LEARNING AND ADAPTIVE ECONOMIC BEHAVIOR[†]

Designing Economic Agents that Act Like Human Agents: A Behavioral Approach to Bounded Rationality

By W. BRIAN ARTHUR*

Most economists accept that there are limits to the reasoning abilities of human beings—that human rationality is bounded. The question is how to model economic choices made under these limits. Where, between perfect rationality and its complete absence, are we to set the "dial of rationality," and how do we build this dial setting in to our theoretical models?

One approach to this problem is to lay down axioms or assumptions that suppose limits to economic agents' computational ability or memory, and investigate their consequences. This is useful, but it begs the question of how humans actually behave. A different approach (the one I suggest here) is to develop theoretical economic agents that act and choose in the way *actual* humans do. We could do this by representing agents as using parametrized decision algorithms, and choose and calibrate these algorithms so that the agents' behavior matches real human behavior observed in the same decision context. Theoretical models using these "calibrated agents" would then, we could claim, furnish predictions based on *actual* rather than idealized behavior.

It is unlikely there exists some yet-to-be-defined decision algorithm, some "model of man," that would represent human behavior in all economic problems—an algorithm whose parameters would constitute universal constants of human behavior. Different *contexts* of decision making in the economy call for different actions; and an algorithm calibrated to reproduce human learning in a search problem might differ from one that reproduces strategic-choice behavior. We would likely need a repertoire of calibrated algorithms to cover the various contexts that might arise.

Nevertheless, for a particular context of decision making, calibrating theoretical behavior to match human behavior would allow us to ask questions that are not answerable at present under the assumption of either perfect rationality or idealized learning. We might want to know whether a given neoclassical model with human agents represented by "calibrated agents" will result in some standard asymptotic pattern—a rational-expectations equilibrium, say. We might ask whether agents calibrated to learn as humans do converge to some form of optimality, or interactively to a Nash equilibrium.[1] And we might want to study the speed of adaptation in a particular economic model with human agents represented by calibrated agents.

What would it mean to calibrate an algorithm to "reproduce" human behavior? The object would be algorithmic behavior that reproduces statistically the characteristics of human choice, including the distinctive errors or departures from rationality that hu-

[†] *Discussants*: Ken Binmore, University of Michigan; Drew Fudenberg, MIT; John Geanakoplos, Yale University.

*F.R.I. Stanford University, Stanford, CA 94305-6084, and Santa Fe Institute, 1120 Canyon Road, Santa Fe, NM 87501. I thank Kenneth Arrow, Frank Hahn, John Holland, Richard Herrnstein, and David Lane for stimulating and discussing these ideas, without implicating them in the views expressed here, and Drew Fudenberg and John Geanakoplos for helpful criticism.

[1] Drew Fudenberg and David Kreps (1988) and Paul Milgrom and John Roberts (1991) take a different, but parallel approach. They show that if human learning behavior fulfills certain axioms, asymptotic behavior will result in standard outcomes.

mans make, in the given context. Ideally, the algorithm could pass a Turing test of being indistinguishable from corresponding human behavior in the same context, to an observer who was not informed whether the behavior was algorithm generated or human generated (Alan Turing, 1956). This of course would be asking a lot.

This paper reports on and discusses my recent work (1990) that explores this idea of calibrating an algorithm to reproduce human behavior. It develops and calibrates a learning algorithm for a commonly encountered and simple decision context, that of agents choosing repeatedly among discrete actions with initially unknown, random consequences.

I. A Parametrized Learning Automaton

Consider the problem of iterated choice under uncertainty, in which a decision maker chooses one of N possible actions at each time that have random payoffs or profits drawn from a stationary distribution that is unknown in advance. This would be the case, for example, where a firm, government agency, or research department is faced each period with a choice among N alternative pricing schemes, or policy options, or research projects, each with consequences that are poorly understood at the outset and that vary from "trial" to "trial". The agent chooses one alternative at each time, observes its consequence or payoff, and over time updates his choice as a result. What makes this iterated choice problem interesting is the tension between *exploitation* of high-payoff actions that have been undertaken many times and are therefore well understood, and *exploration* of seldom-tried actions that potentially may have higher average payoff.

