The Case for Strategic Paranoia

An honors thesis for the Department of Economics

Liam F. Clegg

Tufts University, 2011

Contents

Notation

This paper will use a single system of notation for describing mathematical models of learning and decision making. The system below has been chosen to be flexible enough to reproduce all of the various models cited.

$\mathscr{A}$	a set of possible actions or pure strategies for an agent,
A	an arbitrary element of $\mathscr{A}$,
n	the number of elements of $\mathscr{A}$,
θ	a vector in $\mathbb{R}^n$ of the probabilities that an agent will perform each action in $\mathscr{A}$,
$\theta_t(A)$	the probability that an agent will choose action A at time t,
$\mathscr{S}$	a set of possible states of the world,
S	an arbitrary element of $\mathscr{S}$,
e_A	the probability vector which assigns certainty to A,
$\pi(A, \cdot)$	the payoff to an agent of doing A, given some other agent's action or the outcome of some random event,
$u_t(A)$	the expected payoff to an agent of doing A at time t,
ϕ_t	the expected payoff of a strategy θ, equal to $\sum_{A \in \mathscr{A}} \theta_t(A) u_t(A)$,
$L : \theta_t \mapsto \theta_{t+1}$	a discrete time learning rule,
$\dot{\theta}_t$	a continuous time learning rule.

When more than one agent is considered, they are indexed by superscripts. So, for example, the expected payoff to the Row player R of playing T in the first round of the matching pennies game in Figure 1 is given by

$$u_1^R(T) = e_T^R \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \theta_1^C = \theta_1^C(L)\pi^R(A, L) = \theta_1^C(L)2, \tag{1}$$

where vectors are understood to be rows or columns depending on whether they appear on the left or the right of a matrix.

		Column	
		L	R
Row	T	$2, -2$	$0, 0$
	B	$0, 0$	$1, -1$

Figure 1: Normal form of a biased matching pennies game

1 Introduction

Blaise Pascal contrasted the mathematical mind (*esprit de géométrie*) and the strategic mind (*espirit de finesse*). I claim that this distinction is not a real one, and the two minds are in fact one. Thence, behaviors which appear to reflect errors in judging probabilities, or poor decision making in the face of uncertainty, may actually reflect implicit strategic considerations.

This thesis proceeds as follows. Section 2 provides a brief intellectual history of strategy and uncertainty, and introduces some ideas from evolutionary biology. Section 3 presents my main argument. Sections 4, 5, and 6 show three different ways in which I have attempted to model endogenous reinforcement and its effects on learning in strategic and non-strategic situations. Section 7 summarizes my results and suggests directions for future research.

2 Backgound

2.1 Strategy and Competition

2.1.1 Military Men of Old

Perhaps the oldest known work on strategy is Sun Tzŭ's *Art of War*, written in China during the 5th or 6th century B.C. For Sun Tzŭ, 'war' meant not only fighting and killing, but everything a state did in relation to its neighbors. Much of the counsel he presents is so general that it applies to the whole class of non-cooperative games:

- All warfare is based on deception. (1:18)
- If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle. (3:18)

- That general is skillful in attack whose opponent does not know what to defend; and he is skillful in defense whose opponent does not know what to attack. (5:8)

- Though the enemy be stronger in numbers, we may prevent him from fighting. Scheme so as to discover his plans and the likelihood of their success. Rouse him, and learn the principle of his activity or inactivity. Force him to reveal himself, so as to find out his vulnerable spots. (6:22-23)

Superficially, the line quoted from chapter 5 seems remarkably similar to von Neumann's Minimax strategy (see Section 2.1.4). However, the lines from chapters 1, 3, and 6 show that Sun Tzŭ took a naive view of strategic uncertainty which supposes that everything is known by someone. Thence, he advises concealment, deception, prediction, and spying. He does not suggest anything like a mixed strategy, in which one's actions are truly random.[1] Such was the strategy of World War II submarine commanders who used dice as a navigational tool to create unpredictable paths though the water (Miller, 1997).

In his *Discourses on Livy* (1531), Machiavelli touches briefly on strategy in a few places. He has a chapter entitled "Nothing Is More Worthy of a Commander Than to Anticipate the Decisions of the Enemy" (III:18); however, in the chapter itself he offers no examples of commanders who have actually done this. This omission suggests that anticipating an enemy's decisions is more difficult in practice than we might be inclined to think when imagining an ideal commander. Machiavelli later praises the role of deceit in warfare, calling it "a glorious thing" (III:40). Like Sun Tzŭ, he assumes that actions are effectively determined beforehand, and that the keys to success in battle are therefore concealing one's own future actions while trying to learn those of the enemy.

[1] Niou and Ordeshook (1994) suggest an interpretation of the words *ch'i* ("uncommon") and *cheng* ("common") in chapter 5 which allows that Sun Tzŭ had mixed strategies in mind. Given the weakness of the evidence they present relative to their general admiration for Sun Tzŭ's 'anticipation' of twentieth-century advances in game theory, it is difficult to take this suggestion seriously.

2.1.2 Cournot

Augustin Cournot (1838/1897, Ch. 7) offered one of the first formal studies of strategy in economics when he described a duopoly situation in which two producers produce indentical goods, and both produce less than possible to keep the price up. If the two firms colluded, they could cut production even further and make an even greater profit; if there were more firms in the market, total production would be greater and prices and profits therefore lower. The optimal quantities supplied by two producers form a Nash Equilibrium, but such a concept did not exist at the time. Rather, Cournot notes the stability of the resulting equilibrium (under what we would call best-response dynamics), and imagines producers being pushed back to producing equilibrium quantities by myopic profit maximization if they happen to deviate from it (Cournot, 1838/1897, p. 81 and Fig. 2). While Cournot published a treatise on probability in 1843, he does not consider uncertainty anywhere in his chief work of economics.

2.1.3 The Gestation of Game Theory

The formal study of games began in earnest with a series of notes published by Emile Borel between 1921 and 1927.[2] In the first note (Borel, 1921/1953), Borel introduces the concept of a mixed strategy, which he calls a 'rule' for choosing actions. He develops the idea of a rule superior to all others, which will allow a player in a fair game to expect to break even against any opponent, even if the opponent knows his strategy. He (incorrectly) states that such a rule does not exist for all games. Lacking such a rule, he suggests that a player "should so vary his plans that the probabilities attributed by an outside observer to his different manners of playing may never be defined" (p. 100). In his closing paragraph, he speculates that playing such must require "a complete incoherence of mind."

Borel's second note (Borel, 1924/1953) was first published as an appendix to the 1924

[2]The notes were published in French; what follows is based on the translations of the most important ones done by Leonard J. Savage and published in *Econometrica* in 1953.

edition of his book on probability. In this paper, Borel considers the game "known in Japan as the game of paper, scissors, and stone" (p. 102n). He shows that the best strategy is to randomize evenly between the three options, while the best response to any other strategy is a pure strategy. However, he recognizes that playing such a strategy "would induce [the other player] to modify his play" (p. 104), indicating that he has not yet abstracted away from repetition and learning. He then solves for the optimal rule when the payoffs differ depending on the manner of winning, and when one player has an overall advantage. He concludes by repeating his mistaken belief that in games of sufficient complexity, there is no rule which cannot be beaten by an opponent who knows it.

2.1.4 John von Neumann and the Minimax Theorem

In 1926, John von Neumann first presented a proof of his Minimax Theorem. The proof was published a few years later (von Neumann, 1928). The theorem states what Borel had tried to show, but thought impossible: that in any finite game, there is a minimax strategy for both players, such that a player using the strategy cannot be exploited even by an opponent who knows the strategy. In most interesting examples, the minimax strategy is a mixed strategy, such as the strategy of randomizing uniformly between the three actions in rock-paper-scissors. Von Neumann's proof is long and painful, an example of Rózsa Péter's assertion that "von Neumann proved what he wanted to prove" (Leonard, 2010, p. 64).

A decade later, von Neumann and the economist Oskar Morgenstern began work on a volume that would bring together von Neumann's work on game theory along with applications both to games and to social and political problems. The result was the epic *Theory of Games and Economic Behavior* (von Neumann and Morgenstern, 1944). The book contains an axiomatic framework for studying games, a simpler proof of the minimax theorem, and solutions of several games, as well as a theory of cardinal utility for choice under uncertainty.

2.1.5 John Nash and his Equilibrium

While in graduate school at Princeton, John Nash used the recent fixed-point theorems of Brouwer and Kakutani to prove quite concisely that there is at least one 'equilibrium point' in every finite 2-player game (Nash, 1950). Such a point, which became known as a Nash Equilibrium, consists of minimax strategies for both players.

Following Nash's work, game theory became dominated by proofs of the existence of equilibria (Hanappi, 2008).

2.1.6 Evolutionary Game Theory

During the 1970's, several evolutionary biologists began using ideas from game theory to model evolution, based on the idea that some traits evolve because they help animals in competition with other animals. John Maynard Smith provided the first formal synthesis in *Evolution and the Theory of Games* (Maynard Smith, 1982). In the book, he points out that game theory is in some ways better suited to evolutionary biology than to economics, since biological agents maximize fitness, which is measurable, while economic agents maximize utility, which is not. He offers several examples of mixed strategies that arise in nature, such as optimal sex ratios. He also puts forth the concept of an Evolutionary Stable Strategy, an equilibrium concept suited to an evolutionary environment in which species mutate and move location over time. An Evolutionary Stable Strategy is equivalent to a Nash Equilibrium that is isolated and trembling-hand perfect (Bomze, 1986).

The most basic model in evolutionary game theory considers a homogeneous population of agents, each of which is programmed to play a specific strategy. The agents have repeated pairwise interactions which each other, and the strategies reproduce in proportion to their success in these interactions. For a population with n strategies, this leads to a system of $n-1$ differential equations of the form

$$\dot{\theta}_t(A) = \theta_t(A)[u_t(A) - \phi_t], \tag{2}$$

called the replicator dynamic (RD), where θ_t is the vector of population proportions for each strategy, $u_t(A)$ is the fitness of a strategy A, and ϕ_t is the average fitness of the population.

2.1.7 Learning in Games

During the 1980's several game theorists became interested in applying mathematical learning models to agents in games. Drew Fudenberg and David K. Levine, two of the leaders in this project, collect the early findings in their book, *The Theory of Learning in Games* (Fudenberg and Levine, 1998).

