

# A Note on Exploring Rationality in Games

Steven O. Kimbrough  
University of Pennsylvania  
Philadelphia, PA 19004  
kimbrough@wharton.upenn.edu

15 March 2004

## **Abstract**

The received concept of strategic (game-theoretic) rationality is attended by a formidable list of paradoxes, anomalies, and empirical failures. This paper reviews three well-known and problematic decision contexts and diagnoses as a common source of difficulty the failure in the received view of rationality to adequately recognize risk-return tradeoffs. This diagnosis is then supported by computational experiments and analysis that demonstrate the effectiveness in strategic contexts of specific and plausible forms of decision making that do better justice to risk-return tradeoffs. The paper suggests that this kind of rationality, called an exploring rationality, be considered as an alternative to the received view.

Rational agents, according to the standard account of rationality in classical game theory,<sup>1</sup> are omniscient, deductively omnipotent maximizers of (their own) expected utility. Games are ‘solved’—their outcomes predicted—by ascertaining their equilibria (particularly Nash equilibria and refinements thereof) and showing how rational agents, so defined, will reach these equilibria.

These statements, of course, constitute a caricature bordering on parody of a thoughtful, vibrant, and rich literature. Agents are not required to be generally omniscient and deductively omnipotent, they merely need to have common knowledge of the game (including the rationality of their counter-players) and be able reliably to make all relevant deductions. (See Shubik [34] for an especially forthcoming account.) And forays beyond these stringent assumptions are very much in play. (At the textbook level see, e.g., [2, 12].) Even so, this characterization—of the standard account of rationality in games (SARG)—is sufficiently accurate to serve present purposes. These purposes are to motivate and conduct in small part an examination of alternate concepts of rationality in contexts of strategic interaction (games, in the technical sense of game theory).

The *prima facie* implausibility of the standard account of rationality in games (SARG) has not, indeed has never, gone unnoticed. Accumulating experimental evidence over the last 40 years and more, however, has led to reconsideration and search for alternatives on the part of many researchers in game theory. (See [4, 19] for useful compendia of results.) The following passage from a recent textbook on game theory is representative of the thoughts of at least many researchers.

Ironically, game theory is often hoisted on its own pétard: many of its most fundamental predictions—predictions that would have been too vague to test with any confidence in the pre-game-theoretic era—are *decisively and repeatedly disconfirmed*, in laboratory settings, with substantial agreement among experimenters, regardless of their theoretical priors. [12, page xxiv]

Many other worries can be added to a list of legitimate concerns. I am especially struck by two of a computational nature. First, equilibria in games are often so computationally complex as to be intractable from any realistic perspective. Second, as proved in [13], predicting outcomes of games can be undecidable, even for relatively simple games (such as spatialized Prisoner’s Dilemma). In either case, rationality—or the standard account thereof—degenerates to a counsel of perfection, impossible to follow.

One response to this situation is to eschew rationality as an explanatory, or even relevant, concept. Instead of agents having strategies (and having them wisely, rationally, or not etc.), the focus is on the behavior of the strategies themselves. Agents are secondary, they merely carry strategies. The action focuses on the dynamics of (evolutionary) interplay among the strategies. That is the tack taken by Gintis (and others). Here is his motivating description of the programme.

... *game theory is about the emergence, transformation, diffusion, and stabilization of forms of behavior.* Traditionally, game theory has been seen as a theory of how “rational agents” *do* behave, and/or how the rest of us *should* behave. Ironically, game theory which for so long was predicated upon agent rationality, has shown us, by example, the shakiness of the concept. For one thing, the centipede game and others like it show that there is nothing substantively “rational” about even so simple a thing as eliminating dominated strategies . . . . Moreover, the solution to some games (even when unique) is often so sophisticated that it is implausible that ordinary people would be willing to spend the resources to discover it. This supports the evolutionary notion that good strategies diffuse across populations of players rather than being learned by “rational optimizers.” Finally, experimental studies of dictator, ultimatum, and public goods games indicate that if people are “rational,” it must be in a sense far more sophisticated than the simple, self-interested, maximization of expected utility.

---

<sup>1</sup>The literature is huge. Important exemplars include [10, 25, 34].

It is better to drop the term “rational” altogether, which is what we do in this book . . . .

In the same vein, we do not follow classical game theory in asking how agents “learn” to play optimal strategies, because the cognitive processes involved in “learning” are probably, under most conditions, much less important than the forms of imitation underlying the replicator dynamic. . . and cultural transmission. . . . In short, evolutionary game theory replaces the idea that games have “solutions” that agents “learn,” with the idea that games are embedded in natural and social processes that produce agents who play effectively.

Dispensing with the rationality postulate does not imply that people are *irrational* (whatever that means). The point is that the concept of “rationality” does not help us understand the world. [12, pages xxv-xxvi]

I wish to motivate and begin to develop here an alternative response, or programme of research, regarding rationality in contexts of strategic interaction. To this end, I shall discuss in §1 three decision contexts that, I argue, *suggest* something important about rationality and that motivate subsequent investigation. §2 then accepts the suggestion and describes a series of computational experiments involving learning in games. I conclude in §3 with a summary and brief comments.

First, a (mostly) terminological note. The word *rational* should not be ceded to any particular theory of rationality. Instead, I shall make free use of *rational*, *rationality* and so on in an atheoretic, ordinary language way, relying on context or an explicit device (e.g., *SARG*) to indicate when I am referring to a particular theory of rationality. I note that at least one economist of repute shares this practice.

Rationality is interpreted here, broadly, as the discipline of subjecting one’s choices—of actions as well as of objectives, values and priorities—to reasoned scrutiny. Rather than defining rationality in terms of some formulaic conditions that have been proposed in the literature (such as satisfying some prespecified axioms of “internal consistency of choice,” or being in conformity with “intelligent pursuit of self-interest,” or being some variant of maximizing behavior), rationality is seen here in much more general terms as the need to subject one’s choices to the demands of reason. [33, page 4]

## 1 Three Decision Contexts

Our first decision context is a game of Repeated Prisoner’s Dilemma (RPD), in which two players play a one-shot Prisoner’s Dilemma (PD) game multiple times. The one-shot Prisoner’s Dilemma game—serving as the *stage game* in the repeated game, RPD—involves two players each with two strategies: *C* (coöperate) and *D* (defect). In strategic (aka: normal) form the game is: with the requirement that  $T > R > P > S$

	C	D
C	R	S
D	T	P

Figure 1: Schema for 2×2 Symmetric Games

and that  $2 \cdot R > T + S$ . Mnemonically: *R*=Reward for mutual coöperation; *T*=Temptation to defect; *S*=the Sucker’s payoff; and *P*=the Penalty for mutual defection. Typically, even usually, in experiments  $T = 5, R = 3, P = 1$ , and  $S = 0$ . Since  $T > R$  and  $P > S$ , there is only one equilibrium point (EP): both

players play  $D$ . The dilemma, of course, is that if both players played  $C$ , both would be better off, since  $R > P$ .