The classic multi-arm-bandit version of this problem is to design a learning algorithm or automaton that maximizes some criterion—such as expected average payoff. Our problem is different. It is to design a learning algorithm or learning automaton that can be tuned to choose actions in this iterated choice situation the way humans would.

Consider then a learning automaton that represents a single agent who can undertake one action of N possible actions at each time. We may think of "learning" in this iterated-choice context as updating the probabilities of taking each action on the basis of the payoffs or outcomes experienced. Action i brings reward $\Phi(i)$ that is unknown to the agent in advance, positive, and distributed randomly with a stationary distribution. The automaton (our artificial gent) "learns" via the following simple algorithm. It associates a vector of *strengths*, S_t, with the actions 1 through N, at each time t. The current sum of these strengths is C_t (the component sum of S_t), and the initial strength vector S_0 is strictly positive. The vector p_t represents the agent's probabilities of taking actions 1 through N at time t. At each time t, the agent:

1) Calculates the probability vector as the relative strengths associated with each action. That is, it sets $p_t = S_t / C_t$.

2) Chooses one action from the set according to the probabilities p_t and triggers that action.

3) Observes the payoff received and updates strengths by adding the chosen actions's j's payoff to action j's strength. That is, where action j is chosen, it sets the strengths to $S_t + \beta_t$ where $\beta_t = \Phi(j)e_j$; (e_j is the jth unit vector).

4) Renormalizes the strengths to sum to a value from a prechosen time sequence. In this case, it renormalizes strengths to sum to $C_t = Ct^{\nu}$.

This last step allows us to set the rate and deceleration of the learning via the parameters C and ν that are fixed in advance. The rate of learning, it turns out, is proportional to $1/(Ct^{\nu})$. Parameters C and ν thus define a two-parameter family of algorithms that can be used to calibrate the automaton.

The algorithm has a simple behavioral interpretation (at least when $\nu = 0$). The strength vector summarizes the current confidence the agent or automaton has learned to associate with actions 1 through N. Confidence associated with an action increases according to the (random) payoff it brings in when taken. The automaton chooses its action with probabilities proportional to its current confidence in the N actions, and learning takes place as these probabilities of actions are updated. The summed confidence in all actions is constrained to be constant. S_0, the initial confidence in the actions, represents prior beliefs, possibly carried over from past experience.

It also has a machine-learning interpretation. A Holland-type *classifier* is a condition/action couple ("if object appears in left vision field/turn toward object"), where the action is allowed to be activated only if the condition is fulfilled (John Holland et al., 1987). If several classifiers have the same condition and that condition is fulfilled, they "compete" to be the one activated. Our algorithm can be viewed as a set of N classifiers each competing to be activated, where classifier j is the simple couple "if it is time to act/choose alternative j." As is standard in classifier systems, strengths are associated with the classifiers; one classifier is triggered on the basis of current strengths; and the chosen classifier's strength is updated by the associated reward.

The algorithm is nonlinear in that actions that are frequently taken are further strengthened or reinforced, as in the classic Hebb's rule (Donald Hebb, 1949). And it is stochastic in that actions are triggered randomly on the basis of current probabilities, and rewards are drawn randomly from a distribution. Nonlinearity allows for the exploitation of "useful" actions—ones that pay well tend to be strengthened early on and therefore to be heavily emphasized. And the stochastic property (triggering actions randomly on the basis of their strength) allows for exploration: if a little used action brings in a "jackpot," it may be strengthened sufficiently to become a frequent action.

What can we say about the long-run properties of the learning implicit in this algorithm? Will it "discover" the maximal expected-payoff action, k say, and learn over time to activate it only in the limit? This is not obvious. There are two contradictory tendencies. On the one hand, if an inferior high-payoff action j is triggered early, it may gain in strength and be triggered ever more often until it dominates. Learning may then fall into action j's "gravitational orbit" without escaping. On the other hand, if exploration does not die away too fast, the algorithm will eventually uncover the fact that k is better and home in on it.