Börgers and Sarin (2000)[3] consider an individual decision problem, in which an agent makes repeated choices and earns random payoffs conditional on her actions. They begin with a discrete reinforcement learning model, but then assume that the learning rate is slow enough relative to the rate of play that a law of large numbers applies and the agent's behavior evolves deterministically according to expectations. The resulting learning model in the continuous time limit is given by

$$\dot{\theta}_t(A) = \theta_t(A)[\alpha_t(A) - \alpha_t] + \alpha_t(A)[\sigma_t(A) - \theta_t(A)], \tag{3}$$

where

a_t is an aspiration level,

$\alpha_t(A)$ $= u_t(A) - a_t$ is the expected payoff to action A relative to the aspiration level,

α_t $= \phi_t - a_t$ is the average payoff relative to the aspiration level, and

$\sigma_t(A)$ $= P[\dot{\theta}_t(A) > 0]$ is the total probability of $\theta_t(A)$ increasing.

$\theta_t(A)$ could increase either because the agent selected action A and received a favorable outcome, or because the agent selected some other action and received an unfavorable outcome. The aspiration level in their model changes too, but that piece is less relevant here.

The first summand in the right hand side of Equation 3 is equivalent to Equation 2 below, the continuous time replicator dynamic of Evolutionary Game Theory, although the

[3] A mimeo form of this paper, dated 1995, was widely cited in the late nineties.

interpretation differs somewhat between the two models. Börgers and Sarin demonstrate the different influences of the two summands in Equation 3 by reference to two boundary cases. First, when the aspiration level is less than all of possible payoffs, there can be no unfavorable outcomes. Therefore, $\sigma_t(A) = \theta_t(A)$, and the right-hand side of the equation reduces to the standard replicator equation

$$\dot{\theta}_t(A) = \theta_t(A)[u_t(A) - \phi_t],$$

which converges asymptotically to the optimal action from any interior starting point. Second, when there are only two possible payoffs, and the aspiration level is at the midpoint with distance c from both of them, the first summand is zero. In this case, probability matching is a fixed point, and the system converges to this fixed point whenever $\sigma_t(A)$ converges. While the authors do not mention it, this learning model rationalizes at once two groups of two-armed bandit experiments: those which used no real incentives and found probability matching, and those which did use real incentives and found maximization.

Börgers and Sarin (1997) consider the reinforcement learning model orgininally proposed by Bush and Mosteller (1955) (see Section 2.2.6 below). While Bush and Mosteller present their model as a stochastic learning model for an individual facing an uncertain environment, Börgers and Sarin consider a two-player game. They first show that, in continous-time limit using a law of large numbers, the expected movement of this learning model is equivalent to the continous time replicator dynamic. They then show that this system converges asymptotically. Their proof relies heavily on a result of Norman (1968), which says roughly that any learning model in which the probability updating function (the "event operator") is a contraction mapping on the state space converges.

2.2 Choice under Uncertainty

The mathematical study of probability began in the early seventeenth century with the study of gambling. The probabilities involved in various games were calculated and considered in isolation, without reference to their part in any larger strategy. Early probabilists such as Blaise Pascal thought of chance events as being like the roll of a die, which was often literally true.

The classical economists recognized the role of uncertainty in economics only where it was impossible to miss. Again, they viewed chance events as acts of God or accidents of nature, fundamentally no different from the roll of a die.

2.2.1 Probability: The First Century

In 1654, Blaise Pascal and Pierre de Fermat took up an old problem, known as the unfinished game: suppose two gamblers are playing a game of pure chance with multiple rounds, such as tossing a coin until one player wins three tosses. If they abandon the game before either player has won, what is the fair way to divide the stakes, based on how close each player is to winning? Answering this question requires computing a mathematical expectation, which at the time required the invention of such a concept.

Three years later, Christiaan Huygens published the first book on probability, *Libellus Ratiociniis in Ludo Aleae (The Value of Chances in Games of Fortune)*. In this brief work, he offers a general formula for the expected value of a gamble, a general solution to the unfinished game problem, and solutions to several particular problems involving dice. While the work is pregnant with references to games, adversaries, and the like, he makes no mention whatsoever of strategy. For Huygens, uncertainty is all in the dice.

In a 1713 letter, Nicolas Bernoulli pointed out a problem with evaluating gambles strictly in terms of their expected value, which became known as the St. Petersburg Paradox: Suppose Peter flips a coin until he gets heads. If he gets heads on the first flip, he pays Paul 1 ducat; if on the second flip, 2 ducats; if on the third, 4 ducats; in on the fourth, 8 ducats;

etc. The expected value of this gamble to Paul is infinite[4], but few people would pay more than 2 ducats to be in Paul's position. A solution to this problem was given by Nicolas Bernoulli's cousin, Daniel Bernoulli, in a paper published in 1738.[5] Bernoulli's solution is to evaluate gambles in terms of expected *utility* rather than expected *value*. Utility here means roughly the perceptible benefits conferred on a person by a gain or loss. He assumes that "any increase in wealth, no matter how insignificant, will always result in an increase in utility which is inversely proportionate to the quantity of goods already possessed"[6] (Bernoulli, 1738/1954, p. 25). So, winning twenty dollars yields greater utility to a man who has only ten dollars than to one who has a thousand. Thus began the formal consideration of risky events in relation to an individual's overall situation.

2.2.2 Probability in the Enlightenment

Following the publication of Isaac Newton's *Principia* in 1687, eighteenth century Europe saw an explosion of work in mathematics, science, and philosophy. Pierre Simon, Marquis de LaPlace, made significant contributions in all three fields, and also compiled and consolidated the advances of others. In 1819, he published *A Philosophical Essay on Probabilities*, intended to introduce a lay audience to his and others' work on probability. He credits games of chance with motivating and illustrating many advances in probability (p. 53), but he devotes little effort to the study of games. Rather, he presents probability theory as pure mathematics, and mainly discusses its application to Enlightenment-era questions of knowledge and prediction. LaPlace presents here the Principle of Insufficient Reason: when one knows nothing about the odds involved in a situation, one should assume all outcomes are equally likely.

[4] $\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 4 + \cdots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty$

[5] Bernoulli acknowledges that Gabriel Cramer produced a nearly identical solution a few years before he did.

[6] That is, for all $x > 0$, $u'(x) > 0$ and $u''(x) < 0$.

2.2.3 Uncertainty in Classical Economics

In *Wealth of Nations*, Adam Smith briefly considers the role of uncertainty in a few situations where its importance is undeniable: in the choice of profession by young men[7], and in the purchase of lottery tickets and insurance. Nowhere in his discussions of uncertainty do strategic concerns come into play. Smith does recognize the intrinsic reward of success against long odds in his discussion of professional choice:

> To excel in any profession, in which but few arrive at mediocrity, is the most decisive mark of what is called genius or superior talents. The public admiration which attends upon such distinguished abilities, makes always a part of their reward; a greater or smaller in proportion as it is higher or lower in degree. It makes a considerable part of that reward in the profession of physic; a still greater perhaps in that of law; in poetry and philosophy it makes almost the whole. (Smith, 1776/1937, p. 107)

Like Smith, Alfred Marshall briefly considers the role of uncertainty in his *Principles of Economics*, in his discussions of insurance (p. 398), wages (p. 554), and the returns to capital (p. 613). Based on Bernoulli's idea of diminishing marginal utillity, he points out in passing that gambling, even at fair odds, always entails an expected loss (p. 135n.; cf. also LaPlace, 1819/1951, p. 24). However, following his proof of this statement in the mathematical appendix, he concedes that this does not necessarily make gambling unattractive:

> It is true that this loss of probable happiness need not be greater than the pleasure derived from the excitement of gambling, and we are then thrown back upon the induction that pleasures of gambling are in Bentham's phrase "impure"; since experience shows that they are likely to engender a restless, feverish character, unsuited for steady work as well as for the higher and more solid pleasures of life. (Marshall, 1890/1920, p. 843)

[7]Few women faced such a choice in Smith's day

2.2.4 The Savage Axioms

Leonard J. Savage (1954) laid out a collection of "Postulates of a Personal Theory of Decision" in his *Foundations of Statistics*. In spite of its title, the book is really a series of rules for making choices under uncertainty to optimally satisfy a consistent set of preferences. Savage presents the theory as being applicable to decision making in day-to-day life, and as a source of guidance for decision makers striving to behave more rationally.

Savage's theory considers of possible states, their subjective probabilities, acts, and consequences. Agents choose acts based on their preferences for consequences, which are functions of acts and states.[8]

Some of the more controversial of Savage's Axioms are the following:

(P1) Well-ordering: There exists a weak ordering $\geq$ over all acts.

(P2) Independence: Preference between acts should be independent of support which the acts share.

2.2.5 Allais, Ellsberg, Kahneman & Tversky

At a conference in Paris in 1952, Maurice Allais famously performed an impromtu experiment on the assembled experts in choice under uncertainty which revealed a regularity of preferences that violated the Independence Axiom.[9] Based on this, and several other empirical observations, he rejects what he calls the neo-Bernoullian formulation which ranks acts by some convex combination of the possible consequences of each act. Instead, he proposes a more general formulation in which utility is an arbitrary functional on the probability distribution of neo-Bernoullian expected utilities. (Allais, 1953)

[8]Since my aim here is to present the philosophical assumptions implicit in Savage's theory, I omit his topological notation and present the axioms only in rough summary form.

[9]The experiment asked participants to state their preference in each of two pairs of lotteries. The first was between I. a 100% chance of \$500,000, or II. a 10% chance of \$2,500,000, an 89% chance of \$500,000, and a 1% chance of 0. The second was between III. an 11% chance of \$500,000 and an 89% chance of 0, or IV. a 10% chance of \$2,500,000 and a 90% chance of 0. Most people prefer I to II, and IV to III. However, the two pairs are identical in the parts of their support that differ; III and IV are constructed by replacing an 89% chance of \$500,000 with an 89% chance of 0 in I and II, respectively.

Allais is much more famous for the paradox revealed by his experiment than for any alternative he proposed to neo-Bernoullian approaches. While he does present several compelling criticisms of these approaches, he does not stray from considering choice undertainty in real life as if it were a choice between lotteries.