In the one-shot PD game each player has a dominant strategy, which is to play  $D$ . The outcome  $(D, D)$  is a Nash equilibrium because neither agent/player has an incentive to regret its decision, given the decision by the other player. Case closed according to SARG; both players, if rational, will defect. The situation is more complicated when the game (PD or indeed any game) is repeated. There are two cases: infinite or indefinite repetition, in which after each stage of play there is a non-zero probability of playing another game; and definite or finite repetition, in which it is announced to the players that a specific number of stages will constitute the repeated game. In the case of indefinite repetition, the famous (and genuine) Folk Theorem of game theory applies: nearly any sequence of outcomes (other than uniformly the worst possible outcome for one player) can be obtained by a Nash equilibrium.<sup>2</sup> In short, the number of equilibria explodes, so that ‘solving’ the game by predicting its outcome will be at an equilibrium point is not especially helpful.

Classical game theory is much more specific for definitely repeated Prisoner’s Dilemma (DRPD): there is one (subgame-perfect)<sup>3</sup> equilibrium and that is that both players always play  $D$ . The argument is by backwards induction. If there is only one game left to play, each player should defect, by the argument from dominance, since we are effectively in the one-shot case again. If there are two games left to play and on the last game both players will defect, then the next to last game should be seen under the one-shot view, . . . , and so on. At least as early as the mid-1950s this conclusion by backwards induction for DRPD made game theorists uneasy. Luce and Raiffa [25] note the argument with approval and avowed unease. They confess they would very often be inclined to behave ‘irrationally’.

Our second decision context is the well-known Surprise Exam Paradox (aka: Surprise Hanging Paradox).<sup>4</sup> The scene is a classroom. The teacher announces that there will be a surprise examination given at one of the following six meetings of the class. The exam will be a surprise in the sense that on the morning of the day of the exam the students will not have enough information to know that the exam will occur that day with a probability at or above  $l$ , a given threshold (level or line), say  $l = \frac{2}{5}$ . The students find this puzzling and they reason as follows. “There cannot be a surprise exam on the sixth, the last, day, since we would know that morning that the exam had not yet taken place and hence that it would have to be given that day. But if we know that the surprise exam cannot happen on the sixth day, by similar reasoning it cannot happen on the fifth day. If the morning of the fifth day arrives without our having had the exam, we would know that it has to be given then, since it can’t happen on the sixth day. Therefore, the exam cannot be given on the fifth day either. By continuing this reasoning process we find that it is impossible for the teacher to give us a surprise exam during the next six meetings of the class.”

The students’ reasoning, of course, relies on backward induction, as does the game-theoretic reasoning to universal defection in DRPD, discussed above. As noted by Grim et al. [13, page 163]

The similarity of this reasoning to that of the argument for dominant defection throughout a series of known finite length is worth noting because of course the Surprise Examination is treated standardly in the philosophical literature as a *paradox*, thought to hide some fallacious piece of logical legerdemain. That the same form of reasoning is thought of as valid in the theoretical economics literature, though perhaps inapplicable in some practical sense, indicates that important work remains to be done in bridging the two bodies of work.

Should we conclude that backwards induction is a glory of game theory and a scandal of philosophy?<sup>5</sup>

<sup>2</sup>Binmore [2] has a clear and accessible presentation and proof.

<sup>3</sup>A technical matter that need not detain us. The interested philosopher will find a good discussion of the concept in the *Stanford Encyclopedia of Philosophy*, at <http://plato.stanford.edu/entries/game-theory/>.

<sup>4</sup>The literature is voluminous. A sample: [1, 3, 5, 7, 14, 29, 32, 37]. There is a useful bibliography at <http://www.magnolia.net/~leonf/paradox/hanging.txt>.

<sup>5</sup>Apologies to C.D. Broad who is reputed to have said of induction of the Humean type that it is “the glory of science and the

Our third decision context is non-strategic; it is not a game. The agents play only against nature, who acts without interests or foresight. Here is a description of the original experiment, which has been profusely replicated and confirmed in its results.

In the Humphreys' (1939 [17]) light-guessing experiment, a subject is seated before two bulbs and is instructed to predict which one will illuminate on each of a series of trials. Only one bulb lights on any trial. After his prediction, the subject is able to see whether or not he was correct by simply observing which light subsequently illuminates. The probability that a given bulb will light is fixed in advance and is typically figured for a block of trials. . . .

At the start of this experiment subjects typically distribute their choices equally between the two lights. As the experiment continues, they tend to increase their choices of the more frequently reinforced light. After one to two hundred trials, the behavior of the subjects generally stabilizes and they choose each light with the same probability with which the bulb illuminates; this is generally called a *matching strategy*. [28, pages 11–2]

The subjects' behavior is not optimal, even though they learn accurately the probabilities in question. Instead of matching, subjects would do better always to guess that the light with the higher (estimated) probability of being lit will light next.

How might we explain such behavior? Here is one *form* of explanation:

One theory and model... of decision making in the light-guessing experiment is based upon the assumption that the subject behaves as if he is maximizing his expected utility. The term, *utility*, as it is used here does not refer simply to the reward associated with each alternative, but also to any other considerations which may increase the subjective value of a particular choice. . . . Siegel has considered two sources of utility in the light-guessing experiment. The first is the utility of a correct choice, that is, the utility of the reward received for correctly predicting which light will illuminate on a trial. The second source of utility is that of choice variability resulting from the intrinsic boredom of a pure strategy (choosing the same light constantly), as well as from the greater satisfaction connected with being able to predict the less frequent light correctly. [28, page 12]

In other words, it is assumed that agents maximize expected utility. If their behavior indicates otherwise, sources of utility are added to the model until it fits the data. Ofshe & Ofshe [28], from whom these passages are quoted, extend this move to explain behavior in certain games. It is posited that subjects have utility functions for a number of factors, which they estimate, then act upon accordingly. The move on display—assuming that agents maximize utilities and adding sources of utility until models fit the data—is representative, and hardly peculiar to Siegel or Ofshe & Ofshe. Models thus produced are in principle testable and usable for generalizations, although results have not been impressive. It is at least worth asking whether any very different form of explanation for the light-guessing experiments might be apt.

Now to some analysis of the three cases. What I have to say is not aimed at giving a definitive and exclusive treatment of the three examples. Given that each is open to multiple interpretations, I doubt any such venture could succeed. My goal, instead, is to draw out a certain theme that I think unites these cases and that bears on the concept of rationality. I'll begin with the Surprise Exam paradox.

