incomes, not its policy. (More will be said later about how the estimates depend on experience, that is, about the learning rule.) This system has no policy-changing rule. It shifts investment back and forth between the options according to a fixed rule, which takes as its input the estimates of currently expected incomes. These estimates determine the parameters of the stochastic process that terminates visits (limits the expected duration of an investment in an option). Matching in this system is purely feed-forward. Hence, nothing limits the abruptness with which behavior can shift from one stable investment pattern to another. A large step change in the estimates of the currently expected incomes can produce an equally large and equally step-like change in expected visit durations. The latency between a step change in the environment (the reward schedules) and the answering change in behavior is determined by the properties of the income estimator (the learning mechanism). The more efficient this estimator is, the shorter the latencies will be.

2. A critical experiment

We (Gallistel et al., 2001) determined which kind of process—feedback or feedforward—mediates matching by measuring the time course of the behavioral adjustment to step changes in the schedules of reward. As just noted, the feedforward account allows for these adjustments to occur with step-like abruptness, at a latency determined only by the efficiency with which the income-estimating mechanism can detect changes in the expected incomes. By contrast, a feedback process like melioration cannot adjust abruptly, because matching is the result of a process of equilibration. To get from the pre-change equilibrium to the post-change equilibrium, the process must spend enough time in several intermediate states to obtain reliable (option-differentiating) estimates of the returns. How long it spends in these intermediate states and how many of them there are depend on the width of the return-averaging windows and the magnitudes of the policy adjustments following each assessment of relative returns. As already noted, an empirically successful specification of these dynamic parameters has eluded attempts to elaborate a model of matching based on melioration. For present purposes, however, it does not matter what values these dynamic

parameters might be assumed to have. No assessment of relative returns can be made until the subject has tried both options at least once. Estimates of return based on single visits are extremely noisy (Gallistel et al., 2001, Fig. 14). If the subject made adjustments to its option probabilities based on the relative returns from single visits, the adjustments would often be in the wrong direction (opposite the direction dictated by the true values of the expected returns). Thus, if subjects use a maximally narrow window for estimating returns, they will often waste time moving in the wrong direction in policy space. If, to avoid this, they get more reliable estimates of the expected returns by widening their averaging windows, it will take more time to obtain those estimates. Either way, it seems inescapable that the shift from one equilibrium to a radically different one can only proceed slowly; the change must be spread over a great many visit cycles.

The subjects in our experiment were rats with electrodes implanted in their brains at a locus that produces an intense, non-satiating rewarding experience. (For the latest theorizing about the relation of that experience to naturally occurring experience, see Shizgal, 1997.) The environment was a Plexiglas box with two levers. They were located in the recesses of two different alcoves, so that it took the rats at least a second and a half to shift from one lever to the other. Holding down the levers generated rewards on concurrent, independent variable interval schedules. The subjects were tested in daily two-hour sessions.

The experiment had several phases, but in the phase of interest here, we used five pairs of schedules (VI 7.1-s/VI 62.5-s, VI 8.55-s/VI 25.64-s, VI 12.82-s/VI 12.82-s, VI 25.64-s/VI 8.55-s, and VI 62.5-s/VI 7.1 s). All pairs summed to the same overall rate. The corresponding ratios of scheduled rates of reward were: 9/1, 3/1, 1/1, 1/3, and 1/9. The sum of the scheduled rates in each pair (the sum of the reciprocals of the VI's) is 9.4 rewards per minute. Because the scheduled rates were the primary determinants of the rates of rewards actually experienced, the overall rate of experienced reward was approximately the same regardless of which pair of schedules was in effect.

In the phase of interest here, the pair of schedules in force at the beginning of each session could not be predicted from the pair in force at the end of the preceding session. Moreover, at an unsignaled point somewhere in the middle 80 minutes of each 2-hour session, the schedules initially in force were replaced by a different and equally unpredictable pair. Thus, subjects in this phase encountered frequent and unpredictable step changes in the relative rates of reward permitted by the two schedules, both between and within sessions. Some of these changes were large; some were small. The questions are, how abruptly did subjects adjust to these changes and how long did it take for these adjustments to appear? We focused particularly on the within-session changes, because they were unsignaled: the subject had no way of judging when a change had occurred except by processing the sequence of experienced rewards.