Daniel Ellsberg introduced a similar experiment which likewise generated systematic violations of Savage's Independence and Well-ordering Axioms. His chief finding was that most people prefer gambles with known probabilities to those with unknown probabilities, in ways that cannot be rectified by any subjective probabilities. He argued that people are naturally averse to ambiguity, and proposed as a utility function for uncertain gambles a weighted average of the expected payoff and minimum payoff, with the weight of the worst-case proportional to the ambiguity of the situation (Ellsberg, 1961). K
"uhberger and Perner (2003) suggests that paranoia may play a role in the Ellsberg Paradox.

While Savage, Allais, Ellsberg, and others relied on informal experiments and thought experiments, beginning in the 1970s the Israeli psychologists Daniel Kahneman and Amos Tversky conducted a series of controlled experiments in which they solicited preferences over lotteries from hundreds of subjects (mostly Israeli undergraduates). Based on their findings in numerous survey experiments, they constructed an expected utility function for lotteries which consisted of both a value function $v(x)$ for outcomes and a probability weighting function $\pi(p)$ for probabilities. The value function is characterized by risk-aversion over gains ($v'' < 0$ for $x > 0$), risk-seeking over losses ($v'' > 0$ for $x < 0$), and loss aversion ($|v(-x)| > v(x)$ for $x > 0$). The probability weighting function is inverse-sigmoidal, so that it overweights very small probabilities and underweights very large ones. They termed this model Prospect Theory (Kahneman and Tversky, 1979).

Prospect Theory is revolutionary in two ways. First, it is purely descriptive, and is offered without any normative implications. Second, it is based on controlled experiments with naive subjects, rather than casual observation or thought experiments. Kahneman and Tversky made scant philosophical or mathematical innovation in modeling choice under uncertainty,

but they revolutionized the motivations and processes for building and fitting models.

2.2.6 Probabilistic Learning Theories

At the same time that Savage, Samuelson, and others were debating the optimal way to approach uncertainty, many experimental psychologists were conducting experiments with animals and humans to find out how these subjects dealt with uncertainty. Following on the success of experiments with animals (Thorndike, 1905), these experiments generally provided subjects with very little information, but gave them the chance to learn the probabilities associated with events through repeated experience. Psychologists then constructed mathematical models of learning and choice under uncertainty to fit the data they collected in their experiments.

Bush and Mosteller (1955) develop one such mathematical model of learning. Their formulation relies heavily on operators which are defined in terms of scalar constants, of the general form

$$\theta_{t+1} = Q\theta_t,$$

where Q is a stochastic matrix (i.e., a square matrix whose columns sum to unity). Thus, for $\mathscr{A} = \{A_1, A_2\}$, the updating rule associated with A_1 can be written

$$\begin{aligned} Q_1\theta(A_1) &= (1-b_1)\theta(A_1) + a_1(1-\theta(A_1)) && \text{for some scalars } a_1, b_1 \\ &= a_1 + \alpha_1\theta(A_1) && \text{where } \alpha_1 = 1 - a_1 - b_1 \\ &= \alpha_1\theta(A_1) + (1-\alpha_1)\lambda_1 && \text{where } \lambda_1 = a_1/(a_1+b_1) \\ &= \lambda_1 + \alpha_1(\lambda_1 - \theta(A_1)). \end{aligned} \tag{4}$$

When one such operator is applied repeatedly to θ, we see that

$$\lim_{t\to\infty} \theta_t = \begin{pmatrix} \lambda_1 \\ \\ \lambda_2 \end{pmatrix}.$$

That is, θ converges exponentially to a fixed point. Many psychologists at the time thought that this was a good approximation of the behavior of human and animal subjects in various choice experiments. Indeed, the model seems to have been chosen for its ability to model this phenomenon. However, as some game theorists complained at the time, the model can converge to a sub-optimal strategy, namely if it converges to an interior fixed point when there is a strictly dominant pure strategy. While it could be extended by allowing the operator to vary, this is not the authors' intent; they describe at length procedures for estimating a and b or α and λ from an experimental dataset. Performing such estimations assumes that the operator is fixed.

Cross (1973) applies the Bush and Mosteller model to firms, and adds a random state variable which influences payoff magnitude. This leads to the following updating equation when action A is taken at time t, and state S is realized:

$$\theta_{t+1}(A) = \theta_t(A) + \pi(A, S)(1 - \theta_t(A)), \tag{5}$$

where the π function is restricted to $[0, 1]$. This is equivalent to the first form of Equation 4, with $b = 0$ and the parameter a replaced with the payoff function.

Sutton and Barton (1998) extend the intuition of reinforcement learning to a model that includes action selection. Their resulting model, temporal difference (TD) learning, combines the exploration of Monte Carlo methods with the bootstrapping approach of dynamic programming. The TD agent maintains a policy θ for selecting actions, and a value function V mapping current states to future reward streams:

$$\begin{aligned} V_t(S|\theta) &= \sum_{k=0}^{\infty} \gamma^k \pi_{t+k+1}(A, S) \\ &= \pi_{t+1}(A, S) + V_{t+1}(S|\theta), \end{aligned} \tag{6}$$

where the action A taken in each future period is chosen according to θ, and $\gamma \in (0, 1]$ is a discount rate. When the agent takes an action, and learns the payoff $\pi_{t+1}(A, S)$, an error

signal δ_t is generated, given by

$$\delta_t = \pi_{t+1}(A, S) + V_{t+1}(S|\theta) - V_t(S|\theta). \tag{7}$$

This error term is then the input to a policy updating function, the specifics of which I omit here, such that

$$\theta_{t+1}(A) > \theta_t(A) \text{ iff } \delta_t > 0.$$

2.3 Evolutionary Fundamentals

Evolutionary theorizing requires a "strange inversion of reason" which I will indulge here. Rather than asking why people behave suboptimally (or 'irrationally'), I assume that people must be acting optimally, or almost optimally, and that systematic evidence to the contrary casts doubt upon the definition of optimality in question. This is in keeping with Orgel's Second Rule: *Evolution is cleverer than you are* (Dennett, 1984).

Given that the world is full of hungry predators, and that many of these predators are quite clever, it is not a bad idea to be a bit paranoid. The cost of mistaking a gust of wind for a sabertooth tiger is a few wasted calories; the cost making the opposite mistake is being converted to calories. Nature being thus, it is not surprising that humans have evolved "Hyperactive Agent Detection Devices" (Barrett, 2000).

3 Endogenous Reinforcement for Protean Behavior

Protean Behavior (Chance and Russell, 1959) is a biologist's term for acting randomly to avoid having one's actions predicted and exploited by a predator. For example, an ostrich grazing in the savannah in Kenya must occasionally pull its head out of the grass and look around to see if any lions are coming (Bertram, 1980). The timing of these scans is nearly a Poisson process, which once seemed sub-optimal to many biologists. However, lions do not attack at random, but hide in the bush and wait for the ideal opportunity to strike. Scannell

et al. (2001) show that the protean timing of scans is therefore optimal, in that it does not create any points in time at which the lion knows that the ostrich will not be looking up soon. Miller (1997) offers several instances in which protean behavior might be observed in primates, including humans, but there has been little evidence collected on this.

It is obvious from introspection that the satisfaction of success in the face of uncertainty is greater the less one expects to succeed beforehand. More recently, the finding by neuroscientists that dopamine encodes *reward prediction error* (Schultz et al., 1997) offers a mechanism for the common experience that winning at long odds is more fun. Dopamine RPE can facilitate reward learning by facilitating long-term potentiation in the brain, and this is the hypothesized role of dopamine in current neurobiological accounts of reward learning (Montague et al., 1996; Hazy et al., 2010). Such accounts make no use, however, of the phenomenology to individuals of the thrill of winning.

Many have suggested that this thrill of winning is adaptive in that it encourages individuals to explore new potential sources of food and other goods. This may be the case, but I do not consider this argument here. Furthermore, there are well-documented costs to this aspect of dopamine. In extreme cases, people can become hooked on the rush of unexpected gains, leading to compulsive gambling (Ross, 2009). For a compulsive gambler, it is not the winnings which are exciting, but rather the thrill of winning itself.

3.1 Main Argument

I argue that the thrill of winning, which I call Endogenous Reinforcement, is adaptive in that it prevents agents from adopting pure strategies in competitive environments, which would leave them open to exploitation. Thus, Endogenous Reinforcement can facilitate Protean Behavior. This means that humans' learning and decision making apparatus is implicitly strategic, in a way that may generate sub-optimal behavior in the fact of pure randomness. What appears to be bounded rationality may therefore be strategic paranoia.

The next three sections show three very different attempts I have made at modeling

endogenous reinforcement, and its ability to help agents play mixed strategies, or at least avoid pure strategies.

4 Stability of Replicator Dynamics in Rock-Paper-Scissors

4.1 Previous Work

Consider a population of agents playing the Rock-Paper-Scissors game depicted in Figure 2. Let $\phi_t(A)$ be the number of players playing pure strategy A at time t, or the sum of the probabilities associated with A over all agents playing mixed strategies with A in their support. Let $\theta_t(A) = \phi_t(A)/\sum_{A'} \phi_t(A')$ be the fraction of the population this number represents. I will be concerned here only with fractions of the population, and will not model the growth or shrinkage of the population as a whole. Let θ_t be the vector of population fractions. Then the expected payoff of playing strategy A at time t is

$$u_t(A) \equiv \sum_{A'} \theta_t(A') u(A, A'),$$

and the average expected payoff in the population is

$$\bar{u}_t = \sum_{S} \theta_t(A) u_t(A).$$

The replicator dynamic is based on the idea that each strategy is the subject of evolution, with each pure strategy "reproducing" in proportion to its success in each round. I model this as a continuous dynamical system,

$$\dot{\phi}_t(A) = \phi_t(A) u_t(A), \tag{8}$$

yielding

$$\dot{\theta}_t(A) = \frac{\dot{\phi}_t(A) \sum_{A'} \phi_t(A') - \phi_t(A) \sum_{A'} \dot{\phi}_t(A')}{(\sum_{A'} \phi_t(A'))^2} = \theta_t(A)[u_t(A) - \bar{u}_t]. \tag{9}$$

	R	P	S
R	$0,0$	$-1,1$	$1,-1$
P	$1,-1$	$0,0$	$-1,1$
S	$-1,1$	$1,-1$	$0,0$

Figure 2: Rock-Paper-Scissors

Now, given a vector θ_t of population proportions of strategies at date t, the expected payoffs of the three possible actions are given by

$$\begin{aligned} u_t(R) &= \theta_t(S) - \theta_t(P) \\ u_t(P) &= \theta_t(R) - \theta_t(S) \\ u_t(S) &= \theta_t(P) - \theta_t(R) \end{aligned}$$

