The Dynamics of Behavior: Review of Sutton and Barto: *Reinforcement Learning: An Introduction* (2nd ed.).

J. E. R. Staddon

Duke University

jers@duke.edu

There are two main theoretical approaches to behavioral data: *parsimonious* and *comprehensive*. Examples of parsimonious approaches are Weber's law and the matching law: an empirical regularity is described by an equation with one or two or even no free parameters. Comprehensive theory seeks to reproduce complex phenomena with as many free parameters as necessary. Examples are evolutionary theories (e.g., McDowell, 2018; Edelman, 1987) and many types of neural net modeling, which may incorporate reinforcement learning but may also adapt in other ways (supervised and unsupervised learning). The 3Blue1Brown web tutorials give an excellent account of reinforcement learning in the context of neural networks. A brief review of the various types of computational learning theories is Moldakarimov & Sejnowski, 2008.

Artificial neural networks have many arbitrary features (i.e. free parameters): the number of elements ('neurons'), the number of layers, the learning rules, etc. These features are arbitrary in the sense that the system can learn adequately with a wide variety of configurations. Comprehensive theories like this are not *unique*. On the other hand, like Darwin's theory, they may correctly identify the process — evolution by natural selection — while being open as to the details — genetic transmission (which was discovered a bit later). They are in some ways closer to engineering than science: they aim to predict and simulate rather than explain. Reinforcement learning is an intermediate case, not very parsimonious (there are several possible update rules and neural network configurations), but sometimes simple enough to qualify as a parsimonious theory.

Richard Sutton and Andrew Barto's *Reinforcement learning: An introduction*[1] (Cambridge: MIT Press, 2018) is the 526-page second edition of a 322-page book first published in 2002. The

[1] An electronic draft version is available at: http://incompleteideas.net/book/bookdraft2017nov5.pdf. An accessible text that overlaps with this one is the classic *Parallel Distributed Processing: Explorations in the microstructure of cognition*, by D. E. Rumelhart and James McClelland. https://academiaanalitica.files.wordpress.com/2016/11/david-e-rumelhart-james-l-mcclelland-pdp-research-group-parallel-distributed-processing_-explorations-in-the-microstructure-of-cognition_-foundations-vol-11986.pdf

books follow a line of work pioneered by U.S. Air Force research scientist A. Harry Klopf and summarized in his 1982 *The hedonistic neuron*. The "hedonism" reference, the idea that artificial systems are capable of goal-directed behavior, can perhaps be traced to an even older, seminal paper "Behavior, purpose, and teleology" by Rosenblueth, Weiner, and Bigelow (1941).

Reinforcement learning (RL) is learning via consequences. To this extent, the topic has relevance to operant conditioning, which is defined as change in behavior caused by the contingent presentation of a reinforcer or punisher. The focus of this book is largely mathematical, a textbook of tools — and sometimes spectacular simulations (Chapter 16). It begins with three and a half pages of notation definitions. The following 17 chapters cover topics such as Tabular Solution Methods, Finite Markov Decision Processes, Dynamic Programming, Temporal-Difference Learning, On-policy Prediction with Approximation, Eligibility Traces and two chapters explicitly concerned with psychology and neuroscience. Several example problems are discussed in Chapter 16: Samuel's checkers player, the game of Go, and others.

With similar aims to RL, but not referenced in these books (despite a wealth of other historical material), is W. Ross Ashby (e.g., his 1960 book *Design for a brain*). Ashby's *law of requisite variety* describes a kind of behavioral homeostasis based not on a single variable, but on a range of variables, such as the problem of regulating body weight via foraging in a complex environment. In every case, the idea is that an adaptive system operates by selecting — either among a set of actions or among possible values of a single action — an option that promises to improve some measure of the system's state. Ashby's law simply recognizes the fact that if, say, body temperature is affected by many variables, then the adapting body must be able to cope with all of them.

Ashby's view is Darwinian. He recognizes that the secret ingredient in all adaptive behavior is *variation*, the repertoire the organism has available to cope with challenges offered up by the environment. Without variation there may be no effective behavior from which the environment can select. Hence a theoretical scheme that is almost exclusively concerned with selection, like RL, is limited in its application to Darwinian problems. I begin by describing reinforcement learning in a bit more detail and contrasting it with the Darwinian alternative via an example.