The changes we observed were abrupt, as can be shown by two different graphic displays. In the first (Fig. 1), we plotted the cumulative duration of the visits to one side against the cumulative duration of the visits to the other. The slope of this cum–cum plot at any point is the ratio of the average visit durations at that point. A change in the slope indicates a change in the expected durations of the visits to the two levers. Abrupt changes in the expected durations of the visits give rise to sharply defined inflection points in the cum–cum function. The examples in Fig. 1, which are representative of the more than 100 within-

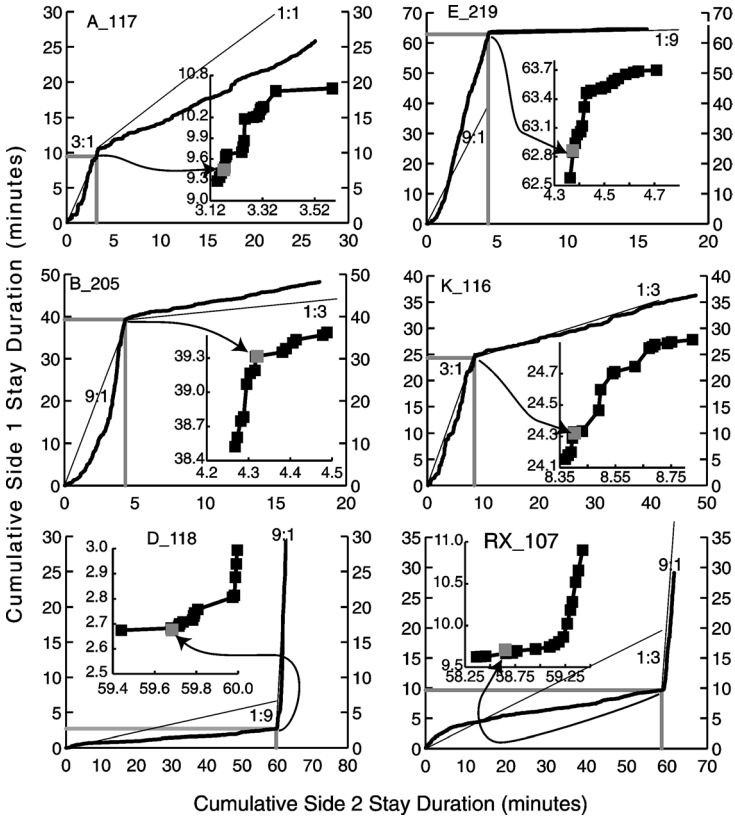

Fig. 1. Six examples of abrupt behavioral change in response to a step change in the schedules of reward: the cumulative duration of the visits on Side 1 plotted against the cumulative duration of the visits to Side 2. The slope of this plot at any point is the ratio of the expected visit durations at that point. Abrupt changes in slope indicate abrupt changes in the ratio of these expectations. The thin lines with ratios near them (e.g., 1:1) indicate the ratios of the programmed rates of reward. When the plot parallels this line, the subject is approximately matching. The insets show the plots in the vicinity of the abrupt slope changes on a visit-by-visit scale. Each data point marks the end of a visit cycle (one visit to each side). The gray vertical lines on the main plots and the gray squares on the insets indicate the point at which the reward schedules changed (the step change in the environmental forcing function).

session transitions we observed in this phase of the experiment, are extremely abrupt. The insets show the function at a visit-cycle-by-visit-cycle level of resolution; successive points are the cumulative durations at the conclusion of successive visit cycles. From the insets, it may be seen that the transition from one stable set of expected visit durations to a very different stable set often occurred over a span of fewer than 5 visit cycles. Sometimes, the entire transition appeared to occur within a single visit cycle.