Using the fact that $\theta_t(S) = 1 - \theta_t(R) - \theta_t(P)$, and the fact that $\bar{u}_t = 0$ for all t, I can write the replicator dynamics as a two-dimensional dynamical system:

$$\begin{aligned} \dot{\theta}_t(R) &= \theta_t(R)[1 - \theta_t(R) - 2\theta_t(P)] \\ \dot{\theta}_t(P) &= \theta_t(P)[-1 + 2\theta_t(R) + \theta_t(P)] \end{aligned}$$

This system has one fixed point, at $(\frac{1}{3}, \frac{1}{3})$, corresponding to the Nash equilibrium in mixed strategies of the game. Linearizing about the equilibrium yields the Jacobian

$$\begin{bmatrix} -\frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

This matrix has eigenvalues $\lambda = \pm\frac{1}{3}\mathbf{i}$ with zero real part, so the equilibrium is not asymptotically stable and is surrounded by cycles. (Fudenberg and Levine, 1998, Ch. 3)

4.2 Results of Adding Endogenous Reinforcement

Suppose we make the following change to this system. While the payoff matrix remains the same, each agent in each round earns an effective reward which is the sum of the exogenous reward, dictated by the payoff matrix and the result of each stage game, and an endogenous reinforcement, dependent on the results of the stage game given her perceived chances of winning with the action she plays. I add two parameters: δ_1, representing the thrill of winning, and δ_2, representing the agony of defeat. The expected payoffs of the three strategies are now as follows:

$$\begin{aligned} u_t(R) &= \theta_t(S)(1+\delta_1(1-\theta_t(S))) - \theta_t(P)(1+\delta_2(1-\theta_t(P))) \\ u_t(P) &= \theta_t(R)(1+\delta_1(1-\theta_t(R))) - \theta_t(S)(1+\delta_2(1-\theta_t(S))) \\ u_t(S) &= \theta_t(P)(1+\delta_1(1-\theta_t(P))) - \theta_t(R)(1+\delta_2(1-\theta_t(R))), \end{aligned}$$

while $\bar{u}_t = \sum_A \theta_t(A)u_t(A)$. This will not normally equal zero. However, the endogenous payoffs I have introduced do not change the nature of the game, but only the way in which agents choose strategies.

Again, using the fact that $\theta_t(S) = 1 - \theta_t(R) - \theta_t(P)$, we get

$$\begin{aligned} u_t(R) &= 1 - \theta_t(R) - 2\theta_t(P) + \delta_1\theta_t(R) + \delta_1\theta_t(P) - \delta_2\theta_t(P) \\ &\quad - \delta_1\theta_t(R)^2 - 2\delta_1\theta_t(R)\theta_t(P) - \delta_1\theta_t(P)^2 + \delta_2\theta_t(P)^2 \\ u_t(P) &= -1 + 2\theta_t(R) + \theta_t(P) + \delta_1\theta_t(R) + \delta_2\theta_t(R) - \delta_2\theta_t(P) \\ &\quad - \delta_1\theta_t(R)^2 + \delta_2\theta_t(R)^2 + 2\delta_2\theta_t(R)\theta_t(P) + \delta_2\theta_t(P)^2 \\ \bar{u}_t &= -\delta_2\theta_t(R) + \delta_1\theta_t(P) + \delta_1\theta_t(R)^2 + 2\delta_2\theta_t(R)^2 + \delta_1\theta_t(R)\theta_t(P) - \delta_2\theta_t(R)\theta_t(P) - 2\delta_1\theta_t(P)^2 \\ &\quad - \delta_2\theta_t(P)^2 - \delta_1\theta_t(R)^3 - \delta_2\theta_t(R)^3 - 3\delta_1\theta_t(R)^2\theta_t(P) + 3\delta_2\theta_t(R)\theta_t(P)^2 + \delta_1\theta_t(P)^3 + \delta_2\theta_t(P)^3. \end{aligned} \tag{10}$$

Now, the dynamical system is characterized by

$$\begin{aligned}
\dot{\theta}_t(R) =&\theta_t(R) - \theta_t(R)^2 - 2\theta_t(R)\theta_t(P) + \delta_1\theta_t(R)^2 + \delta_2\theta_t(R)^2 - \delta_2\theta_t(R)\theta_t(P) - 2\delta_1\theta_t(R)^3 \\
&- 2\delta_2\theta_t(R)^3 - 3\delta_1\theta_t(R)^2\theta_t(P) + \delta_2\theta_t(R)^2\theta_t(P) + \delta_1\theta_t(R)\theta_t(P)^2 + 2\delta_2\theta_t(R)\theta_t(P)^2 + \delta_1\theta_t(R)^4 \\
&+ \delta_2\theta_t(R)^4 + 3\delta_1\theta_t(R)^3\theta_t(P) - 3\delta_2\theta_t(R)^2\theta_t(P)^2 - \delta_1\theta_t(R)\theta_t(P)^3 - \delta_2\theta_t(R)\theta_t(P)^3 \\
\dot{\theta}_t(P) =& - \theta_t(P) + 2\theta_t(R)\theta_t(P) + \theta_t(P)^2 + \delta_1\theta_t(R)\theta_t(P) - \delta_1\theta_t(P)^2 - \delta_2\theta_t(P)^2 - 2\delta_1\theta_t(R)^2\theta_t(P) \\
&- \delta_2\theta_t(R)^2\theta_t(P) - \delta_1\theta_t(R)\theta_t(P)^2 + 3\delta_2\theta_t(R)\theta_t(P)^2 + 2\delta_1\theta_t(P)^3 + 2\delta_2\theta_t(P)^3 + \delta_1\theta_t(R)^3\theta_t(P) \\
&+ \delta_2\theta_t(R)^3\theta_t(P) + 3\delta_1\theta_t(R)^2\theta_t(P)^2 - 3\delta_2\theta_t(R)\theta_t(P)^3 - \delta_1\theta_t(P)^4 - \delta_2\theta_t(P)^4.
\end{aligned} \tag{11}$$

Observe[10] that $\left(\frac{1}{3}, \frac{1}{3}\right)$ is still a fixed point of the system. Furthermore, linearizing about this equilibrium now yields the Jacobian

$$\begin{bmatrix} -\frac{1}{3} - \frac{1}{9}\delta_1 & -\frac{2}{3} - \frac{1}{9}\delta_1 - \frac{1}{9}\delta_2 \\ \frac{2}{3} + \frac{1}{9}\delta_1 + \frac{1}{9}\delta_2 & \frac{1}{3} + \frac{1}{9}\delta_2 \end{bmatrix}.$$

This matrix has eigenvalues $\lambda = -\frac{1}{18}\delta_1 + \frac{1}{18}\delta_2 \pm \left(\frac{1}{\sqrt{3}} + \frac{1}{6\sqrt{3}}\delta_1 + \frac{1}{6\sqrt{3}}\delta_2\right)\mathbf{i}$. Observe that $Re(\lambda) < 0$ iff $\delta_1 > \delta_2$. Thus, the equilibrium is stable and attracting so long as the joy of winning is greater than the agony of defeat. □

4.3 A shorter path to the same result

A more general result, given in Nowak (2006), is that the replicator dynamics for a game with a payoff matrix of the form

	R	P	S
R	0	$-Y$	X
P	X	0	$-Y$
S	$-Y$	X	0

[10] I.e. trust me

is asymptotically stable iff the determinant of the payoff matrix is positive. As

$$\det \begin{vmatrix} 0 & -Y & X \\ X & 0 & -Y \\ -Y & X & 0 \end{vmatrix} = X^3 - Y^3, \tag{12}$$

this happens iff $X > Y$.

In my model, then, the δ_1 and δ_2 terms change the relevant determinant to

$$\begin{aligned} &\det \begin{vmatrix} 0 & -1-\delta_2(1-\theta_t(P)) & 1+\delta_1(1-\theta_t(S)) \\ 1+\delta_1(1-\theta_t(R)) & 0 & -1-\delta_2(1-\theta_t(S)) \\ -1-\delta_2(1-\theta_t(R)) & 1+\delta_1(1-\theta_t(P)) & 0 \end{vmatrix} \\ &= -(1+\delta_2(1-\theta_t(P)))(1+\delta_2(1-\theta_t(S)))(1+\delta_2(1-\theta_t(R))) \\ &\quad + (1+\delta_1(1-\theta_t(S)))(1+\delta_1(1-\theta_t(R)))(1+\delta_1(1-\theta_t(P))) \\ &> 0 \text{ iff } \delta_1 > \delta_2. \end{aligned} \tag{13}$$

4.4 On the Benefits of Convergence

One may reasonably ask what good a stable equilibrium does anybody in this game. The answer depends on a particular interpretation of mixed strategies and the replicator dynamic. Assume that there is a large populations of agents, each of whom plays a mixed strategy. Agents match randomly and compete against each other for one round only, and all agents observe the outcomes of all interactions. Then, as the learning rule is deterministic and depends only on the behavior of the whole population, all agents will play the same mixed strategy at all times.

The result of this interpretation is that payoff variance is experienced at the level of individuals. This is the level at which strategies are implemented, and is thus the level at

which strategic mechanisms are selected for. Using the fact that $\bar{u}_t = 0 \quad \forall t$,

$$\begin{aligned} Var_t[\theta] &= \sum_A \sum_{A'} \theta_t(A)\theta_t(A')(\pi(A, A'))^2 \\ &= 2\theta_t(R)\theta_t(S) + 2\theta_t(P)\theta_t(R) + 2\theta_t(S)\theta_t(P) \end{aligned} \tag{14}$$

Now, using the fact that expected payoff is constant at zero, maximizing a concave mean-variance utility function is equivalent to

$$\min_{\theta \in \Delta^3} Var[\theta], \tag{15}$$

yielding the LaGrangian

$$\mathscr{L} = \theta(R)\theta(S) + \theta(S)\theta(P) + \theta(P)\theta(R) - \lambda \left(\sum_A \theta(A) - 1 \right). \tag{16}$$

Thence,

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial \theta(R)} = 0 \qquad &\implies \theta(S) + \theta(P) = \lambda \\ \frac{\partial \mathscr{L}}{\partial \theta(P)} = 0 \qquad &\implies \theta(R) + \theta(S) = \lambda \\ \frac{\partial \mathscr{L}}{\partial \theta(S)} = 0 \qquad &\implies \theta(P) + \theta(R) = \lambda \\ \frac{\partial \mathscr{L}}{\partial \lambda} = 0 \quad &\implies \theta(R) + \theta(P) + \theta(S) = 1 \end{aligned} \tag{17}$$

Combining these four equalities yields

$$\theta(R) = \theta(P) = \theta(S) = \frac{1}{3}, \tag{18}$$

and thence

$$Var[\theta^*] = Var\left[\begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \right] = \frac{2}{9} + \frac{2}{9} + \frac{2}{9} = \frac{2}{3}. \tag{19}$$

Additionally, variance is zero for all pure strategies, since in this case the result of every

interaction is a tie.