The basic idea behind both Ashby's law and RL is that all goal-directed behavior embodies some kind of *negative feedback.* In RL, the feedback is usually immediate (contiguous; in some versions delayed feedback may also have an effect via memory decay called an *eligibility trace*). As the environment changes, the system acts in ways that somehow reduce the discrepancy between its current state and a desired state.

Sutton and Barto begin with a very general definition:

> Reinforcement learning is learning what to do…so as to maximize a numerical reward signal….Reinforcement learning…is simultaneously a problem, a class of solution methods that work well on the problem, and the field that studies this problem and its solution methods. (p. 1)

This definition is illustrated (p. 10) with an equation describing *temporal-difference* (TD) learning:

$$V(S_t) \leftarrow V(S_t) + \alpha[V(S_{t+1}) - V(S_t)]. \quad (1)$$

In other words, at each time step, the value of the current state is updated by a fraction, α, of the difference between the previous state, $V(S_t)$, and the new state, $V(S_{t+1})$.

Adapting to temporal schedules

Here is a simple example: adapting to a fixed-interval (FI) schedule. A timer is assumed. The task of the learner is to adapt to the prevailing interval by adjusting wait time to a fixed fraction of the interval. A simple model tracks the interreinforcement interval (IRI) according to Equation 1, updating its estimate at each reinforcement. If y is the model's estimate of the time it should wait after reinforcement before beginning to respond, the *wait time*, and x is the actual IRI at the point of reinforcement, then:

$$y(t + 1) = y(t) + \alpha(x(t + 1) - y(t)) \quad (2)$$

which is Eq. 1 applied, reinforcement by reinforcement, interval by interval, to this situation.

Figure 1 is a simulation; it shows a typical cumulative record where $x(0) = x(t) = 100$ (arbitrary scale), $\alpha = 0.5$, responding occurs at a fixed rate after the wait time has elapsed, and reinforcements occur at the points of inflection. The wait time, short initially, soon converges on 50% of the interreinforcement interval. Reinforcement rate is constant.

This model is incomplete (the timer is assumed) and much too simple, of course. Nevertheless, it does predict some dynamic features of FI responding. For example, suppose the inter-reinforcement interval is not constant (i.e., not a fixed-interval schedule), but depends on the preceding wait time: $x(t + 1) = \beta y(t)$ (an *autocatalytic schedule*). Then, the model will either wait longer and longer or shorter and shorter, depending on whether β is greater or less than one.

Figure 2 shows a simulation of the first few intervals of such a schedule when the first inter-reinforcement interval is 10, $\alpha = 0.5$, and $\beta = 2$. The intervals, and the wait times, get longer and longer. Reinforcement rate continues to decline. This behavior is clearly maladaptive; after all,

the model simply needs to wait for a short time every now and then — to show some variability — to reset the inter-reinforcement interval to a shorter value. It fails to do this, of course, because this version of reinforcement learning is entirely deterministic: there is no variable (stochastic) term in the model.

But real organisms may be no smarter. **Figure 3** shows 16 experimental sessions of waiting time data from one of four pigeons, all well-trained: first on FI and then on an autocatalytic schedule like the one just described (Wynne & Staddon, 1988, Fig. 9). The labels indicate the value of parameter[2] β: A, B > 1; c, d < 1. This bird (and three others not shown here) showed results similar to the prediction of this very simple RL model. When $\beta > 1$, waiting times get longer and longer; when < 1, shorter and shorter. Evidently, the pigeon, like the model, is tracking successive inter-reinforcement intervals and like the model, seems not to sample by occasionally waiting a short time.

Naïve animals might well show more variability and might, therefore, not fall into the trap offered by an autocatalytic schedule (unfortunately this experiment has not been done). But the pigeons in this experiment were very well trained, so neither the pigeon nor the model in this example are really *learning*. The pigeon has *already* learned, not a response or even a time interval, but a *program* (Staddon, 1981), a rule it follows interval-by-interval. This *habit* (an older, and in some ways better, word for *operant*) is quite stable, as the bird's vulnerability to the autocatalytic schedule shows.