Suppose, more stringently, the teacher says, "Class, tomorrow I will give you a surprise exam." Is it possible for this to be true? If it is to be true, the exam is given tomorrow (*E*) and it is a surprise to the students (*S*). Determining whether the exam is in fact given is, I shall suppose, unproblematic. What counts as it being a surprise? Let us say that if (and only if), from the students' perspective, the probability of  

---

scandal of philosophy."

the exam's being given is below a critical level,  $l$ , and the exam is given, then the students are surprised. Formally,

$$(P(E) < l \wedge E) \leftrightarrow S \quad (1)$$

For the sake of the discussion, let us set  $l$  to  $\frac{2}{5}$ ; but any value  $l > 0$  will do. The teacher can give a surprise exam by defining a chance setup and a random variable, say  $E'$ , that takes on a value of 1 such that  $P(E' = 1) < l$ . The teacher then commits herself to giving the exam if and only if in a specified trial of the chance setup  $E' = 1$ . For example, the chance setup might be a single roll of a fair die and  $E' = 1$  if and only if a 1 or a 2 comes up on the specified trial. Then  $P(E' = 1) = \frac{1}{3} < l$ . The teacher rolls the die today and keeps the outcome secret. If and only if the die comes up 1 or 2, the teacher gives the exam the next day and the students are surprised (and so is the teacher, if asked before the die is rolled). Summarizing:

1.  $l = \frac{2}{5}$  (or any value  $> 0$ )
2.  $(P(E) < l \wedge E) \leftrightarrow S$
3.  $P(E' = 1) < l$  (by design of the chance setup)
4.  $P(E' = 1) = P(E)$  (by commitment of the teacher)
5.  $E' = 1 \leftrightarrow E$  (by commitment of the teacher)

$$\models (E \wedge S) \vee (\neg E \wedge \neg S)$$

And  $(E \wedge S)$  with probability  $P(E) = P(E' = 1) < l$ . The teacher can give a surprise exam tomorrow, but she runs some risk of speaking falsely when making the announcement. In our example, the teacher's risk,  $R_T$ , is  $(1 - P(E' = 1)) = \frac{2}{3}$ . That is a lot of risk for the teacher to take, but if she is willing to take that risk, it becomes possible to announce and succeed in giving a surprise exam tomorrow. The students' error is in failing to see this, in failing to appreciate that the teacher faces a risk-return tradeoff and may (dare we say rationally?) choose to make the tradeoff in a manner that accepts substantial risk for the benefit of having a chance at giving a surprise exam tomorrow. Depending on  $l$ , which itself depends on the students' attitudes, the teacher may be saying something that is probably false. My point is that what she says might be true and that it is wrong to infer that it can't be true.

More realistically, the teacher may announce a surprise exam sometime during the next  $n$  days and in doing so reduce her risk of speaking falsely. That risk cannot be eliminated, but by increasing  $n$  it may be reduced to below any fixed positive number. (I leave the details to the reader.)

The theme I wish to draw out of the Surprise Exam paradox is the (rational) tradeoff between risk and reward. Failure to recognize this tradeoff is, I claim, sufficient to undo the backwards induction argument in this case. I am not claiming to have 'solved' the Surprise Exam paradox, for it is open to too many interpretations to admit of any one solution. My purpose is served by noting that the under-appreciated tradeoff between risk and reward yields under-appreciated empowerment.

Returning now to DRPD, the definitely-repeated Prisoner's Dilemma, recall the 'unease' recorded by Luce and Raiffa over the conclusion that universal defection is the only rational course. One's unease may be strengthened by considering a parameterization of the PD stage game as in Figure 2. Let  $\epsilon > 0$  and for simplicity, let  $S = 0$ . Suppose  $B = \$1,000$ ,  $\epsilon = 1\text{¢}$ , and  $n$  (the number of repetitions) is 100. Knowing you are playing another human being, would you ever try cooperating to see if  $(B, B)$  might be obtained and even sustained? If not, what about  $B = \$1,000,000$ ,  $\epsilon = 0.001\text{¢}$ , and  $n = 10,000$ ? Surely at some point, contrary to SARG, it would be irrational not to venture a little cooperation in hopes of inducing jointly-beneficial outcomes.

On the other hand, if the stage game were parameterized as in Figure 3 with  $B$  large and  $n$  and  $\epsilon$  small,

		C	D
		$B$	$B + \epsilon$
C	$B$	$S$	$S + \epsilon$
D	$B + \epsilon$	$S + \epsilon$	

Figure 2: Parameterized PD Stage Game, Encouraging Coöperation

		C	D
		$B$	$B + \epsilon$
C	$B$	$S$	$B - \epsilon$
D	$B + \epsilon$	$B - \epsilon$	

Figure 3: Parameterized PD Stage Game, Discouraging Coöperation

it is hard see why one would ever risk playing  $C$ . There is very little potential benefit and very much risk.

SARG does not distinguish these two cases. Both (by construction) conform to a PD and in DRPD there is always only one (subgame-perfect) equilibrium: everyone always defects. Of course, as proved in [24], if you believe your counter-player in DRPD is irrational, then it may be rational on your part to play  $C$ , at least sometimes. The suggestion here is that if a theory of rationality leads us to this point, then it is worth considering alternative theories of rationality. Surely, an account of rationality should recognize risk-return tradeoffs in DRPD. If the rewards are high and the risks are low, might that not encourage risk-taking because the counter-player is thought to be *rational*?

The light-guessing experiment is seemingly a very different case. It is not strategic and it does not involve backwards induction. There are two kinds of explanation for the subjects' behavior: (1) they are maximizing some non-obvious utility functions, and (2) they are out and out dumb or irrational. A third type of explanation is available: misapplied heuristic.

Consider another version of the experiment. Instead of two lights flashing at some unknown rate, a resource, such as food, appears at two different locations, again at some unknown rate. Further, assume there is competition for the resource and that the subject's chance of getting the resource is in direct proportion to the rate of appearance of the resource (analogous to the rate of flashing) times the number of competitors for the resource (a factor *not* present in the light-guessing experiment). Concretely, imagine a pond with ducks as subjects. At one end of the pond an experimenter (Mr. Red, analogous to the red light) throws bread crumbs into the water at rate  $r$ ; at the other end another experimenter (Ms. Green) throws crumbs at rate  $g$ . We can call this the (prototypical) foraging experiment. The experiment has been done and the ducks will segregate, with  $r/(r + g)$  of them at Mr. Red's end of the pond and  $g/(r + g)$  of them at Ms. Green's end. If the rates change or the size of the crumbs changes, the ducks will rapidly adjust (within a few minutes), maintaining an equal expectation of reward per duck in the pond. In short, the ducks do probability matching (as a special case; they are sensitive to the size of the rewards), or more generally, expectation matching. See [11, chapter 11] for a review of the literature.

If a duck applied its foraging heuristic—expectation matching—in the context of a light-guessing experiment, the duck's behavior would resemble (match?) that of humans in light-guessing experiments. Are the ducks then foraging irrationally? Not at all. In the foraging context (with competition for resources), expectation matching is optimal (*ceteris paribus*). If, instead, a breed of ducks concentrated exclusively on the more productive source of food, that breed could be evolutionarily invaded by an expectation-matching

breed, and eventually swamped to extinction. Expectation matching in the foraging context is an ESS, an evolutionarily stable strategy (see [26] for the concept of an ESS).

Putting the point another way, our theme of risk-return tradeoff returns because expectation matching is directly and explicitly a policy for making risk-return tradeoffs. In the foraging context it is an excellent strategy from an evolutionary perspective; it is what you'd like to give your kids. As should be expected, it is well entrenched in the animal kingdom (again, see [11, chapter 11]). Birds have it, bees have it, we have it. Subjects in the light-guessing experiments have it and, I submit, may be incorrectly using it to drive their choices. Incorrectly, but perhaps with a little charity not irrationally. The subjects are (I suggest) using a successful and well-entrenched heuristic in an unfamiliar situation bearing similarity to contexts in which the heuristic is correctly applied. Do we really want to declare this irrational, especially during the first few dozen trials?