In my earlier paper (1990), I show that the algorithm has stochastic dynamics:

$$(1)\quad p_{t+1}(i) = p_t(i) + \alpha_t\left\{ p_t(i)\left[\phi(i) - \sum_j \phi(j) p_t(j)\right] + \xi_t\right\}.$$

The probability of choosing action i grows at a rate proportional to the difference between i's expected payoff $\phi(i)$ and the average payoff at current probabilities, plus an unbiased perturbation term ξ. The step-size α_t is $1/(Ct^{\nu})$. Further analysis settles optimality. If $\nu < 1$, the step-size remains large enough for inferior action j, emphasized early by chance, possibly to build up sufficient strength to shut k out. Optimality is in this case not guaranteed. If, on the other hand, $\nu = 1$, optimality *is* guaranteed. The step-size falls off at rate $1/t$; this delays movement to a nonoptimal action and retains exploration for a long enough time for k to be repeatedly activated and to dominate.

II. Calibration Against Human Subjects

We now want to calibrate the parameters C and ν against data on human learning. Here we are interested in three things: the degree to which the calibrated algorithm represents human behavior; whether the measured value of ν lies within the range that guarantees asymptotically optimal choices; and the general characteristics of learning (such as speed and ability to discriminate) that the calibrated values imply.

To calibrate the algorithm, I use the results of a series of two-choice bandit experiments conducted by Laval Robillard at Harvard in 1952–53 using students as subjects (reported in Robert Bush and Frederick Mosteller, 1955). I would prefer to calibrate on more recent experiments, but these have gone out of fashion among psychologists, and no recent, more definitive results appear to be available. I therefore use Robillard's data as an expedient, interpreting the resulting calibration as a good indication of human behavior in situations of choice rather than a definitive statement.

Robillard set up seven experiments, each with its own payoff structure, and allocated groups of ten subjects to each. Each subject could choose action *A* or *B* repeatedly in 100 trials; and the experiments differed in the probabilities with which unit payoffs occurred. For each experiment, Robillard reported the proportion of *A* choices in each sequential block of 10 trials, averaged over the group of ten subjects (for the data, see my 1990 paper).

I proceed by allowing "groups of artificial agents" (computer runs of the algorithm) to reproduce the equivalent of Robillard's data for fixed values of C and ν. The artificial agents produce *stochastic* sequences of choices or frequencies of choosing action *A*; hence goodness of fit to Robillard's data (under a suitable criterion) for fixed parameters is a random variable. I calibrate the parameters C and ν by minimizing the expected sum of errors squared between the automata-generated frequencies and the corresponding human frequencies for each particular experiment, totaled over the seven experiments. This results in $C = 31.1$ and $\nu = 0.00$. Note immediately that ν lies in a region where asymptotic optimality is far from guaranteed.

Figures 1 and 2 show the artificial agents' learning plotted against the human subjects' in four of the seven experiments, using these calibrated values. (The other experiments are similar in fit.) Judged by eye, the results are encouraging. The automata learn with roughly the same rate and variation as the humans in each of the experiments. Further statistical work (see my 1990 paper) shows that other data sets besides Robillard's produce similar fits, and that six of the seven Robillard learning trajectories could have been produced by the calibrated automata in the sense that each fits well within a distribution of 100 corresponding computed automata trajectories. (The outlier experiment, pictured second in Figure 2, has