The real learning in situations like this is the development of the program, which has two ingredients: variability: the ingredients to be assembled into the habit must occur so that configurations closer to the optimum can be selected. The second ingredient is of course the mode of selection which, for a hungry pigeon is some kind of learning through contiguity. The process of shaping by successive approximations begins to give some idea of the ingredients of the final habit; but for tasks that are at all complex, we really know very little how the bits arise and how they are put together.

RL deals only with the selection aspect. The program is defined for the selector ahead of time, by the structure of the problem such as game, a *k*-armed bandit, a maze or some other problem with a well-defined tree structure of possible outcomes. I return to this issue in a moment.

---

[2] These parameters are re-scaled to the model; see Fig. 9 for the actual values.

Optimality

The idea that a problem solution is "optimal" is always part of Sutton and Barto's discussion of various forms of RL. The book is, after all, a discussion of a set of tools for control problems, not a list of learning theories. Usually, some limitations are recognized. For example, in Chapter 6 on temporal-difference learning (TD) in a section titled "Optimality of TD (0)" they write: "…TD (0) converges deterministically to a single answer independent of the step-size parameter…as long as [it] is chosen to be sufficiently small" (p. 126), which recognizes a limitation on supposedly 'rational' solutions that is often neglected.

In this book and other places, such as economics,[3] there is sometimes a confusion between optimality as a *goal* and as a *process.* A process may have an optimal solution; *but it never by itself defines what it means to be rational*. A method for achieving a goal is not the goal itself. Here is well-known example, the traveler's dilemma (Basu, 2007). where this limitation was not fully acknowledged:

> Lucy and Pete, returning from a remote Pacific island, find that the airline has damaged the identical antiques that each had purchased. An airline manager says that he is happy to compensate them but is handicapped by being clueless about the value of these strange objects. Simply asking the travelers for the price is hopeless, he figures, for they will inflate it.
>
> Instead he devises a more complicated scheme. He asks each of them to write down the price of the antique as any dollar integer between 2 and 100 without conferring together. If both write the same number, he will take that to be the true price, and he will pay each of them that amount. But if they write different numbers, he will assume that the lower one is the actual price and that the person writing the higher number is cheating. In that case, he will pay both of them the lower number along with a bonus and a penalty – the person who wrote the lower number will get $2 more as a reward for honesty and the one who wrote the higher number will get $2 less as a punishment. For instance, if Lucy writes 46 and Pete writes 100, Lucy will get $48 and Pete will get $44.
>
> What numbers will Lucy and Pete write? What number would you write?

---

[3] For example, the popular math site Math24 (https://www.math24.net/optimization-problems-economics/) describes optimality in this way: "Example 3: The demand function for a certain product is linear and defined by the equation $p(x)=10-x/2$, where x is the total output. Find the level of production at which the company has the maximum revenue." The problem is to be solved by differential calculus. This a process to *discover* a goal, whereas optimality is more usually defined as *setting* the goal ahead of time. The goal may then be attained in several ways.

The game-theory solution to this dilemma is $2, via an "if A does this I should do that", repeated until they drop down to the lowest number, which is the 'rational' solution. Basu writes "To see why 2 is the *logical choice*, consider a plausible line of thought that Lucy might pursue: her first idea is that she should write the largest possible number, 100, which will earn her $100 if Pete is similarly greedy…Soon, however, it strikes her that if she wrote 99 instead, she would make a little more money…" [emphasis added] and so on, an obviously sub-optimal (irrational) solution arrived at by supposedly rational means.

Sutton and Barto point out that their TD-learning example only works if the increments permitted are small enough. On the other hand, small increments in the traveler's dilemma problem lead to a very non-optimal result; but subtracting 5 instead of 2 before comparing yields a very different result, closer to the common-sense optimum, $100 — because $95 and $97 both look worse than $100 to both players. Step size matters in problems like this which have local maxima.

Conclusion: Both TD learning and the game-theory "hill-climbing" analysis of the traveler's dilemma are just processes. They may arrive at a final state that is optimal from one point of view, but they don't define what it means to be optimal. Optimality is an outcome, not a process, even though economists, particularly, very often confuse the two. Just because, for example, the optimal allocation of a budget between two or more goods can be discovered by equating marginal utilities does not mean that the process the consumer deploys to arrive at her choice involves anything of the kind.