Now, as pure strategies are subject to exploitation and thence negative expectations at the level of individual interacions, the equilibrium mixed strategy is the unique optimum for agents with long-term concave utility functions. Given a sufficiently long time-frame or a sufficiently broad context, we may reasonably assume that all agents are risk averse, and therefore have concave utility functions. One possible interpretation is that short-term utility corresponds to hedonic utility, while long-term utility corresponds to evolutionary fitness.

The interesting result here, then, is that long-term risk-aversion is implemented by a mechanism which encourages risk-seeking in the short term. Endogenous reinforcement raises the expected utility of risky prospects in proportion to their riskiness, creating risk-seeking behavior at the level of individual interactions. However, such behavior leads agents to converge to optimal mixed strategies. With all agents playing the optimal mixed strategy, this strategy is asymptotically stable, so that payoff variance is minimized.

5 Probability Matching as a Minimax Strategy

5.1 Introduction

Probability matching (PM) was widely documented by experimental psychologists of the mid-twentieth century. For example, in a 'two-armed bandit experiment' in which subjects were asked to predict which of two lights would come on in many repeated trials, if the light on the left actually came on in 75% of trials, subjects came to predict 'left' 75% of the time (Grant et al., 1951). Economists and game theorists saw PM as irrational, since subjects would have been correct more often if they had selected 'left' (in this example) every time after learning the probabilities (e.g. Arrow, 1958). This debate was largely resolved when experimenters figured out how to get subjects to maximize, selecting the superior option on all later trials, as the theorists thought they should (Edwards, 1956). Interest in PM

has re-arisen in recent years (Vulkan, 2000), and yet more experiments have documented contexts in which it does and does not occur (Shanks et al., 2002). However, no one has yet produced a satisfactory explanation of why PM happens in the first place.

Many explanations of PM have focused on the fact that PM is the only way that a subject in a two-armed bandit experiment can possibly get 100% correct (e.g. Goodnow, 1955). This fact plays no role in my explanation. Rather, I argue that subjects continue to choose the inferior option sometimes because they fear that if they choose the superior option every time, it will somehow 'catch on' and cease to be superior.

I define *experiment* to mean an interaction between a single individual and a source of pure randomness, in which the individual's level of success is determined jointly by her actions and the outcome of random events. I use *game* in the usual sense of an interaction between multiple individuals each of whose success is determined jointly by her own actions and those of the others. I assume that games are noncooperative.

In reviewing the literature, I have been unable to locate a single two-armed bandit study which questions the assumption that subjects know they are in an experiment. This is particulary surprising in the case of a study by Suppes and Atkinson (1960), in which pairs of subjects were told they were each in an independent experiment but were actually in a game against each other.[11]

This paper offers an explanation of PM as follows. First, I suggest that it is likely that participants in the studies in question act as if they are in a game rather than an experiment. Next, I show that PM is an optimal strategy in a particular game, and therefore not an unreasonable behavior. Then, I show how a subject in an experiment who thinks she is in a game can fail to learn that it is in fact an experiment, since the expected feedback in the game and the experiment are equivalent. Finally, I suggest that PM, and perhaps many other empirical phenomena uncovered by decision making experiments, are not cases

[11] "The participants were read the following instructions: 'We always run subjects in pairs because this is the way the equipment has been designed and also because it is the most economical procedure. Actually, however, you are working on two completely different and independent problems."' (Suppes and Atkinson, 1960, p. 81)

of bounded rationality, but rather of strategic paranoia.

My argument is teleological and theoretical, and thus does not necessarily conflict with any empirical or mechanistic account of PM. For instance, the combination of experience, feedback, and incentives (Edwards, 1956) could be sufficient for participants to learn they are in an experiment, rather than a game, and to give up PM in favor of maximizing. Similarly, the generation of random responses could be implemented by a complex, bilateral process in the brain (Wolford et al., 2000; Miller et al., 2005) which requires significant glucose (McMahon and Scheel, 2010) and is experienced consciously as the generation and testing of hypotheses. Whether my argument changes the plausibility rankings of such accounts is a question for another paper.

5.2 Probability Matching is an Optimal Strategy in Certain Games

		Column	
		L	R
Row	T	$X, -X$	$0, 0$
	B	$0, 0$	$Y, -Y$

Figure 3: Biased matching pennies, $X > 0, Y > 0$

Consider the 'matching pennies' game in Figure 3. A strategy θ^R for Row is fully defined by $p = \theta^R(T)$; for the sake of simplicity I will refer here to p and $q = \theta^C(L)$ rather than θ^R and θ^C.

For any pair of strategies (p, q), the expected payoff to Row is given by

$$\pi^R(p, q) = pqX + (1 - p)(1 - q)Y \tag{20}$$

Under complete information, both players' minimax strategies are given by

$$p^* = q^* = \frac{Y}{X + Y}. \tag{21}$$

When both players play their minimax strategies, the expected payoff to Row is

$$\pi^* := \pi^R(p^*, q^*) = \frac{XY}{X+Y} \tag{22}$$

Consider the following twist. Suppose Column knows the payoffs X and Y, while Row does not. They play the game repeatedly, and while Row observes Column's actions, she does not learn the payoffs (this could happen if the payoffs are themselves random variables, of which only Column knows the distributions). How is Row to prevent herself from being exploited, given her ignorance of the game?

Theorem 1. *In this game, probability matching is a minimax strategy for Row.*

Proof. Suppose Row matches Column's play, so that $p = q$. Then her expected payoff is

$$\pi^R(q, q) = q^2 X + (1-q)^2 Y. \tag{23}$$

Now, Column must respond by solving

$$\min_q \pi^R(q, q), \tag{24}$$

which yields the first order condition

$$\begin{aligned} & \frac{d}{dq}\pi^R(q,q) = 0 \\ \implies & \frac{d}{dq} q^2 X + (1-q)^2 Y = 0 \\ \implies & 2q^{**}X - 2(1-q^{**})Y = 0 \implies q^{**} = \frac{Y}{X+Y} = q^*. \end{aligned} \tag{25}$$

As

$$\frac{d^2}{dq^2} = 2X + 2Y > 0, \tag{26}$$

q^{**} is indeed a minimum, and as $q^{**} = q^* = p^*$,

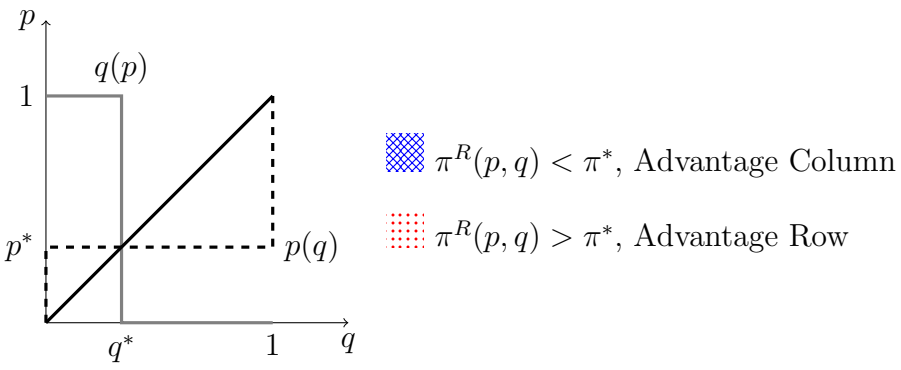

Figure 4: Row's matching strategy

$$\pi^R(q,q) \geq \pi^* \quad \forall q. \tag{27}$$

□

This can also be seen in Figure 4, which shows that the line $p = q$ lies in a region of the pq-plane where $\pi^R(p,q) \geq m$. Therefore, if Row can match Column's strategy, she is guaranteed an expected payoff of at least m. The best that Column can then do is to play q^*, his minimax strategy under complete information.

5.3 Feedback Equivalence

Consider an experiment in which a forecaster must choose between two predictions, A_1 and A_2, based on her true belief that A_1 is correct with probability p and A_2 with probability $1-p$. Let S_1 denote the state of the world in which A_1 is correct, and S_2 that in which A_2 is correct, where $S_2 \equiv \neg S_1$. If she plays A_1 with probability p, i.e. she probability-matches, then there are two crucial similarities between this experience and the experience of playing a minimax strategy in the game in Figure 3.

Theorem 2. *In both probability matching, and playing the minimax strategy in the matching pennies game in Figure 3, a player's losses are distributed evenly between the two actions.*

Proof. First, in the experiment, suppose that S_1 occurs with probability p. Then if a player plays A_1 with probability p,

$$\begin{aligned} P[\text{loss}, A_1] = P[S_2, A1] &= (1-p)p, \text{ and} \\ P[\text{loss}, A_2] = P[S_1, A2] &= p(1-p). \end{aligned} \tag{28}$$

Thence, Bayes' Rule yields

$$P[A_1|\text{loss}] = P[A_2|\text{loss}] = \frac{1}{2}. \tag{29}$$

Now, in the game in question, recall that there is a unique equilibrium in which $p^* = q^*$. From the payoff matrix, we see that in equilibrium

$$\begin{aligned} P[\text{loss}, T] = P[R, T] &= (1-q^*)p^* = (1-p^*)p^*, \text{ and} \\ P[\text{loss}, B] = P[L, B] &= q^*(1-p^*) = p^*(1-p^*). \end{aligned} \tag{30}$$

Again, Bayes' Rule yields

$$P[T|\text{loss}] = P[B|\text{loss}] = \frac{1}{2}. \tag{31}$$

□

A more interesting similarity between the two experiences arises when we consider endogenous reinforcement. Given what we now know about dopamine (Schultz et al., 1997), suppose we assume that Brackbill and Bravos (1962) were correct to conjecture that the 'payoff' to correctly forecasting an event is inversely proportional to the likelihood of that event occurring. Thus, a participant in a two-armed bandit experiment who receives no

financial performance incentives has a utility function $\Psi(A, S)$ of the form

$$\Psi(A_i, S_j) = \begin{cases} \frac{1}{P[S_j]} & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{32}$$

Theorem 3. *For a participant with a utility function of the form given in* (32)*, the expected payoffs of both actions are equal, just as in playing matching pennies against an opponent who plays minimax.*

Proof. From (32) and given $P[S_1] = p, P[S_2] = 1 - p$,

$$\begin{aligned} E[\Psi(A_1, \cdot)] = &\quad p\tfrac{1}{p} &&= 1, \text{ and} \\ E[\Psi(A_2, \cdot)] = &\quad (1-p)\tfrac{1}{1-p} &&= 1. \end{aligned} \tag{33}$$

□

This last theorem, as stated, relies on a specific functional form for the endogenous reinforcement function, which may be a dubious assumption. However, if we replace 'inversely proportional' with 'negatively correlated', we find that the flavor of the experience remains: one action has a higher payoff, but pays off more rarely, so the expectations somewhat balance out.