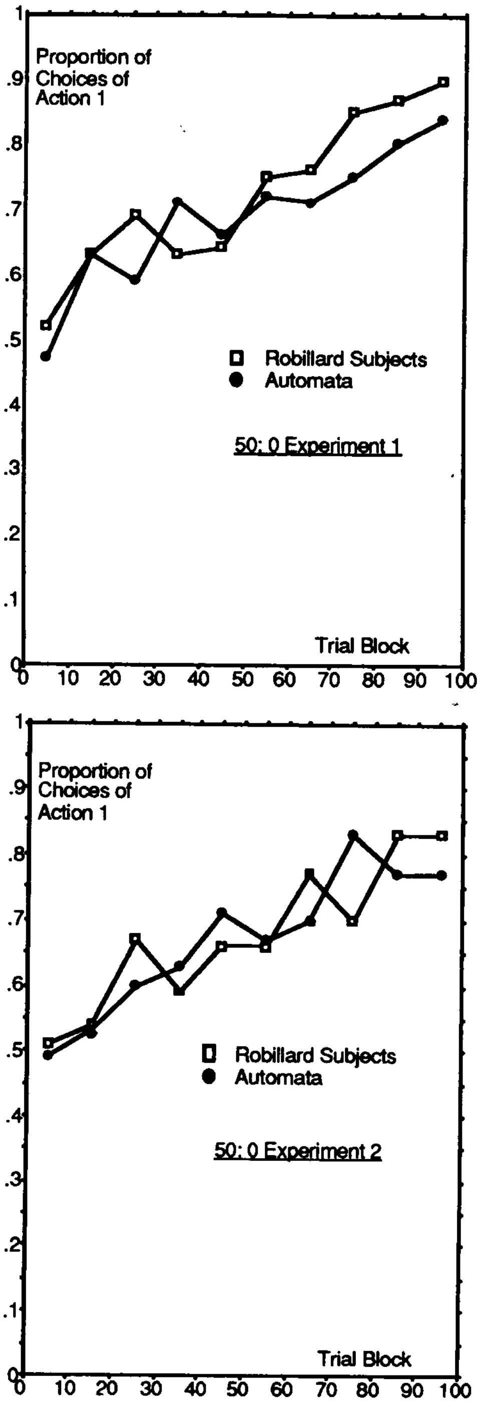

FIGURE 1. CALIBRATED AGENTS' VS. HUMANS' CHOICE FREQUENCIES IN TWO EXPERIMENTS

Note: 50:0 denotes unit payoff with probabilities 50 and 0 percent.

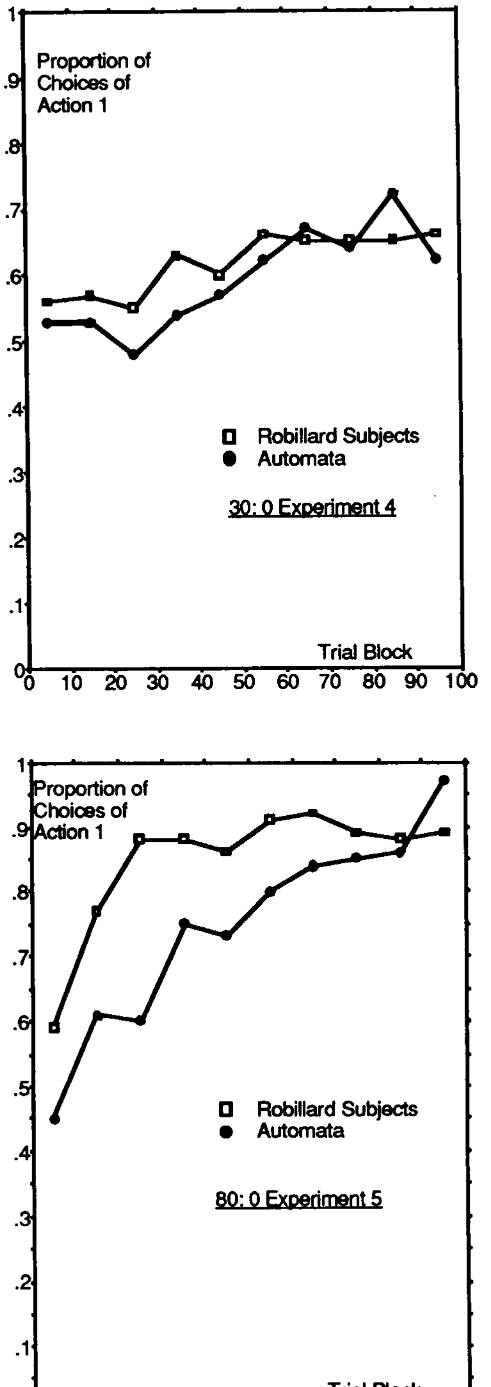

FIGURE 2. CALIBRATED AGENTS VS. HUMANS IN TWO OTHER EXPERIMENTS

close-to-determinate payoffs, something humans notice faster than the algorithm.)