Variation

The idea that (operant) learning is a trial-and-error process does not fit easily into the deterministic algorithms of RL. The authors add exploration to optimization by distinguishing between *greedy* and *nongreedy* moves*:*

> If you maintain estimates of the action values, then at any time step there is at least one action whose estimated value is greatest. We call these the greedy actions. When you select one of these actions, we say that you are exploiting your current knowledge of the values of the actions. If instead you select one of the nongreedy actions, then we say you are exploring, because this enables you to improve your estimate of the nongreedy action's value. (p. 26)

How should the model decide between greedy (optimal) and nongreedy (exploratory) actions? The authors discuss a number of possibilities, ranging from random nongreedy choice to planned optimality: "In any specific case, whether it is better to explore or exploit depends in a complex way on the precise values of the estimates, uncertainties, and the number of remaining steps" (p.

26). Ideally, the structure of the problem should be known in advance before the model decides on whether to sample or not. This is not a useful way to look at any evolutionary process which, almost by definition, lacks foresight.

Sampling — what RL calls "nongreedy" choice — is an add-on, not integral to the RL model. A more evolutionary approach like Ashby's, however, accepts that variation always exists and that organisms only converge on the so-called greedy (optimal) response under extreme conditions: after much training and while experiencing a high reinforcement rate. Under all other conditions, other sources of variability — past history, "instinctive" behaviors, etc. — remain strong and will sometimes displace the optimal response unless it is very strongly selected (Staddon, 2016). In other words, the greedy-vs.-nongreedy decision is handled by competition rather than a pre-arranged strategy or a speed difference (as in Kahneman's (2011) "fast" and "slow" thinking).

Temporal Learning

A single chapter (14) in the Sutton and Barto book (out of a total of 17) deals with the psychology of learning — classical and operant conditioning. Based on a paper on which Sutton was a co-author (Ludvig et al., 2012), the chapter presents a comprehensive model of classical conditioning that embodies the Rescorla-Wagner model, but now including temporal variables, as a special case. The temporal difference approach can duplicate not just the standard blocking effects, but also remote associations and a sort of temporal contrast effect first demonstrated by Egger and Miller (1963). The approach is also intriguing to neuroscientists because of "an uncanny similarity between the behavior of temporal-difference algorithms and the activity of dopamine producing neurons in the brain…" (p. 21) during classical conditioning. This and other applications to neuroscience are discussed in Chapter 15.

The treatment of operant (aka instrumental) learning in the book is cursory, probably because the book is aimed at computer scientists and mathematicians, not experimental psychologists. And perhaps also because it does cover many issues that overlap operant and classical conditioning. The chapter does not treat real-time dynamics. The authors point to, for example, "quantitative precision and real-time performance" of IBM reinforcement-learning-based WATSON, the *Jeopardy* winner; but they nowhere present actual real-time data. On the other hand, it is worth noting that WATSON automatically derives confidence estimates for a set of possible answers each *Jeopardy* question. As one commentator notes: "Machine learning algorithms have metacognition at their core" (Nabavi, 2018). But predicting real-time dynamics should be a "natural" for any real-time theory, as in the autocatalytic example and further discussed below.

One variant of RL that has some relevance for learning psychology allows for an outcome at one instant to affect all the behavior leading up to that point. This is done via what is called an

eligibility trace, which represents an exponentially decaying memory factor that modulates the effect of a later outcome on the value of earlier choices.

Eligibility traces retain a memory of past events, potential time markers. An obvious application of this idea is to timing — temporal discrimination. Focus on this problem might have a restorative effect on theory in operant conditioning. The field began with Skinner's brilliant technical advance in the form of the cumulative record. Skinner's method drew attention to the real-time dynamics of the behavior of individual organisms. But, seduced by the deceptive orderliness of averages, theorists emphasized steady-state models such as the 'ticking-clock' scalar-expectancy theory (SET) that ignored the learning process. Carefully crafted experimental procedures such as time-left (see discussion by de Castro & Machado, 2010) always use a strong time marker, such as a long blackout period. The sensitivity of temporal discrimination to the quality, the memorability, of the time marker therefore went largely unexamined (see extended discussion in Staddon, 2010, Chapter 13).