5.4 Scholium

I offer the following explanation for probability matching:

1. Humans have evolved to see agency in randomness, and to err to the side of paranoia (Barrett, 2000, e.g.).

2. Probability Matching is an optimal behavior in a non-cooperative game (Theorem 1).

3. The endogenous reward system makes probability matching an equivalent experience to playing this game in equilibrium (Theorems 2 and 3).

$\therefore$ Subjects may display probability matching behavior in two-armed bandit experiments.

5.5 Discussion

This explanation of probability matching is fundamentally different from any currently in the literature. As it is teleological, it is compatible with most of the mechanistic explanations currently extant (e.g. Fantino and Esfandiari, 2002; Koehler and James, 2010). It offers a clear example of how endogenous reinforcement could lead to sub-optimal behavior in the face of pure randomness.

6 Learning Simulations in Zero-Sum Games

6.1 Backgound

Erev and Roth (1998) test several learning rules against the results of 12 different experiments in which subjects played games with unique, mixed-strategy equilibria. Their computational approach is well-suited to my question of how endogenous reinforcement can help agents learn and play mixed strategies. Additionally, their method of comparing simulations results to experimental results could be another way to test ER. I focus on their learning rule, with some modifications to represent endogenous reinforcement.

6.2 Erev & Roth

I first attempted to replicate the findings of Erev and Roth, for the first experiment they looked at in Suppes and Atkinson (1960), henceforth S&A2. The payoff matrix for this game is given in Figure 5; payoffs represent the chance of receiving a reinforcement on that round. On all rounds, each player earns a payoff π_t of either 1 or 0; the two players' winnings are calculated independently on each round. Erev and Roth simulated their learning model for 200 trials, the same as in the experiment by Suppes and Atkinson. I ran mine for 2000 trials,

	A_1^2	A_2^2
A_1^1	1/3, 2/3	1, 0
A_2^1	1/2, 1/2	1/6, 5/6

Figure 5: Payoff matrix for S&A2

since the long-run behavior visible after 2000 trials is often quite different from what appears after 200. The results of 100 runs of 2000 rounds each are shown in Figure 6.

The Erev & Roth learning rule uses propensities $q(A)$, which are updated in each round by

$$q_{t+1}(A) = \begin{cases} q_t(A) + \pi_t & \text{if agent played } A \text{ at time } t \\ q_t(A) & \text{otherwise} \end{cases}. \tag{34}$$

Thence,

$$\theta_t(A) = \frac{q_t(A)}{\sum_{A \in \mathscr{A}} q_t(A)} \tag{35}$$

The minimax strategies for S&A2 are $\theta^{1*} = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}$ and $\theta^{2*} = \begin{pmatrix} 5/6 \\ 1/6 \end{pmatrix}$. In the simulations, θ^2 appears to approach θ^{2*} in almost all simulations. In 100 simulations using initial propensities of 20,[12] the average value for $\theta^2_{2000}(A_1)$ was $0.88, SD = 0.030$, just above the minimax value of 0.833. However, θ^1 varies widely, and does not appear to converge to θ^{1^*}. Its average terminal value was $0.70, SD = 0.059$.

[12]This value is qualitatively the same as Erev and Roth's fit parameter $s(1) = 54$

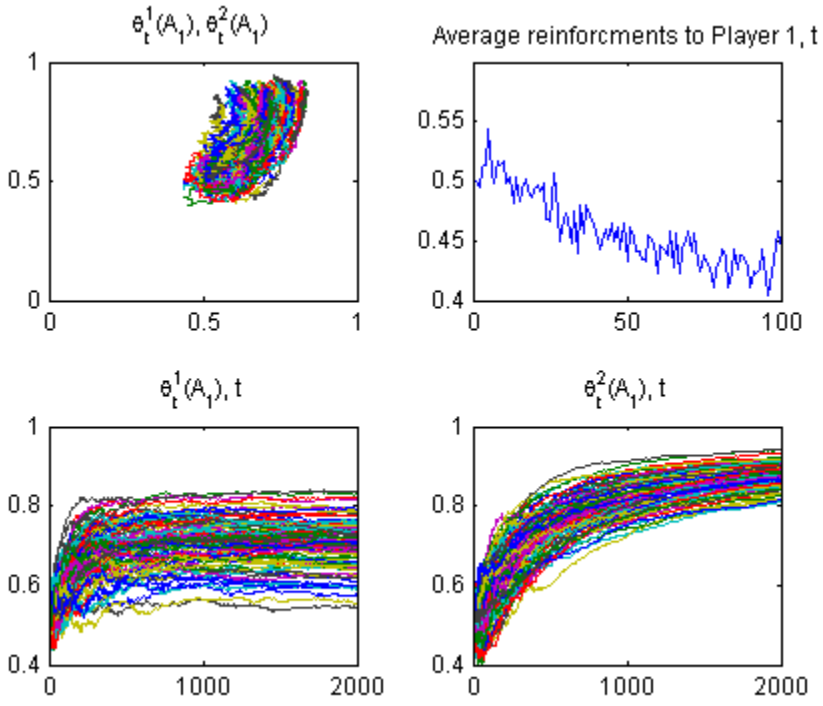

Figure 6: Results of simulations of S&A2 using the Erev-Roth learning rule (34). Different colored lines represent different runs of the simulation. Top Left: Parametric plots of $\theta_t^1(A_1), \theta_t^2(A_1)$. Top right: Average payoff to Player 1 over blocks of 20 rounds. Bottom Left and Right: $\theta_t^1(A_1)$ and $\theta_t^2(A_1)$ as functions of time.

6.3 Endogenous Reinforcement

I tested a modification of the Erev-Roth learning rule, which includes endogenous reinforcement when the action chosen had a probability less than $1/n$ of being chosen, where $n = \frac{1}{|\mathscr{A}|}$:

$$q_{t+1}(A) = \begin{cases} q_t(A) + \pi_t + \delta_t & \text{if agent played } A \text{ at time } t \text{ and } \theta_t(A) < \frac{1}{n} \\ q_t(A) + \pi_t & \text{if agent played } A \text{ at time } t \text{ and } \theta_t(A) \geq \frac{1}{n} \\ q_t(A) & \text{otherwise} \end{cases}, \tag{36}$$

where

$$\delta_t = \frac{\left(\frac{\sum_{A' \in \mathscr{A}} q_t(A')}{n} - q_t(A)\right) \alpha \pi_t}{\sum_{A' \in \mathscr{A}} q_t(A')}, \tag{37}$$

with α a parameter greater than zero. Thence, as before

$$\theta_t(A) = \frac{q_t(A)}{\sum_{A \in \mathscr{A}} q_t(A)} \tag{38}$$

Plots below show the results of simulations for various values of α. In all figures, the four subfigures are the same as in Figure 6. Table 1 shows average values of $\theta(A_1)$ after 2000 trials, averaged over 100 simulations, and standard deviations, for several values of α. With $\alpha = 0$, (36) reduces to (34).

For positive values of α up to .8, the simulations generate cycles on the boundary of the strategy space, as shown in Figure 7. At $\alpha = .8$, simulations yielded some straight lines and some cycles; at $\alpha = .9$ there were almost only straight lines (Figure 8).

It appears, then, that for small values of α, the learning rule (36) greatly delays divergence to the boundary, but tends to flip-flop back and forth between nearly-pure strategies. For larger values, the learning rule tends to lock into specific mixed strategies early on.

α	$\theta^1_{2000}(A_1)$	SD	$\theta^2_{2000}(A_1)$	SD
0	.88	.03	.70	.06
0.1	.64	.48	.80	.40
0.4	.55	.50	.84	.37
0.6	.61	.49	.89	.31
0.8	.70	.29	.62	.22
0.9	.54	.10	.52	.05

Table 1: Terminal average strategies and standard deviations for both players.

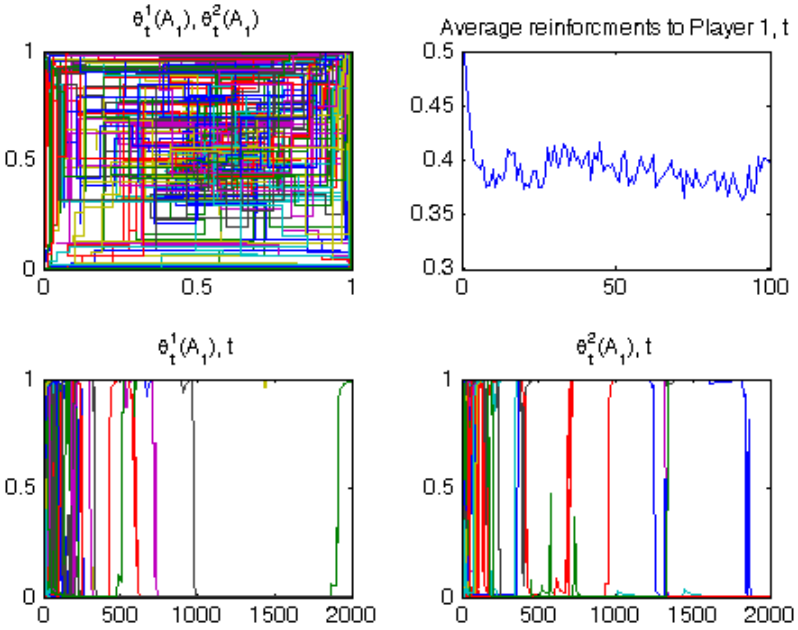

Figure 7: Results of simulations of S&A2 using the Erev-Roth learning rule with endogenous reinforcement (36), with $\alpha = .4$.