More convincing than statistical tests-of-fit are tests of whether the algorithm can replicate human behavior in quite different choice problems than those for which it was calibrated. A different problem is provided by recent experiments of Richard Herrnstein et al. (1990) where the distribution of payoffs to a choice is no longer fixed, but depends instead on the frequency of actions taken. These are of interest because in this frequency-dependent case, human subjects show a well-documented characteristic behavior called *melioration*; their choices converge not to optimal frequencies that maximize expected payoff, but to quite different frequencies that equalize expected payoffs of each action. In reproducing these experiments, I found that the calibrated agents perform similarly to the Herrnstein subjects; they also meliorate. That our algorithm picks up this characteristic of human behavior is not surprising. Both human and artificial agents carry out local search on the efficacy of choices at the current frequency of choice; thus in this frequency-dependent case both deviate from "rational" behavior in the same way. Findings like this give us confidence that we *can* indeed replicate human behavior for particular decision contexts with calibrated learning agents.

III. What About Optimality and Nash Equilibria?

Let us now put our calibrated agents to work. First, what they can tell us about the prospects for human choices converging to long-run optimality (in our standard, frequency-independent case)? Theoretically, we know from the zero-calibrated value for ν that optimality may not be reached on all occasions. But how often might this happen in practice?

To explore this, I set up a series of computer experiments (see my 1990 paper) designed around an iterated decision problem with six choices, each with uniformly distributed payoff from 0.5 to 1.5 times the choice's expected value. Expected payoffs fell off geometrically from action 1 to 6, with action 1 the maximal choice. I found that as long as the best action was set to be more than 15 percent better than the next best, the algorithm locked in to it close to 100 percent of the time, in 100 repeated experiments. But when the best action's expected payoff was set at less than 10 percent or so more than the next best's, chance variations in received payoff caused choices to lock in to less-than-optimal outcomes in a significant percentage of experiments. Human choice, if captured by the calibration, appears to "discover" and exploit the optimal action with high probability, *as long as it is not difficult to discriminate*. But beyond a perceptual threshold, where differences in alternatives become less pronounced, nonoptimal outcomes become more likely.

It might be objected that this finding is merely an artifact of the algorithm I have chosen. I do not believe so. What is crucial to the emergence of the optimal action is a slowing down in speed of convergence, so that learning has time to "discover" the action with largest expected value. The data, not the algorithm, show this slowing down does not occur. I would thus expect the finding that long-run optimality depends on the difficulty of the problem to be validated under other well-fitting algorithm specifications. We could of course invoke an imagined discount rate to render lock-in to an inferior outcome "optimal over time." But this "discount rate" would then appear to be independent of the time between trials, and I find this argument unpersuasive.

A similar finding carries over to the question of whether human agents are likely to converge to a Nash equilibrium in an iterated game. Think now of our calibrated agents representing human agents learning within a normal-form, stage game (see my 1989 paper). The agents can observe their own actions and random payoffs, but are not particularly well informed of other players' actions and payoff functions. An example might be oligopolistic firms choosing among pricing policies in a decentralized market on the basis of observed end-of-quarter profit. Each agent then faces a multichoice bandit problem as before, and our learning context carries over to this wider problem. Of course, in this case, each agent's payoff distribution changes slowly as other agents change their choice probabilities.

We can represent strategic learning here, for each agent separately, by the calibrated stochastic process in equation (1) and apply asymptotic probabilistic analysis to the resulting stochastic, dynamic model. Our results then represent human behavior in this context to the degree that the calibration captures actual human learning.