In the second way of showing the time-course of the transitions (Fig. 2), we used a Bayesian model to calculate the probability density functions for the estimates of the two reward rates and for the estimates of the two leaving rates. The probability density functions for the estimates of the reward rates were calculated after the delivery of each reward.

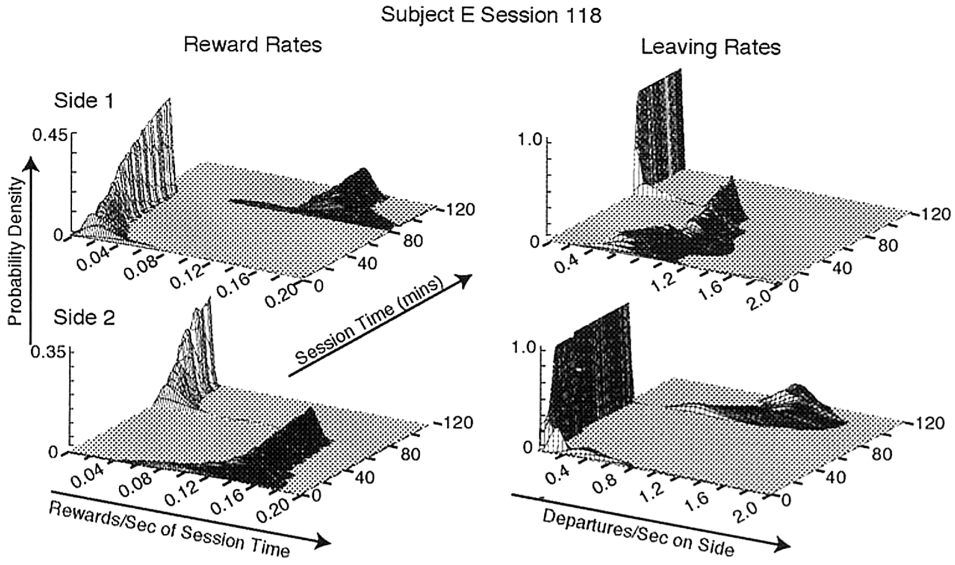

Fig. 2. Probability density functions for the estimates of the rates of reward on the two sites (left two plots) and for the leaving or departure rates from each side (right plots) versus session time. The expected visit duration for a side is the inverse of the departure rate for that side. Notice that the change in the location of the pdfs for the behavioral estimates are just as step-like as the changes in the estimates for the reward rates.

The probability density functions for the estimates of the two leaving rates were calculated at the conclusion of each visit cycle. The Bayesian estimator "knew" a priori that there would be a step change in each pair of rates at some point. Thus, it calculated not only successive probability density functions for each estimate but also the probability that the change had occurred. The calculation was structured in such a way that a probability density function for a rate estimate did not necessarily take into account all of the data up to that point in the session. When the data indicated with high probability that the change in rates had occurred at an earlier point in the session, a probability density function only took into account the data since the estimated change point. Because this sophisticated estimator did not use an averaging window, step changes in the parameters being estimated gave step changes in the location of the probability density functions for the estimates, as may be seen in Fig. 2.

The most important thing to note in Fig. 2 is that the changes in the location of the probability density functions for the estimates of the rats' leaving rates are just as step-like as the changes in the locations of the probability density functions for the estimates of the reward rates. We know that the latter changes were steps. Thus, the observed change in expected visit durations was as abrupt as a known-to-be step change in the expected incomes.

The central conclusion from this experiment for present purposes is that when there is a step change in the environmental forcing function, the resulting change in matching behavior is much more abrupt than would be possible if matching were the equilibrium state of a system that adjusts its policy on the basis of the effects of that policy on the

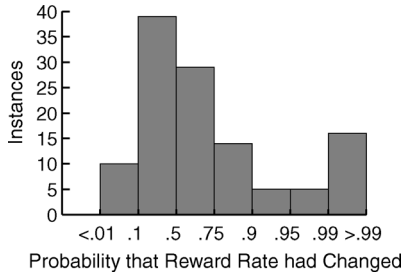

Fig. 3. Histogram showing results of calculation in which, for each of more than 100 change sessions, we took the mode of the probability density function for the behavioral change (the time in the session at which the behavioral change was maximally likely to have occurred) and calculated what the probability was at that point in the session that the reward rate had changed. Most of the behavioral changes occurred while this probability was still low, that is, they occurred as soon as there was any evidence for a change in reward rates.

relative returns. The behavioral changes approximate steps. Whatever the process is that produces matching, it must be capable of producing step changes in the expected visit durations in response to step changes in the environment.