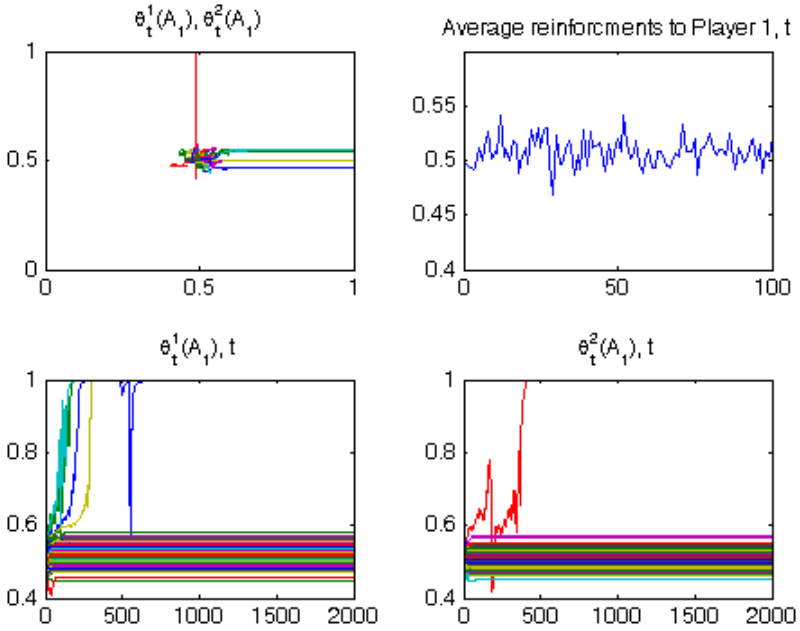

Figure 8: Results of simulations of S&A2 using the Erev-Roth learning rule with endogenous reinforcement (36), with $\alpha = .9$.

6.4 Arousal as a Separate Variable

The Erev-Roth learning rule has the strength of storing only one value for each possible action. This simplicity makes it more likely that such a rule could map onto actual cognitive functions. However, for several decades now, emotion researchers have recognized two primary dimensions of mood: affect (or pleasure/displeasure) and arousal (Zevon and Tellegen, 1982; Bradley et al., 2001). The most common model of emotion consists of positive affect and negative affect as two unipolar components, both correlated with arousal (see Figure 9, from Bradley et al., 2001). Positive and negative affect are associated with approach and withdrawl behavior, respectively (Watson et al., 1999).

I constructed a learning model in which arousal, r_t, is a single variable which varies with time, and each action has a valence $q_t(A)$ associated with it which is equivalent to propensity in the Roth-Erev learning model. Now, in addition to a vector of propensities equal in dimension to the action space, each agent must store a single scalar. Endogenous

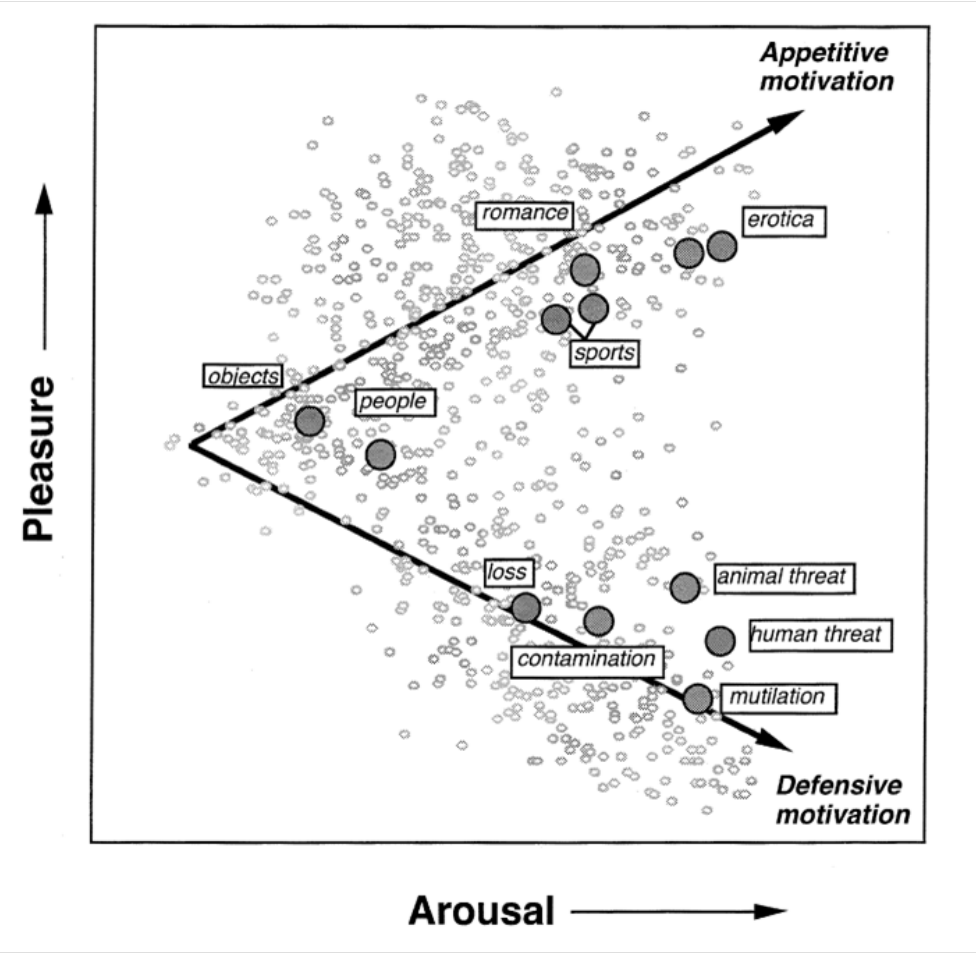

Figure 9: Average pleasure and arousal ratings for a variety of pictures. From Bradley et al. (2001).

reinforcement raises arousal, which also decays with time. Arousal mediates propensity updating. The model is as follows:

Arousal is updated first after each round by

$$r_t = \gamma r_{t-1} + \left| \frac{q_t(A_t)}{\sum_{A' \in \mathscr{A}} q_t(A')} - \pi_t \right|, \tag{39}$$

where A_t is the action played at time t.

Next, valences are updated by

$$q_{t+1}(A) = \begin{cases} q_t(A) + \pi_t r_t & \text{if agent played } A \text{ at time } t \\ q_t(A) & \text{otherwise} \end{cases}. \tag{40}$$

In this model, the crucial parameter is γ, the decay constant of arousal. Figures 10 and 11 show the results of 100 runs of this model with $\gamma = .3$ and $\gamma = .9$, respectively. In the former case, the model generates high variability at first, but gradually converges to a narrow range of the strategy space. In the latter, while player 2's strategy approached something near the minimax strategy in most runs, player 1's strategy was all over the place. In both cases, and for all other values of γ tested, the addition of arousal creates great variability early in every run, before the valence terms grow large enough that any change has minimal impact.

This two-factor learning model yields promising results. It fits well with the affective neuroscience literature. It generally prevents agents from diverging to the boundary of the strategy space, but does not have a built-in internal asymptote, as many reinforcement learning rules do (Bush and Mosteller, 1951, e.g.). It yields the same sort of initial exploration followed by settling into a stable strategy that is often observed in experiments.

However, it is far from perfect. It does not even approximately converge to the minimax strategy for player 1. This illustrates one of the difficulties in converging to a Nash Equilibrium in zero-sum games: once one player learns to play her minimax strategy, the other

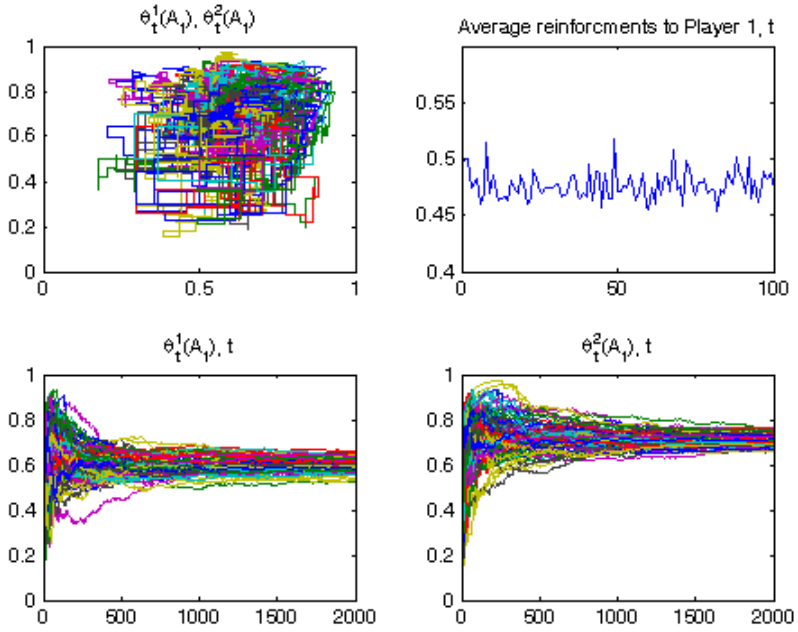

Figure 10: Results of simulations of S&A2 using the Erev-Roth learning rule with arousal (40), with $\gamma = .3$.

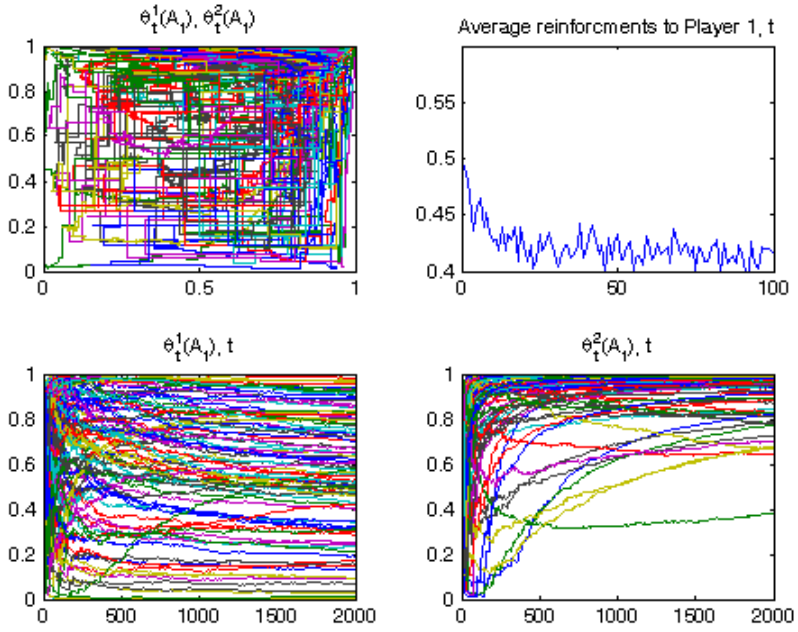

Figure 11: Results of simulations of S&A2 using the Erev-Roth learning rule with arousal (40), with $\gamma = .9$.

player becomes indifferent between all actions, and so may not learn his minimax strategy. This learning rule also appears highly sensitive to the choice of decay parameter γ.