Summing up, it seems a truism to say that rational decision making requires attention to, and principled response to, risk-return tradeoffs. Yet failure to do so, I would argue, underlies the Surprise Exam paradox, the insistence on the rationality of universal defection in Definitely Repeated Prisoner's Dilemma, and much interpretation of the light-guessing experiments. This raises the question of how behavior in games might be explained, and even justified rationally, when agents take into account risk-return tradeoffs. In the next section we begin to address this question by examining computational (hence transparent) agents in strategic contexts. The agents learn and adapt using simple, plausible algorithms that are responsive to risks and rewards.

## 2 Reinforcement Learning in Games

Thorndike's "law of effect" makes the apparently obvious assertion that behavior attended by reward viewed positively by the organism tends to recur (or be reinforced) and behavior attended by reward viewed negatively tends not to recur. The idea was articulated by behavioral psychologists into a theoretical framework for learning as based on "selection by consequences" [35]. Whether this kind of theory can be adequate for learning in general may be, and has been, doubted (see [11] for a compendium of work from the computational-representation alternate perspective). The "selection by consequences" view, however, does have the considerable virtue of affording simple operationalizations. This has led to a prospering sub-field in machine learning, called *reinforcement learning*.<sup>6</sup> Standard reference works are [39] and [18].

In non-strategic contexts, reinforcement learning has been proved to be attended by a number of attractive properties, and has shown considerable success in applications. Its success to date in games played by artificial agents [6, 16, 31], and in modeling behavior of players in games [4, 9, 30], has been more modest. In the latter case, modeling of human subjects, reinforcement learning models easily outperform predictions from classical game theory, but models with additional factors appear to be improvements on pure reinforcement learning models.

These are relevant, but somewhat peripheral, considerations for purposes to hand. These purposes are (1) to present and discuss the performance of a simple (and typical) reinforcement learning model as prototype of a kind of rationality that explicitly is responsive to risk-reward tradeoffs, and (2) to present and discuss a new form of reinforcement learning model as a more sophisticated and plausible prototype of this kind of rationality. §2.1 focuses on point (1), followed by §2.2 and a discussion of point (2).

---

<sup>6</sup>Inspired in part by but not to be confused with reinforcement learning in the psychology literature. Actually, reinforcement learning in the machine learning sense is most closely allied with dynamic programming and Markov decision processes.

repeat forever:

1. Observe the current state,  $s_t$ .
2. Select the current action,  $a_t$ , from  $Q(s, a)$ .
3. Take action  $a_t$  and obtain reward  $r_t$ .
4. Update  $Q(s, a)$  based on  $r_t$ .

loop

Figure 4: Pseudo-Code for Q-Learning

## 2.1 Q-Learning: Learning in State Space

In so-called Q-learning [39, 18], an agent finds itself in  $s$ , one of many (two or more) possible states, is presented with a number (two or more) of actions it might take, chooses an action,  $a$ , and receives a return (positive or negative). In consequence, the agent updates its estimate of taking action  $a$  in state  $s$ ,  $Q(s, a)$ . See Figure 4 for a summary with pseudo-code.

Choice of action depends in a nondeterministic way on  $Q(s, a)$ . For the sake of simplicity, discussion here is limited to  $\epsilon$ -greedy action selection. Results are robust to reasonable action selection methods. Under  $\epsilon$ -greedy action selection, action  $a'$  is selected with probability  $(1 - \epsilon)$ , where  $a' = \arg \max_a Q(s, a)$ . In short,  $a'$  is an action that maximizes the current estimated  $Q$ -value for the present state; the agent takes an action that currently seems to give a highest return. Recognizing the need to explore and to make a risk-return tradeoff, the agent picks with probability  $\epsilon$  randomly among the available actions that do not maximize  $Q(s, a)$ . Whichever action is taken, after a return is obtained  $Q(s, a)$  is updated. For present purposes, the update rule is a simple linear learning rule, often used in psychology, having the form

$$\text{NewEstimate} = \text{CurrentEstimate} + \text{StepSize}\{\text{return} - \text{CurrentEstimate}\}$$

Stated in terms of the  $Q$  function,

$$Q'(s, a) = Q(s, a) + \alpha(r - Q(s, a)) \quad (2)$$

The step size, or  $\alpha$ , is typically set somewhere between 0.2 and 0.4. This update rule is an abstraction of the running average, in which  $\alpha = \frac{1}{n}$ . The effect of setting  $\alpha$  to a constant is to count recent observations more heavily and thereby to be responsive to a changing environment.

The object of learning in Q-learning is the  $Q$  function, which maps state-action pairs to an estimated value. For this reason, Q-learning may said to be an example of learning in state space. For future reference, it is helpful to summarize the Q-learning regime by describing its three principal elements. See Figure 5, for a summary in the context of repeated play of  $2 \times 2$  games with a memory of 1 game.

Consider now play by two Q-learning agents of Indefinitely Repeated Prisoner's Dilemma. The stage game is repeated 100,000 times in each run. This counts as indefinite repetition because the agents have no, and can represent no, information concerning the length of play. We parameterize the stage game, as in Figure 6 (a version of Figure 2, the encouraging scenario, with  $S = 0$ , but  $B$  is fixed at 3).

Summary results for 100 replications of runs of 10,000 plays are given in Table 1. Note that the level of mutual coöperation (CC) declines as  $\delta$  increases, even though throughout the stage game qualifies as a Prisoner's Dilemma. The behavior of these artificial, very simple, unminded agents accords with at least some intuitions of rationality. As  $\delta$  increases, the risk of playing  $C$  increases and the comparative benefit decreases, and *vice versa*. These agents do quite a good job of sorting this out. The right-most column in Table 1 is particularly significant. If both agents played  $C$  during the last 100 plays of all 100 replications (10,000 stage games in all), the agents of would each extract a total of  $3 \cdot 10000 = 30000$  points. This

1. Alternative (or consideration) set.

In the  $2 \times 2$  case, conditioning on the last play by the counter-player,  $\mathcal{A} = \{[0, 0], [0, 1], [1, 0], [1, 1]\}$  for each player. Interpretation of  $[x, y]$ : if on the last play the counter-player played  $x$ , on this play I play  $y$ , where 1=C and 0=D.

2. Attractiveness estimation.

E.g., linear updating rule for  $A^i$ ,  $i \in \mathcal{A}$ :

$$A_{t+1}^i = A_t^i + \alpha\{r_t^i - A_t^i\}$$

$$\text{NewEstimate} = \text{CurrentEstimate} + \text{StepSize}\{\text{reward} - \text{CurrentEstimate}\}$$

$$\text{In the case of Q-learning, } Q'(s, a) = Q(s, a) + \alpha(r - Q(s, a))$$

3. Choice/exploration policy.

$\epsilon$ -greedy. By default  $\epsilon = 0.05$ .

Figure 5: Summary of Elements in a Q-Learning Regime for  $2 \times 2$  Games

	C	D
C	(3,3)**	(0, 3+ $\delta$ )*
D	(3+ $\delta$ , 0)*	( $\delta$ , $\delta$ )#

Figure 6: A Parameterized PD Stage Game. #=Nash; \*=Pareto

is in an important sense the maximum value that can be expected to be extracted from the game; and it is a constant, regardless of  $\delta$ . The column entries show what percentage of the 30000 potential points were actually gleaned by the row player. (Results are not significantly different for the column player.) The agents are remarkably effective and consistent in extracting value from the game. The learning regime with its  $\epsilon$ -greedy exploration policy (a policy for making risk-reward tradeoffs) has robustly succeeded in returning value to our agents. Looking at the stage game, Figure 6, we observe that the Pareto optimal outcome,  $CC$ , is a much better predictor of what the agents will realize in returns than is the Nash equilibrium,  $DD$ .