It is of interest to know not only how abrupt the behavioral changes are but also what their latency is. How quickly do rats detect changes in the expected incomes from the options they are sampling? To answer this question, we compared the rat's behavior to the behavior of a Bayesian ideal detector. The above-described Bayesian estimator for the reward rates and leaving rates also gave a probability density function for the temporal location of the change-point (probability density versus session time). We took the location of the mode of the probability density function for the behavioral change point as the point in the session at which the behavioral change was maximally likely to have occurred. We then applied our Bayesian change detector to the sequence of experienced rewards to calculate the probability that the (anticipated) change in the reward rates had occurred, as of that moment in the session. Figure 3 is the histogram of the results of this calculation. What it shows is that the changes in behavior occurred as soon as there was any appreciable likelihood that the reward rates had changed. The rats could not have adjusted any sooner if they had been getting real-time advice from a professional statistician. They approximated ideal detectors of changes in rates of reward. Although the rat is seemingly blind to the consequences of its own behavior, it is exquisitely sensitive to the state-of-the world revealed through that behavior.

3. The learning rule

The locus of learning in a feedforward model is in the processes that update its model of the world. Changes in the parameter estimates in this model produce changes in behavior. The learning rule my collaborators and I proposed has two components—one for detecting changes in incomes and one for estimating the incomes currently expected.

Detecting a change in a random rate process is equivalent to detecting an inflection point in the cumulative record of the events it has generated (Fig. 4). When the rate is constant, the slope of the cumulative record is constant; when it changes, the slope changes. The

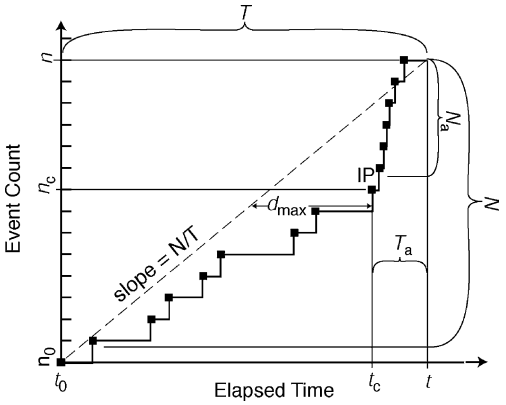

Fig. 4. Detecting a change in the rate parameter of a random rate process is equivalent to detecting a change in the slope of the cumulative record of the events generated by that process (event count versus elapsed time). This diagram shows the quantities involved in the computation. For the computation itself, see text.

detection of an inflection point proceeds in two steps. First, as each new event is added to the cumulative record, the system estimates the putative inflection point, which is the past event at which a change in slope, if such there be, most likely occurred. From purely geometric considerations, this event must be at or close to the event at which the cumulative record deviates maximally from a straight line from the origin of the record to the point on the record corresponding to the current moment (the dashed line in Fig. 4). The slope of this line is N/T, where N is the event count at the moment and T is the duration of observation (the interval over which events have been counted). In other words, it is the estimate of the rate parameter, on the assumption that it has been constant. The point on the cumulative record that deviates farthest from this line is the putative inflection point.