For some game-payoff structures, it turns out, strategies may not converge at all. The fact that each agent changes his choice probabilities (strategy profile) as other agents change *theirs* may cause strategy profiles to cycle. In games where learning *does* converge, the analysis shows a Nash outcome is likely but not assured. Nash requires that each agent converge to best reply; but with $\nu = 0$, there may not be sufficient exploration of strategies, and Nash is not guaranteed. In practice, as before, the likelihood of convergence to Nash depends on the difficulty of discrimination among the action payoffs.

How might we use calibrated agents to represent actual human adaptive behavior in other standard neoclassical models? My paper in progress with Holland, Palmer, and Tayler explores convergence to rational expectations equilibrium using calibrated agents in an adaptive version of the Lucas (1978) stock market. We find that the calibrated agents learn to buy and sell stock appropriately, and that the stock price indeed converges to small fluctuations around the rational expectations value. However, we also find that speculative bubbles and crashes occur—a hint that under realistic learning technical analysis may emerge.

IV. Conclusion

I conclude from this exploratory exercise that we can indeed design artificial learning agents and calibrate their "rationality" to replicate human behavior. Not only does the learning behavior of our calibrated agents vary in the way human behavior varies as payoffs change from experiment to experiment in this repeated multichoice context, but it also reproduces two stylized facts of human learning well-known to psychologists: that with frequency-dependent payoffs, humans meliorate rather than optimize; and there is a threshold in discrimination among payoffs below which humans may lock in to suboptimal choices. Most usefully perhaps, the calibrated algorithm has a convenient dynamic representation that can be inserted into theoretical models.

To the degree that the algorithm replicates human behavior, it indicates that human learning most often adapts its way to an optimal steady state or, interactively, to a Nash outcome. But it also shows that humans systematically underexplore less-known alternatives, so that learning may sometimes lock in to an inferior choice when payoffs to choices are closely clustered, random, and difficult to discriminate among. Thus the question of whether human learning adapts its way to standard economic equilibria depends on the perceptual difficulty of the problem itself.

For choices among actions with initially unknown, random payoffs, it appears that behavior does not settle down much before 40 to 100 or more trials. This implies that there is a *characteristic learning time* for human decisions in the economy that depends both on the payoff structure of the problem and on the frequency of observed feedback on actions taken. There is also a time horizon over which the economic environment of a decision problem stays relatively constant. For some parts of the economy, the learning time may be shorter than the problem time horizon. These would be at equilibrium—albeit a slowly changing one. For other parts, learning may take place more slowly than the rate at which the problem shifts. These parts would be always transient, always tracking changes in their decision environment, and never at equilibrium.

REFERENCES

Arthur, W. Brian, "A Learning Algorithm that Mimics Human Learning," Santa Fe Institute Working Paper 90-026, 1990.

______, "Nash-Discovering Automata for Finite-Action Games," mimeo., Santa Fe Institute, 1989.

______ **et al.,** "A Stock Market with Artificially Intelligent Agents," paper in progress, Santa Fe Institute, 1991.

Bush, Robert and Mosteller, Frederick, *Stochastic Models for Learning*, New York: Wiley & Sons, 1955.

Fudenberg, Drew and Kreps, David M., "Learning, Experimentation, and Equilibrium in Games," mimeo., MIT, 1988.

Hebb, Donald O., *The Organization of Behavior*, New York: Wiley & Sons, 1949.

Herrnstein, Richard et al., "Maximization and Melioration," mimeo., Department of Psychology, Harvard University, 1990.

Holland, John H. et al., *Induction: Processes of Inference, Learning, and Discovery*, Cambridge: MIT Press, 1987.

Lucas, Robert E., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, *46*, 1429–45.

Milgrom, Paul and Roberts, John, "Adaptive and Sophisticated Learning in Repeated Normal Form Games," *Games and Economic Behavior*, February 1991, *3*.

Turing, Alan M., "Can a Machine Think?," in John R. Newman, ed., *The World of Mathematics*, Vol. 4, New York: Simon and Schuster, 1956, 2009–2123.