The findings reported here for Repeated PD generalize well to other  $2 \times 2$  games [21]. They do not, however, generalize well in more difficult contexts. Cournot duopoly games serve as illustration. In a Cournot duopoly game (see any textbook on microeconomics) two firms supply a product to a market and must decide through reasoning and play (not by private negotiation) how much to produce. A monopolist, without competition, would produce an amount,  $q_m$ , that will maximize its economic profits. If instead competition is free among a large number of producers, the total output,  $q_f$ , will be much larger and none of the firms will receive a profit in the economic sense of extraordinary returns. (Remember: this is a model.) Finally, in the duopoly case, total output,  $q_c$ , will be at the Cournot equilibrium (it is also the Nash equilibrium but Cournot lived in the 19th century and gets precedence on this model), which is between the free competition equilibrium and the monopoly equilibrium. The duopolists will receive economic profits ('rents') but not as much as a monopolist will. Put differently, the consumers prefer free competition to duopoly and duopoly to monopoly.

The Cournot duopoly model figures large in economic reasoning and policy making. It is usually treated as a one-shot game, yet in many cases (such as spot markets for electricity) it is played daily or even more often. Surprisingly, very little experimental work has been done on human subjects in Cournot games. The best study I have found, by Holt [15], finds that human subjects (in the case of a simple model of costs and

CC	CD	DC	DD	$\delta$	Row's % CC
9422	218	183	177	0.05	0.963
9036	399	388	150	0.5	0.963
5691	738	678	2693	1	0.931
3506	179	275	6040	1.25	0.972
1181	184	116	8519	1.5	0.930
2	98	103	9797	1.75	0.805
97	114	91	9698	2	0.735
0	100	92	9808	2.5	0.839
2	96	94	9808	2.95	0.986

Table 1: Summary of Results for Prisoner’s Dilemma.  $\epsilon$ -greedy action selection. Totals are for the last 100 rounds of 100 series of 10,000 plays.

demands) chose production levels very near the predicted Cournot levels, shaded slightly down in favor of the monopoly (or collusion) level. Simulation of this game with Q-learning agents produced essentially the same result [20]. Both results seem odd. One would think, contrary to standard fare in economics textbooks, that rational agents in a repeated Cournot game should be able to learn to collude. If collusion can be learned in PD, why not in a Cournot game? Leaving aside for the moment the human behavior in the experiment, if, as is apparent, Q-learners cannot (very easily) learn to collude in repeated Cournot games, perhaps there is another principled form of learning that can, while recognizing the risk-return tradeoff.

## 2.2 Learning in Policy Space

One alternative to learning the value of state-action pairs, as in Q-learning, is to learn the values of the *policies* in a consideration set. By *policy* in a repeated game context, I mean a local strategy of play. For example, a policy might condition its action on the unfolding of play over the last one or two rounds. The idea is that a player selects a policy and turns play over to it, at least for a time. The policy “sees” the current state of play and generates the player’s action (deterministically or not, but here I consider only deterministic rules). To illustrate, in  $2 \times 2$  games (recall Figure 1), taking into account only the play by the counter-player in the last round of play, there are eight possible policies for each player. They may be coded as follows.

- 000 Play  $D$  on the first round of play using this policy, and play  $D$  thereafter.
- 001 Play  $D$  on the first round of play using this policy. After that, if the counter-player played  $D$  on the previous round, play  $D$  on the next round, and if the counter-player played  $C$  on the previous round, play  $C$  on the next round.
- 010 Play  $D$  on the first round of play using this policy. After that, if the counter-player played  $D$  on the previous round, play  $C$  on the next round, and if the counter-player played  $C$  on the previous round, play  $D$  on the next round.
- 011 Play  $D$  on the first round of play using this policy. After that, if the counter-player played  $D$  on the previous round, play  $C$  on the next round, and if the counter-player played  $C$  on the previous round, play  $C$  on the next round.
- 100 Play  $C$  on the first round of play using this policy, and play  $D$  thereafter.

repeat forever:

1. Select a policy  $\pi_i \in \Pi$ , where  $\Pi$  is the consideration set of policies.
2. Pick a length of play,  $k$ , for policy  $\pi_i$ .
3. Play the next  $k$  rounds of the game using  $\pi_i$ .  
Note: At each round,  $\pi_i$  will observe the current state,  $s_t$ , take an action  $a_t$  and obtain a reward  $r_t$ .
4. Update  $V^{\pi_i}$  based on the individual-round rewards,  $r_t$ s, obtained during the  $k$  rounds of play of policy  $\pi_i$ .

loop

Figure 7: Pseudo-Code for Policy-Space-Learning in Games

- 101 Play  $C$  on the first round of play using this policy. After that, if the counter-player played  $D$  on the previous round, play  $D$  on the next round, and if the counter-player played  $C$  on the previous round, play  $C$  on the next round. (This policy is called TIT FOR TAT.)
- 110 Play  $C$  on the first round of play using this policy. After that, if the counter-player played  $D$  on the previous round, play  $C$  on the next round, and if the counter-player played  $C$  on the previous round, play  $D$  on the next round.
- 111 Play  $C$  on the first round of play using this policy, and play  $C$  thereafter.

Figure 7 is the policy-space learning analog of Figure 4, which is for state-space learning. The processes are quite similar, with two key differences. In state-space learning (Q-learning), the consideration set consists of the state-action pairs recognized by the agent. In policy-space learning, the consideration set consists of policies (local, myopic instructions for play, *not* strategies for the entire game), which policies recognize states and actions. Each policy is complete in the sense that no matter what happens the policy is able to recommend an action. State-action pairs need not have this property, and do not in our examples above; at any given time, which state-action pairs ( $Q(s, a)$ ) are available depends on the current state. Policies, on the other hand, are responsive to all possible states.

The second way in which policy-space learning differs from state-space learning lies in the fact that policies are selected and played for a number of rounds,  $k$  in Figure 7, before their values (attractivenesses) are updated and selection of another policy is considered.

Figure 8 is the policy-space counterpart of Figure 5. Note especially that attractiveness estimation and choice/exploration policy are the same. What differs is the object of learning, the consideration set. Table 2 is the policy-space learning counterpart of Table 1. The results in the two cases are broadly equivalent.

Results differ in the much more complex Cournot duopoly game. In Holt's model, if each player produces 6, the monopoly position is achieved and each player makes a profit of 36. If each player produces 12 the competitive position results and each player makes a profit of 0. Finally, the Cournot/Nash equilibrium occurs when each player produces 8, for a profit of 32 each. The following five policies constitute a plausible consideration set for Holt's version of the Cournot duopoly game [22]. Note that production levels are real numbers (floating point numbers in a computer simulation), making the state-action space dense for practical purposes. Part of what policies do for the agents is to categorize state-action space in a way that makes learning and choice more manageable.