6.5 Lessons from the Simulations

Endogenous reinforcement can have a variety of effects on the behavior of a learning rule for a noncooperative game. It can have the originally desired effect of preventing agents from adopting pure strategies. However, when it fails to prevent mixed strategies, it can push agents into them more quickly, or can lead to flip-flopping between pure strategies rather than true mixing. Even with a single choice of learning rule and initial conditions, endogenous reinforcement can lead to wildly different behavior depending on the parameters used.

7 Discussion

My main goal here has been to advance a philosophical thesis, and then show some possible results of that thesis for the study of decision making in economics and psychology. I have made several points. First, it does not make sense to expect humans to think or behave in accordance with the axioms of statistical decision making. Second, considering strategic parnoia can offer a new class of solutions to traditional paradoxes of choice under uncertainty. Third, the phenomenology of the dopamine learning system may be adaptive in that it facilitates protrean behavior. Fourth, learning models which include such endogenous reinforcement can implement long-term concave utility by short-term risk seeking.

8 References

Allais, M. (1953): "Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine," *Econometrica*, 503–546.

Arrow, K. J. (1958): "Utilities, Attitudes, Choices: A Review Note," *Econometrica*, 26, 1–23.

Barrett, J. (2000): "Exploring the natural foundations of religion," *Trends in Cognitive Sciences*, 4, 29–34.

Bernoulli, D. (1738/1954): "Specimen theoriae novae de mensura sortis," *Econometrica*, 5, 175–192, trans. Sommer, Louise.

Bertram, B. (1980): "Vigilance and group size in ostriches," *Animal Behaviour*, 28, 278–286.

Bomze, I. (1986): "Non-cooperative two-person games in biology: a classification," *International journal of game theory*, 15, 31–57.

Borel, E. (1921/1953): "The Theory of Play and Integral Equations with Skew Symmetric Kernels," *Econometrica*, 21, 97–100, trans. Savage, L.J.

——— (1924/1953): "On Games that Involve Chance and the Skill of the Players," *Econometrica*, 21, 101–115, trans. Savage, L.J.

Börgers, T. and R. Sarin (1997): "Learning Through Reinforcement and Replicator Dynamics," *Journal of Economic Theory*, 77, 1 – 14.

——— (2000): "Naive reinforcement learning with endogenous aspirations," *International Economic Review*, 41, 921–950.

Brackbill, Y. and A. Bravos (1962): "Supplementary report: The utility of correctly predicting infrequent events," *Journal of Experimental Psychology*, 64, 648–649.

Bradley, M., M. Codispoti, B. Cuthbert, and P. Lang (2001): "Emotion and Motivation I: Defensive and Appetitive Reactions in Picture Processing," *Emotion*, 1, 276–298.

Bush, R. R. and F. Mosteller (1951): "A mathematical model for simple learning," *Psychological Review*, 58, 313–323.

——— (1955): *Stochastic Models for Learning*, New York: John Wiley & Sons, Inc.

Chance, M. and W. Russell (1959): "Protean displays: a form of allaesthetic behaviour," *Proceedings of the Zoological Society of London*, 132, 65–70.

Cournot, A. A. (1838/1897): *Researches into the Mathematical Principles of the Theory of Wealth*, New York: MacMillan, trans. Bacon, Nathaniel T.

Cross, J. G. (1973): "A Stochastic Learning Model of Economic Behavior," *The Quarterly Journal of Economics*, 87, 239–266.

Dennett, D. C. (1984): *Elbow Room: The Varieties of Free Will Worth Wanting*, Cambridge, MA: MIT Press.

Edwards, W. (1956): "Reward probability, amount, and information as determiners of sequential two-alternative decisions," *Journal of Experimental Psychology*, 52, 177–188.

Ellsberg, D. (1961): "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 75, 643–669.

Erev, I. and A. E. Roth (1998): "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, 88, 848–881.

Fantino, E. and A. Esfandiari (2002): "Probability Matching: Encouraging Optimal Responding in Humans," *Canadian Journal of Experimental Psychology*, 56, 58–63.

Fudenberg, D. and D. K. Levine (1998): *The Theory of Learning in Games*, Cambridge, MA: MIT press.

Goodnow, J. J. (1955): "Determinants of Choice-Distribution in Two-Choice Situations," *The American Journal of Psychology*, 68, 106–116.

Grant, D., H. Hake, and J. Hornseth (1951): "Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement," *Journal of Experimental Psychology*, 42, 1–5.

Hanappi, H. (2008): "The concept of choice: why and how innovative behaviour is not just stochastic," *Journal of Evolutionary Economics*, 18, 275–289.

Hazy, T. E., M. J. Frank, and R. C. O'Reilly (2010): "Neural mechanisms of acquired phasic dopamine responses in learning," *Neuroscience & Biobehavioral Reviews*, 34, 701 – 720, special Section: Dopaminergic Modulation of Lifespan Cognition.

Huygens, C. (1657/1713): *Libellus de Ratiociniis in Ludo Aleae*, London: T. Woodward.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–291.

Koehler, D. J. and G. James (2010): "Probability matching and strategy availability," *Memory & Cognition*, 38, 667–676.

K
"uhberger, A. and J. Perner (2003): "The role of competition and knowledge in the Ellsberg task," *Journal of Behavioral Decision Making*, 16, 181–191.

Leonard, R. (2010): *Von Neumann, Morganstern, and the Creation of Game Theory*, New York: Cambridge University Press.

Machiavelli, N. (1531/1997): *Discourses on Livy*, Oxford: Oxford University Press, trans. Bondanella, Julia Conaway and Bondanella, Peter.

Marshall, A. (1890/1920): *Principles of Economics*, London: MacMillan & Co., 8th ed.

Maynard Smith, J. (1982): *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

McMahon, A. and M. Scheel (2010): "Glucose promotes controlled processing: Matching, maximizing, and root beer," *Judgment and Decision Making*, 5, 450–457.

Miller, G. F. (1997): "Protean primates: The evolution of adaptive unpredictability in competition and courtship," in *Machiavellian Intelligence II*, ed. by A. Whiten and R. W. Byrne, Cambridge: Cambridge University Press.

Miller, M., M. Valsangkar-Smyth, S. Newman, H. Dumont, and G. Wolford (2005): "Brain activations associated with probability matching," *Neuropsychologia*, 43, 1598–1608.

Montague, P., P. Dayan, and T. Sejnowski (1996): "A framework for mesencephalic dopamine systems based on predictive Hebbian learning," *Journal of Neuroscience*, 16, 1936.

Nash, J. F. (1950): "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences of the United States of America*, 48–49.

Niou, E. M. S. and P. C. Ordeshook (1994): "A Game-Theoretic Interpretation of Sun Tzu's: The Art of War," *Journal of Peace Research*, 31, 161–174.

Norman, M. F. (1968): "Some Convergence Theorems for Stochastic Learning Models with Distance Diminishing Operators," *Journal of Mathematical Psychology*, 5, 61–101.

Nowak, M. A. (2006): *Evolutionary Dynamics: Exploring the Equations of Life*, Cambridge, MA: Harvard University Press.

Pascal, B. and P. d. Fermat (1959): "Correspondence of 1654," in *A Source Book in Mathematics, Vol. 2*, ed. by D. E. Smith, New York: Dover.

Ross, D. (2009): "Integrating the dynamics of multi-scale economic agency," in *The Oxford Handbook of Philosophy of Economics*, ed. by H. Kincaid and D. Ross, Oxford: Oxford University Press.

Savage, L. J. (1954): *The Foundations of Statistics*, New York: John Wiley & Sons, Inc.

Scannell, J., G. Roberts, and J. Lazarus (2001): "Prey scan at random to evade observant predators," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268, 541.

Schultz, W., P. Dayan, and P. R. Montague (1997): "A Neural Substrate of Prediction and Reward," *Science*, 275, 1593–1599.

Shanks, D., R. Tunney, and J. McCarthy (2002): "A re-examination of probability matching and rational choice," *Journal of Behavioral Decision Making*, 15, 233–250.

Smith, A. (1776/1937): *An Inquiry into the Nature and Causes of the Wealth of Nations*, New York: The Modern Library.

Sun Tzŭ (1964): *The Art of War*, Taipei: Literature House, Ltd., trans. Giles, Lionel.

Suppes, P. and R. C. Atkinson (1960): *Markov Learning Models for Multiperson Interactions*, Stanford: Stanford University Press.

Sutton, R. and A. Barton (1998): *Reinforcement Learning*, Cambridge, MA: MIT Press.

Thorndike, E. L. (1905): *The Elements of Psychology*, New York: A. G. Seiler.

von Neumann, J. (1928): "Zur theorie der gesellschaftsspiele," *Mathematische Annalen*, 100, 295–320.

von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Vulkan, N. (2000): "An economist's perspective on probability matching," *Journal of Economic Surveys*, 14, 101–118.

Watson, D., D. Wiese, J. Vaidya, and A. Tellegen (1999): "The Two General Activation Systems of Affect: Structural Findings, Evolutionary Considerations, and Psychobiological Evidence," *Journal of Personality and Social Psychology*, 78, 820–838.

Wolford, G., M. B. Miller, and M. Gazzaniga (2000): "The Left Hemisphere's Role in Hypothesis Formation," *Journal of Neuroscience*, 20, RC64.

Zevon, M. and A. Tellegen (1982): "The Structure of Mood Change: An Idiographic/Nomothetic Analysis," *Journal of Personality and Social Psychology*, 43, 111–122.