The second stage of the change-point detecting algorithm calculates the log of the odds against the null hypothesis that there has been no change in the rate. On the null hypothesis that the N recorded events have been randomly distributed in time, the probability, p_e, that any one of them (ignoring event order) falls in the interval T_a is T_a/T. The probability p_f of observing N_a or fewer events in N "tries" is given by the cumulative binomial probability function, as is the probability P_m of observing N_a or more events. When the number of events in T_a is approximately the expected number $(p_e N)$, then the ratio P_f/P_m is approximately 1, so the log of this ratio is approximately 0. As the observed number of events since the putative time of change becomes improbably low, the ratio becomes very small, and its log approaches minus infinity. As the observed number becomes improbably high, the ratio becomes very large, and its log approaches infinity. The absolute value of the log of this ratio (the logit[2]) is the subject's measure of the strength of the evidence that there has been a change in rate. When this decision variable exceeds a critical value, the

[2] The logit is usually defined to be the ratio of two complementary probabilities. Our ratio is between two overlapping probabilities, which therefore do not sum to 1. We use overlapping probabilities because the resulting measure is better behaved when the expected and observed numbers of events are the same and near or equal to zero. Away from unity or when the expected number of events is $\gg 0$, our ratio approximates the usual ratio.

subject perceives a change. When it perceives a change, it truncates the data at the moment the change is perceived to have occurred (the moment t_c in Fig. 4). The data on which the next perception of a change in rate is based are only those after this moment.

This algorithm is our model of the component that detects changes in rates. It reports not only that a change has occurred but also when in the past it is most likely to have occurred, namely, at the putative inflection point. Our model of the component that estimates the current rate is simple: the moment at which a change in rate is detected is always later than the estimate that the algorithm gives for the moment at which the change occurred. The estimate at that moment of the currently prevailing rate is the number of events recorded since the estimated change-point divided by the interval between that point and the point where the change was detected. If, as can happen, there is no event in that interval, the default value for the event count is 1. In this model, the estimates of the current rates of reward are not the result of running averages; there is no averaging window. They are based on the small sample of events observed in the interval backward from the time at which a change was last detected to the time at which it was estimated to have occurred.

Because there is no averaging window in this model, it is time-scale invariant. It works equally well regardless of the time scale set by the rates of event occurrence. Models with averaging windows, which is to say most other models for matching (Lea and Dow, 1984; Staddon, 1988), impose a time scale when they specify the width of the averaging window. The model then works only for events on that time scale. To get these models to work under more general conditions, one needs to have many different averaging windows. This then poses the problem of deciding which window is appropriate for the current circumstance. For a discussion of the importance of time-scale invariance in learning, see Gallistel and Gibbon (2000, 2002).

This completes the specification of the learning rule in our feedforward model of matching. Whether the model proves in the long run to be right or not, it serves to illustrate the distinction between systems that obey the law of effect and systems that merely appear to do so. The differences between the two kinds of systems tend to be dramatic because the locus of learning is fundamentally different. In the second kind of model, fixed policies depend on models of the state of the world. In the first kind (policy-changing learning rules), flexible experience-derived policies reduce the need for models of the world, which is why this kind of model is preferred by the partisans of empiricist theories of mind. To the empiricist school of thought, the nativism and rationalism implicit in the computations required to generate a good model of relevant aspects of the world are uncongenial, as is the assumption that behavior is generated by innate and more or less immutable policies.

4. Diverse reasons why matching is a good policy

If matching is an innate and immutable policy seen in the behavior of a wide range of broadly successful animal species, then one tends to assume that it must be, all things considered, a good policy. For the case we have been considering—concurrent variable interval schedules—it has been shown to be not the optimal policy but to be so nearly optimal as not to matter from a practical standpoint (Heyman and Luce, 1979). In this environment, the animal increases its total income by sampling the poorer option because

the very small cost of this sampling in terms of a reduced income from the better option are more than offset by the additional income realized from the visits to the poorer option. In appreciating why this is so, it is crucial to remember two points. First, the longer the subject has been away from a poor option, the more likely it is to have a reward waiting to be harvested. Thus, rare visits are strongly rewarded. Second, if animals cycle rapidly enough between the options, then the schedules are income limiting, not the investments. Thus, the reduction in the investment in the richer alternative has little effect on the income realized from it.