0. G-TFT. If  $y_{t-1} > y_{t-2}$ , then  $x_t = x_{t-1} + \delta$ ; else  $x_t = x_{t-1} - \delta$

1. Alternative (or consideration) set of policies:

In the  $2 \times 2$  case, conditioning on the last play by the counter-player,  $\mathcal{A} = \{000, 001, 010, 011, 100, 101, 110, 111\}$  for each player. Form: abc: play a the first time; if last time counter-player played 0, play b; if last time counter-player played 1, play c.

2. Attractiveness estimation: linear updating rule for  $A^i$ ,  $i \in \mathcal{A}$ :

$$A_{t+1}^i = A_t^i + \alpha \{r_t^i - A_t^i\}$$

$$\text{NewEstimate} = \text{CurrentEstimate} + \text{StepSize} \{ \text{reward} - \text{CurrentEstimate} \}$$

3. Choice/exploration policy.

$\epsilon$ -greedy. By default  $\epsilon = 0.05$ .

Figure 8: Summary of Elements in a Policy Space Learning Regime for  $2 \times 2$  Games

$\delta$	Average Payoff	Modal Strategy	Freq.	Est. Value	Row's % CC
0.05	2.7224	5	0.5319	2.885	0.907
0.5	2.7577	5	0.7571	2.901	0.919
1.0	2.8108	5	0.8731	2.926	0.937
1.25	2.8139	5	0.8623	2.933	0.938
1.5	2.8083	5	0.8381	2.932	0.936
1.75	2.7950	5	0.8011	2.935	0.932
2.0	2.7314	5	0.6604	2.918	0.910
2.5	2.5324	0	0.8164	2.613	0.844
2.95	2.9524	0	0.8643	3.056	0.984

Table 2: Summary of Results for Policy-Space Learning in Prisoner's Dilemma. Average Payoff over 800,000 rounds of play. Modal Strategy=most frequently-played strategy; 5 = TIT FOR TAT, 0 = ALL DEFECT

Comments. Mnemonic: GENEROUS TIT FOR TAT. Here and throughout these rules,  $\delta$  is fixed at 0.2. Under this strategy the agent incrementally reduces its production so long as its counter-player has not just increased its production.

1. BESTRESPONSE.  $x_t = 12 - 0.5y_{t-1}$ .

BESTRESPONSE is the strategy hypothesized by Cournot to lead to the Cournot/Nash equilibrium, which occurs at  $x_t = y_{t-1} = 8$  (in Holt's version of the game). In addition,  $\pi(x, y)$  (the profit to the player producing amount  $x$ , given that the other player produces amount  $y$ ) is maximized (for  $x$ ) at  $x = 12 - 0.5y$ . There is only one equilibrium and it is stable.

2. S-TFT.

(a) If  $x_{t-1} < y_{t-1}$  and  $y_{t-2} \leq y_{t-1}$ , then  $x_t = x_{t-1} + \delta$ .

(b) If  $x_{t-1} > y_{t-1}$  or  $x_{t-2} = x_{t-1} = y_{t-1} = y_{t-2}$ , then  $x_t = x_{t-1} - \delta$ .

(c) Else,  $x_t = x_{t-1}$ .

Mnemonic: SUSPICIOUS TIT FOR TAT. A less optimistic and trusting version of G-TFT.

3. COPYCAT.  $x_t = y_{t-1}$ .

Needs no explanation.

4. M-TFT

(a) If  $x_{t-1} = x_{t-2} = y_{t-1} = y_{t-2}$ , then  $x_t = x_{t-1} - \delta$ .

(b) If  $x_{t-1} = x_{t-2} \neq y_{t-1} = y_{t-2}$ , then  $x_t = x_{t-1} + 0.5 \cdot (y_{t-1} - x_{t-1})$

(c) Else:

Update the step size  $s_t$ :

$$s_t = s_{t-1} + \beta \cdot [x_{t-1} - x_{t-2}] / [y_{t-1} - y_{t-2}]$$

Then update  $x_t$ :

$$x_t = x_{t-1} + s_t \cdot [y_{t-1} - y_{t-2}]$$

where  $\beta$  is a positive parameter less than 1 (we used 0.9).

Mnemonic: MURPHY'S TIT FOR TAT. Another cautious form of TIT FOR TAT.

Table 3 records the results of pairwise play of these five policies. Note in particular the results for the BESTRESPONSE policy when played against itself. As proved in general by Cournot, when two players use the BESTRESPONSE strategy they end up at the Cournot equilibrium. This is the strategy sanctioned by SARG; it plays the role in the Cournot game that defection plays in Prisoner's Dilemma.

	G-TFT	BESTRESPONSE	S-TFT	COPYCAT	M-TFT
G-TFT	(36, 36)	(33.138, 28.165)	(36, 36)	(36, 36)	(36, 36)
BESTRESPONSE	(28.165, 33.138)	(32, 32)	(32, 32)	(32, 32)	(32.199, 31.899)
S-TFT	(36, 36)	(32, 32)	(36, 36)	(36, 36)	(36, 36)
COPYCAT	(36, 36)	(32, 32)	(36, 36)	(23.166, 23.166)	(36, 36)
M-TFT	(36, 36)	(31.899, 32.199)	(36, 36)	(36, 36)	(36, 36)

Table 3: Average over 100 runs of average profit realized during the last 100 of 10,000 plays

With all five policies in the consideration sets of the two duopolists, their average profits usually exceed 35 each (recall: 32 each is the Cournot/Nash equilibrium; 36 each is the split of the monopolist's position).<sup>7</sup> Figure 9 shows the running average profits during a typical run of 100,000 iterations of the game. It is evident that the policy-learning agents succeed on the whole in colluding tacitly. This raises the question of why Holt's human subjects were unable to collude tacitly and realize profits closer to the monopolist's. Perhaps the agent results are somehow an artifact. Perhaps, but this has to be resolved with further investigation. Note, however, that for real agents incentives and experience matter. Holt's subjects were not strongly incented and they accumulated only modest experience. One has to ask how, say, two electric power companies would fare repeatedly playing in a daily spot market for electricity, with millions of dollars at stake. At the very least, the agent simulations demonstrate that if the firms can limit their consideration set of policies in this way and if they engage in learning in policy space, then it is quite possible to arrive at tacit collusion.

Returning briefly to the  $2 \times 2$  game, we can gain some analytic insight into learning in policy space. Inevitably some simplification is required. In step 2 of policy-space learning (see Figure 7), the agent picks a length of play,  $k$ . In the simulation results presented above the agents picked their  $k$ 's independently, so that policy switching points were not synchronized. Let us now assume that all agents use a common value

<sup>7</sup>These results are robust for subsets of these five policies.