Criticism of SET was accompanied by some attention to the dynamics of timing in the late 1990s (see Staddon & Higa, 1999 and accompanying commentaries). But elegant data of ancient vintage that could have provided cues to more ambitious real-time theory were for a long time ignored. Ferster and Skinner (1957) show numerous cumulative records of pigeons learning different fixed- and variable-interval schedules after experience with continuous reinforcement (i.e., food for every peck). They summarize the three-step process of FI acquisition: first a burst of responding after each reinforcement; then steady "VI-like" responding; and, finally the "scallop" pattern characteristic of steady-state FI behavior (Figure 117). These real-time basics of acquisition present a promising target for reinforcement learning and eligibility traces. They were not tackled by theorists for many years.

A paper by Machado (1997) that appeared 40 years after Ferster & Skinner describes a model which successfully duplicates the FI-acquisition pattern (Figure 4) and shows how the wait time evolves to be a fixed fraction of the programmed interreinforcement interval. Machado's theory involves an eligibility-trace-like process. Possibly a reinforcement-learning account might also be devised; and perhaps also the two will turn out to be related. In any event, a huge lode of real-time cumulative-record data is ripe for theoretical exploration.

Sutton and Barto's book is possibly too mathematical and AI oriented to be of absorbing interest to *JEAB* readers. On the other hand, the formal elegance of RL should be attractive to people who want to master a powerful theoretical tool. The book may also re-direct the attention of behavior theorists to the possibilities offered by the real-time analysis of behavior dynamics and the rich lode of data available in cumulative and other real-time records.

REFERENCES

3Blue1Brown web tutorial: https://www.youtube.com/watch?v=aircAruvnKk

Ashby, W. R. (1960) *Design for a brain; the origin of adaptive behavior*. New York, Wiley. https://archive.org/details/designforbrainor00ashb

Basu, K. The Traveler's Dilemma. *Scientific American,* May 20, 2007.

de Castro, A. C. Vieira & Machado, A. (2010) Prospective timing in pigeons: isolating temporal perception in the time-left procedure. *Behavioural Processes*, 84(1), 490-499.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York, NY: Basic Books.

Egger, M. D., & Miller, N. E. (1963). When is a reward reinforcing? An experimental study of the information hypothesis. *Journal of Comparative and Physiological Psychology, 56*(1), 132-137. http://dx.doi.org/10.1037/h0040744

Ferster, C. B., & Skinner, B. F. (1957) *Schedules of reinforcement.* New York: Appleton-Century-Crofts. http://www.bfskinner.org/wp-content/uploads/2015/05/Schedules_of_Reinforcement_PDF.pdf

Kahneman, D. *Thinking fast and slow.* New York: Farrar, Strauss and Giroux, 2011.

Klopf, A. H (1982) *The hedonistic neuron: A theory of memory, learning and intelligence* Washington, DC: Hemisphere. For other works by Klopf see https://www.researchgate.net/scientific-contributions/2144940722_A_Harry_Klopf

Ludvig, E. A., Sutton, R. S. & Kehoe, J. (2012) Evaluating the TD model of classical conditioning. *Learning and Behavior,* 40:305–319.

Machado, A. (1997) Learning the temporal dynamics of behavior. *Psychological Review*, 104(2), 241-265.

McDowell, J. (2018) An evolutionary theory of behavior dynamics applied to concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior,* 110:323-335.

Moldakarimov, S. B. & Sejnowski, T. J. (2008) Neural computation theories of learning. https://papers.cnl.salk.edu/PDFs/Neural%20Computation%20Theories%20of%20Learning%202008-4081.pdf

Nabavi, N. (2018) Metacognition in the Data Age. https://medium.com/datadriveninvestor/metacognition-in-the-data-age-89ffdcd92feb

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943) Behavior, purpose, and teleology. *Philosophy of Science,* 10, 18-24.

Staddon, J. E. R. (1981). Cognition in animals: Learning as program assembly. *Cognition*, *10*, 287-294.