But this is only one environment and arguably a very artificial one. What about other environments? What about, for example, environments in which the investment is the income-limiting factor? In such environments, putting all of one's investment in the option that pays off more frequently is the best thing to do. Somewhat surprisingly, the matching policy yields this result in this environment, as was first realized by Herrnstein and Loveland (1975). In the laboratory, this environment is implemented with concurrent variable ratio schedules. A variable ratio schedule looks not at the time that has elapsed since the last reward was harvested but rather at the number of responses that have been made. In other words, it rewards investment at some random rate. Variable ratio schedules differ in the rate at which investment is rewarded. Under these conditions, investment in an option rewarded at a lower rate is counterproductive. Asymptotically, subjects do not invest in the poorer option. They spend all their time investing in the better one.

In such an environment, there is a destabilizing positive feedback between the environment and the matching policy. Whenever the subject increases its investment in the richer option and reduces its investment in the poorer, the difference in the incomes realized becomes greater than it was before the change in behavior. The income ratio, as a proportion of total income, keeps ahead of the investment ratio, as a proportion of total investment, until both ratios reach 1. To see this, one must recall that the income is the number of rewards experienced divided by the time over which they were experienced, not by the time invested in the option. The more a subject reduces its investment in a variable ratio schedule—that is the less frequently it tries that option—the fewer the rewards realized from it. It is a matter of simple algebra to show that the only point in behavioral space at which the subject's relative investment in the better option equals its relative income from that option is when it invests all of its time in that option (Herrnstein and Loveland, 1975). In other words, the matching policy leads to maximizing behavior in (at least some) environments where that is the optimal behavior.

It would probably strike an economist that one of the most artificial (unnatural) aspects of the common laboratory paradigm is the lack of competition; the subject is in the environment by itself, with no competitors. Central to an economic perspective is that foraging for goods is a competitive activity; the optimal strategy depends fundamentally on an estimate of what your competitors may be expected to do. Interestingly, matching has been shown to be an evolutionarily stable strategy for a community of *ideal free* competitive foragers (Fretwell and Lucas, 1969; Orians, 1969).

An *ideal* forager has accurate knowledge of the relative richness of the available environments. A *free* forager is one whose access to those environments is not effectively blocked by other foragers. To say that matching is an evolutionarily stable strategy is to say that:

(1) the adoption of the strategy by the community as a whole does not create a selection pressure favoring an alternative strategy, and
(2) competitors that adopt a divergent strategy are selected against.

Roughly speaking, the reasons for this are the same as those that make matching the equilibrium policy for the melioration process: matching distributes the foragers across the environment in such a way that the number of foragers per available reward is everywhere the same. It equates return per local forager across locations that differ in richness (or, at least, the opportunity to earn a return).

A third reason why matching is a good policy has to do with sampling considerations. In a world where what was a poor option yesterday may be the best option today, a good policy never stops sampling the options. Matching based on income estimates has that property; the estimated income from an option cannot go to zero in a finite period of observation, because the upper limit on the estimate is one over the observation time. Thus, even options that have never paid off get sampled, albeit with ever decreasing frequency.

Sampling considerations may also be relevant to understanding the otherwise puzzling fact that visit durations (the durations of individual investments) are exponentially distributed, which means that the probability of winding up a visit in order to try the other option is independent of the duration of the visit. Because the world is only observable one small part at a time—because the animal can only exercise one option at any one time—estimating the temporal structure of the income schedule for an option is no small challenge. Not all options pay off in the random ways so far considered. Some pay off at predictable intervals. We know that animals are sensitive to temporal predictability (see Gallistel and Gibbon, 2002, for review and modeling; Gallistel, 2003). When a time series is sampled periodically, the sampling may introduce bogus temporal structure. This is the phenomenon of aliasing, which is most familiar through its effects on the apparent direction of wagon wheel rotation in old films. Aperiodic sampling prevents aliasing, and sampling through an exponential distribution of inter-sample intervals is aperiodic, because a sample is equally likely to be taken at any moment.