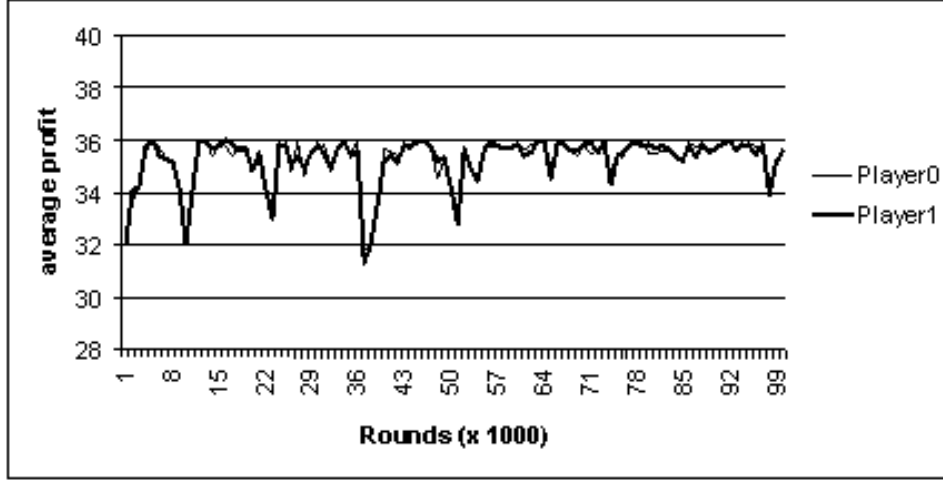


Figure 9: Average profit realized by player over 100,000 repetitions in a typical run, strategies 0–4 available;  $\epsilon$ -greedy strategy selection.

of  $k$  and consequently switch their policies in unison. The return from an episode of  $k$  plays of a policy can now be precisely specified. It will of course depend on the policy, the policy of the counter-player and the value of  $k$ . Tables 4 and 5 record those returns for symmetric  $2 \times 2$  games, coded as in Figure 1.

	000	001	010	011
000	$Pk$	$Pk$	$P + T(k - 1)$	$P + T(k - 1)$
001	$Pk$	$Pk$	$(P + T + S)k/3$	$P + T(k - 1)$
010	$P + S(k - 1)$	$(P + S + T)k/3$	$(P + R)k/2$	$(P + R + T(k - 2))$
011	$P + S(k - 1)$	$P + S + R(k - 2)$	$P + R + S(k - 2)$	$P + R(k - 1)$
100	$S + P(k - 1)$	$S + T + P(k - 2)$	$S + P + T(k - 2)$	$S + T(k - 1)$
101	$S + P(k - 1)$	$(S + T)k/2$	$(S + P + T + R)k/4$	$S + T + R(k - 2)$
110	$Sk$	$(S + R + T + P)k/4$	$Sk$	$S + R + T(k - 2)$
111	$Sk$	$S + R(k - 1)$	$Sk$	$S + R(k - 1)$

Table 4: Payoffs to the Row player from play of  $k$  rounds for the indicated policy pairs

Consider Prisoner's Dilemma again. Recall that in PD we require that  $T > R > P > S$ . In Table 5 under the column for TIT FOR TAT (101) notice that the return  $Rk$ , obtained by policies 101 and 111 (ALLC) is strictly larger than the returns for any of the other strategies (assuming  $k$  is sufficiently large; a weak requirement). Suppose that each agent plays independently 101 with probability  $1 - \epsilon$  and the other strategies with probability  $\frac{\epsilon}{7}$ . With a bit of algebra, which I shall spare the reader, it can be seen that for large (nearly all) ranges of values of  $T$ ,  $R$ ,  $P$ , and  $S$  the data from the episode will in expectation confirm strategy 101 for each player.<sup>8</sup> In short, tacit collusion is achieved because 101 versus 101 is a stochastic equilibrium. The policy-space learning regime of play (with  $\epsilon$ -greedy policy selection) transforms the game, creating an equilibrium that affords the players tacit collusion.

This can be seen as well for the important  $2 \times 2$  symmetric game of Stag Hunt [36], characterized by  $R > T > P > S$ . Stag Hunt as a stage game has two Nash equilibria in pure strategies,  $(R, R)$  and  $(P, P)$ ,

<sup>8</sup>Not in all cases because 111 ALLC does better against 001 SUSPICIOUS TIT FOR TAT than does 101.

	100	101	110	111
000	$T + P(k - 1)$	$T + P(k - 1)$	$Tk$	$Tk$
001	$T + S + P(k - 2)$	$(T + S)k/2$	$(T + R + S + P)k/4$	$T + R(k - 1)$
010	$T + P + S(k - 2)$	$(T + P + S + R)k/4$	$Tk$	$Tk$
011	$T + S(k - 1)$	$T + S + R(k - 2)$	$T + R + S(k - 2)$	$T + R(k - 1)$
100	$R + P(k - 1)$	$R + T + P(k - 2)$	$R + P + T(k - 2)$	$R + T(k - 1)$
101	$R + S + P(k - 2)$	$Rk$	$(R + S + P + T)k/4$	$Rk$
110	$R + P + S(k - 2)$	$(R + T + P + S)k/4$	$(R + P)k/2$	$R + T(k - 1)$
111	$R + S(k - 1)$	$Rk$	$R + S(k - 1)$	$Rk$

Table 5: Payoffs to the Row player from play of  $k$  rounds for the indicated policy pairs

and one Pareto optimal outcome,  $(R, R)$ . Game theorists have tended to focus on  $(P, P)$  as the predicted outcome because playing  $D$  is less risky for each of the players.  $(P, P)$  is said to be the *risk dominant* equilibrium outcome. This may be plausible for the single-shot game (played once), but when the game is repeated why should it be a norm of rationality to minimize risk, regardless of return? Tables 4 and 5 are again helpful. TIT FOR TAT versus TIT FOR TAT is even more favorable and stable in Stag Hunt than it is in Prisoner’s Dilemma, producing  $(R, R)$  on most plays. On the other hand, ALLD (000) versus ALLD is *not* stable (for wide ranges of values on the payoffs) and typically falls in favor of TIT FOR TAT. Again, this results from the transformation induced by the policy-space learning regime, which is robust: simulations under relaxed conditions ( $k$  chosen independently, etc.) lead uniformly to mutual coöperation in Stag Hunt.

### 3 Summary and Discussion

Before commenting on these findings, let me summarize the argument, or story.

1. Standard game-theoretic rationality (SARG) is attended by various paradoxes and problems.
2. With repeated games, and the Folk Theorem applying, SARG fails to make discriminating predictions. Predictions from single-shot models are unreliable.
3. The SARG account of finitely-repeated Prisoner’s Dilemma (predicting universal defection) is descriptively inaccurate and, under some conditions at least, contrary to sensible notions of rationality.
4. Failure to recognize and properly accommodate risk-return tradeoffs underlies the Surprise Exam paradox, the insistence on the rationality of universal defection in Definitely Repeated Prisoner’s Dilemma, and much interpretation of the light-guessing experiments.
5. An alternative form of rationality, which could be called an *exploring rationality* could recognize (the obvious) risk–return (aka: exploration–exploitation) tradeoffs faced by agents and, in doing so, could avoid the ambient paradoxes in a principled and empirically testable way.
6. Reinforcement learning in the state-space sense (Q-learning, e.g.) is a plausible and well-motivated form of minimal exploring rationality. It yields interesting results for games.
7. Reinforcement learning in the policy-space sense (introduced here for games) is a plausible, more powerful (and still testable) form of (potentially more than) minimal rationality. Results to date already suggest reëvaluation of public policy judgments regarding opportunities for tacit collusion in markets.

8. Analysis of simplified forms of learning in policy space explain and predict tacit collusion in repeated games. The policy-space learning regime transforms the game so that collusive outcomes (more generally, nearly Pareto-optimal outcomes) are realized as equilibria. Simulation results indicate that the findings of analysis in the simplified case are robust.