Staddon, J. E. R. (2003/10). *Adaptive behavior and learning*. New York: Cambridge University Press. (Internet edition, 2010). http://dukespace.lib.duke.edu/dspace/handle/10161/2878

Staddon, John (2016) Theoretical Behaviorism, Economic Theory, and Choice. In *Economizing mind, 1870-2016: When economics and psychology met…or didn't.* Marina Bianchi & Neil De Marchi (Eds.) Duke University Press, 2016. Pp. 316-331. https://www.researchgate.net/publication/311446403_Theoretical_Behaviorism_Economic_Theory_and_Choice

Staddon, J. E. R., & Higa, J. J. (1999) Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior, 71,* 215-251. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1284701/

Wynne, C. D. L., & Staddon, J. E. R. (1988) Typical delay determines waiting time on periodic-food schedules: static and dynamic tests. *Journal of the Experimental Analysis of Behavior*, *50*, 197-210. http://dukespace.lib.duke.edu/dspace/handle/10161/3387

FIGURES

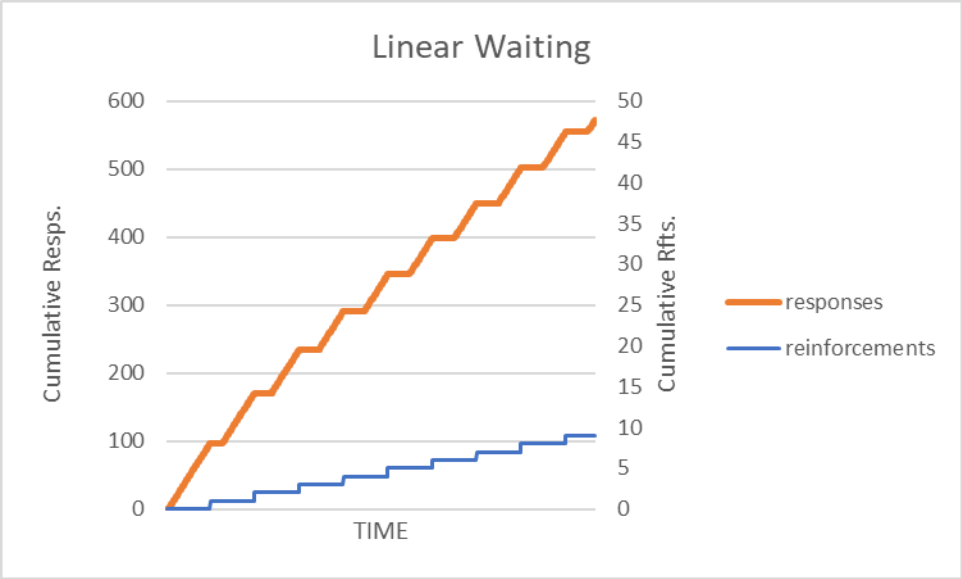

Fig. 1

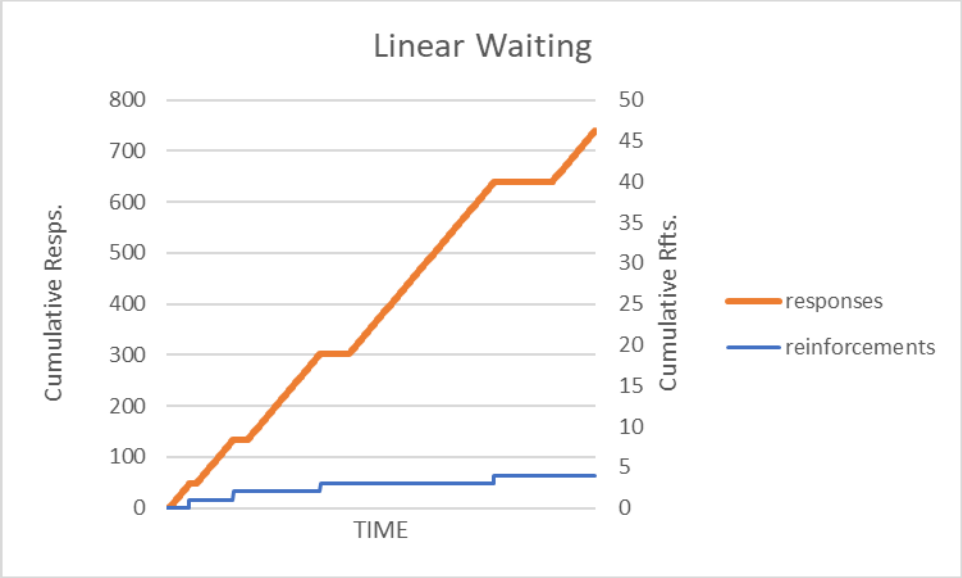

Fig. 2

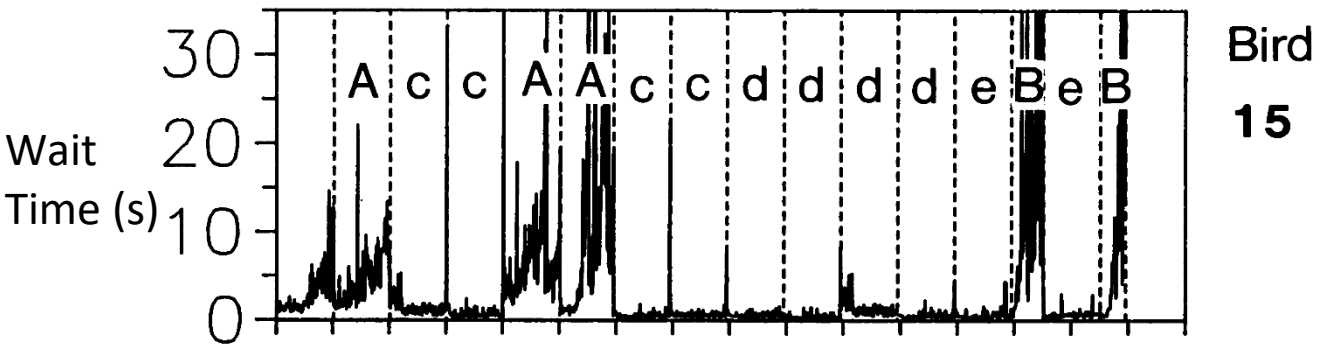

Fig. 3. Wait times, interreinforcement interval by interreinforcement interval in 16 experimental sessions of an autocatalytic schedule. See text for details. (Wynne & Staddon, 1988, Fig. 9, part).

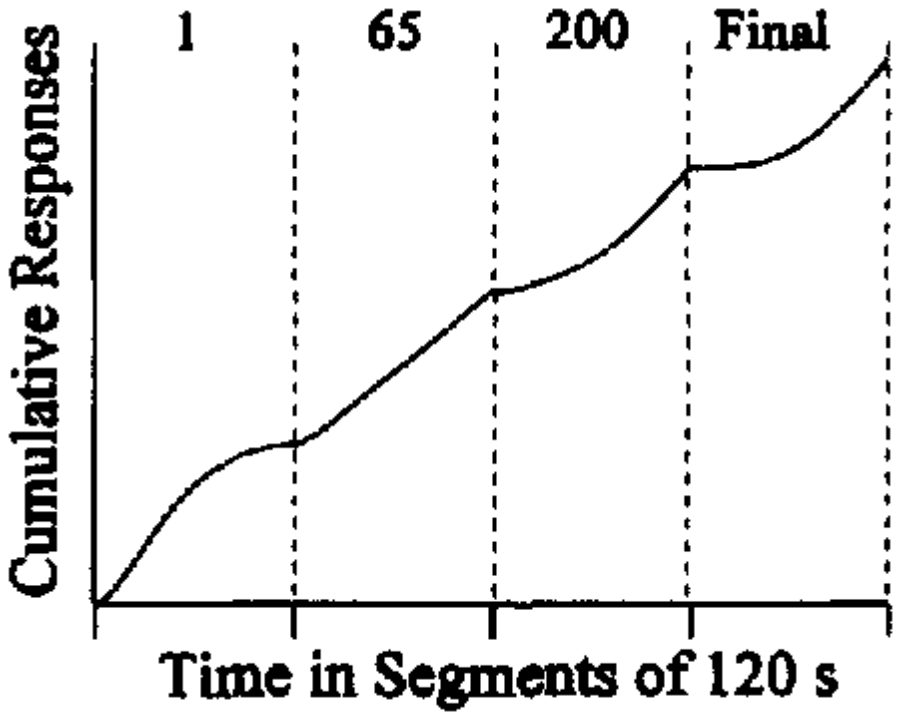

Fig. 4. Machado (1997) Part of Figure 7. Cumulative records of response rate across a 120-s FI schedule during Trials 1, 65, and 200 and at the steady state, according to the BeT model