5. Why innate policies may be better than learned policies

As the above discussion shows, diverse considerations affect what is the best policy to follow in a complicated world. It is hard to know even what all the important considerations are. It may be true that if one uses trial and error to find the return-maximizing policy and if one evaluates returns over many different time windows, then one will eventually converge on the policy that is optimal policy for the world as it really is. However, it may take a very long time to converge on that policy (Wolpert and Macready, 1997), and, in the long run, we are all dead. Environments change-on many different time scales, including time scales longer than a subject's lifetime. Some policies work superbly in some environments and disastrously in others, but you may not know when you have moved from one environment to the other until it is too late. To survive for generation after generation, animal species need policies that work better than most others in most environments actually encountered, and that are not disastrous in any environment with a non-negligible probability. It may take

much more than one lifetime to accumulate enough experience to arrive at such policies by trial and error, even intelligent trial and error (Wolpert and Macready, 1997). That may be why the policies animals follow are built in rather than shaped by their experience. Because they have served animals like that well for tens of thousands of generations, they embody the wisdom of the ages.

References

Davison, M., McCarthy, D., 1988. The Matching Law: A Research Review. Erlbaum, Hillsdale, NJ.

Fretwell, S.D., Lucas, H.L., 1969. Habitat distribution in birds. Acta Biotheoretica 19, 16–36.

Gallistel, C.R., 2003. Conditioning from an information processing perspective. Behavioural Processes 62, 89–101.

Gallistel, C.R., Gibbon, J., 2000. Time, rate and conditioning. Psychological Review 107, 289–344.

Gallistel, C.R., Gibbon, J., 2002. The Symbolic Foundations of Conditioned Behavior. Erlbaum, Mahwah, NJ.

Gallistel, C.R., Mark, T.A., King, A.P., Latham, P.E., 2001. The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. Journal of Experimental Psychology: Animal Behavior Processes 27, 354–372.

Gibbon, J., 1995. Dynamics of time matching: Arousal makes better seem worse. Psychonomic Bulletin and Review 2 (2), 208–215.

Godin, J.-G.J., Keenleyside, M.H.A., 1984. Foraging on patchily distributed prey by a cichlid fish (Teleosti, Cichlidae): A test of the ideal free distribution theory. Animal Behaviour 32, 120–131.

Harper, D.G.C., 1982. Competitive foraging in mallards: Ideal free ducks. Animal Behaviour 30, 575–584.

Herrnstein, R.J., 1961. Relative and absolute strength of response as a function of frequency of reinforcement. Journal of the Experimental Analysis of Behavior 4, 267–272.

Herrnstein, R.J., 1991. Experiments on stable sub-optimality in individual behavior. American Economic Review 81 (2), 360–364.

Herrnstein, R.J., Loveland, D.H., 1975. Maximizing and matching on concurrent ratio schedules. Journal of the Experimental Analysis of Behavior 24, 107–116.

Herrnstein, R.J., Prelec, D., 1991. Melioration: A theory of distributed choice. Journal of Economic Perspectives 5, 137–156.

Heyman, G.M., 1979. A Markov model description of changeover probabilities on concurrent variable-interval schedules. Journal of the Experimental Analysis of Behavior 31, 41–51.

Heyman, G.M., Luce, R.D., 1979. Operant matching is not a logical consequence of maximizing reinforcement rate. Animal Learning Behaviour 7, 133–140.

Lea, S.E.G., Dow, S.M., 1984. The integration of reinforcements over time. In: Gibbon, J., Allan, L. (Eds.), Timing and Time Perception. Annals of the New York Academy of Sciences 423, 269–277.

Orians, G.H., 1969. On the evolution of mating systems in birds and mammals. American Naturalist 103, 589–603.

Shizgal, P., 1997. Neural basis of utility estimation. Current Opinion in Neurobiology 7, 198–208.

Staddon, J.E.R., 1988. Quasi-dynamic choice models: Melioration and ratio invariance. Journal of the Experimental Analysis of Behavior 49, 303–320.

Wolpert, D.H., Macready, W.G., 1997. No free Lunch theorems for search. IEEE Transactions on Evolutionary Computation 1 (1).