In concluding, I wish to offer three comments briefly. First, SARG, including utility theory, is a rigorously-defined, indeed axiomatized, theory. That is no small part of its attraction. More important, however, is that fact that SARG embodies or assumes certain intuitively attractive maxims of rationality. These include transitivity of preference, independence of irrelevant alternatives, the principle of dominance in choice, and so forth. Exploring rationality, in the form of state-space learning and policy-space learning described here, is not axiomatized. It is, however, rigorously specified in particular instances as algorithms executable on computer.<sup>9</sup> Undecidability and computational intractability do not attend the methods described here. Moreover, as I have tried to emphasize, the maxim of prudently attending to risk-return trade-offs is also rationally attractive. When games are repeated it may well trump the principle of dominance. Why, in say RPD, should one pick a dominated strategy in the stage game? To induce coöperation by the counter-player. Why do we think this might work? Because under a plausible regime of play coöperation is reliably achieved by such measures.

Second, a number of themes in the philosophical literature are intriguingly in accord with the, rather micro, views of rationality expressed here. As Robert Nozick has noted, “One way to understand the rationality of a belief is simply to regard it as the result of a certain type of process...” [27, page 80]. Philosophers with reliabilist views of rationality and justification may find support in the demonstration that policy-space learning is so effective in difficult strategic contexts. Philosophers have stressed the role of principles in reasoning, rationality, and moral judgment. Nozick, for example, celebrates principles as bulwarks to temptation [27]. They are what we use now to avoid doing what we would regret later. In a sense, I have been dwelling on the other side of that coin. Principles are like (are generalizations of?) policies, as I have used the term. In learning policies we take risks with an eye to finding better returns; we trade off exploitation (immediate reward) and exploration (deferred reward, in hopes of future return). I note that policies succeed or fail in accordance with the returns they generate. Policies are free—in principle as it were—to be biased, to make metaphysical or political presumptions and generally to employ pragmatic devices. This coheres with recent pluralistic accounts of science, e.g., [38].

Finally, all of this is admittedly, in fact deliberately, speculative. I have tried to sketch a case for the concept of an exploring form of rationality in strategic contexts. The concept itself needs to be explored, clarified, and refined. Further analysis and simulation need to be undertaken, experimental data reconsidered, paradoxes and anomalies revisited. The required effort is enormous. My purpose has been to support the belief that it is likely worthwhile.

## Acknowledgements

Thanks to James D. Laing, Frederic H. Murphy, David H. Wood, and DJ Wu for formative advice and discussions. Ming Lu and Ann Kuo wrote supporting simulation code wisely and well. I wish to thank audiences at Carnegie Mellon University and the Georgia Institute of Technology for listening and responding to earlier versions of these ideas. The work was supported in part by NSF grant number SES-9709548.

---

<sup>9</sup>Epstein and Axtell make an eloquent and persuasive case for constructive, computationally-based explanation in the social sciences generally [8].

## References

- [1] Christina Bicchieri. *Rationality and Coordination*. Cambridge University Press, New York, NY, 1993.
- [2] Ken Binmore. *Fun and Games: A Text on Game Theory*. D.H. Heath and Company, Lexington, MA, 1992.
- [3] Luc Bovens. The backward induction argument for the finite iterated prisoner’s dilemma and the surprise exam paradox. *Analysis*, 57(3):179–86, 1997.
- [4] Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Russell Sage Foundation and Princeton University Press, New York, NY and Princeton, NJ, 2003.
- [5] Timothy Y. Chow. The surprise examination or unexpected hanging paradox. *The American Mathematical Monthly*, pages 41–51, 1998.
- [6] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, Menlo Park, CA, 1998. AAAI Press/MIT Press.
- [7] Andrew M. Colman. Rationality assumptions of game theory and the backward induction paradox. In Mike Oaksford and Nick Chater, editors, *Rational Models of Cognition*, pages 353–371. Oxford University Press, Oxford, UK, 1998.
- [8] Joshua M. Epstein and Robert Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. The MIT Press, Cambridge, MA, 1996.
- [9] Ido Erev and Alvin E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, 88(4):848–881, 1998.
- [10] Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, Cambridge, MA, 1991.
- [11] Charles R. Gallistel. *The Organization of Learning*. The MIT Press, Cambridge, MA, 1990.
- [12] Herbert Gintis. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. Princeton University Press, Princeton, NJ, 2000.
- [13] Patrick Grim, Gary Mar, and Paul St. Denis. *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. The MIT Press, Cambridge, MA, 1998.
- [14] Ned Hall. How to set a surprise exam. *Mind*, 108:647–703, 1999.
- [15] Charles A. Holt. An experimental test of the consistent-conjectures hypothesis. *The American Economic Review*, 75(3):314–325, 1985.
- [16] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Fifteenth International Conference on Machine Learning*, pages 242–250, July 1998.
- [17] L.G. Humphreys. Acquisition and extinction of verbal expressions in a situation analogous to conditioning. *Journal of Experimental Psychology*, 25:294–301, 1939.
- [18] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

- [19] John H. Kagel and Alvin E. Roth, editors. *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ, 1995.
- [20] Steven O. Kimbrough and Ming Lu. A note on Q-learning in the Cournot game. In *WeB 2003: Proceedings of the Second Workshop in e-Business*, Seattle, WA, December 13-14, 2003. Available at <http://opim-sun.wharton.upenn.edu/~sok/sokpapers/2004/cournot-rl-note-final.doc>.
- [21] Steven O. Kimbrough and Ming Lu. Simple reinforcement learning agents: Pareto beats nash in an algorithmic game theory study. *Information Systems and e-Business*, forthcoming 2004.
- [22] Steven O. Kimbrough, Ming Lu, and Frederic Murphy. Learning and tacit collusion by artificial agents in cournot duopoly games. Working paper, University of Pennsylvania, Philadelphia, PA, January 2004.
- [23] Steven O. Kimbrough, Ming Lu, and Soofi M. Safavi. Exploring a financial product model with a two-population genetic algorithm. In *CEC-2004: Congress on Evolutionary Computation*, Portland, OR, June 2004.
- [24] David M. Kreps, P. Milgrom, J. Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoner's dilemma. *Journal of Economic Theory*, 27:245–52, 1982.
- [25] R. Duncan Luce and Howard Raiffa. *Games and Decisions*. John Wiley, New York, NY, 1957. Reprinted by Dover Books, 1989.
- [26] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, New York, NY, 1982.
- [27] Robert Nozick. *The Nature of Rationality*. Princeton University Press, Princeton, NJ, 1993.
- [28] Lynne Ofshe and Richard Ofshe. *Utility and Choice in Social Interaction*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1970.
- [29] Nicholas Rescher. *Paradoxes: Their Roots, Range, and Resolution*. Open Court, La Salle, IL, 2001.
- [30] Alvin E. Roth and Ido Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212, 1995.
- [31] T. Sandholm and R. Crites. Multiagent reinforcement learning in iterated prisoner's dilemma. *Biosystems*, 37:147–166, 1995. Special Issue on the Prisoner's Dilemma.
- [32] Frederic Schick. Surprise, self-knowledge, and commonality. PDF on the World Wide Web, Accessed 2003-02-22. URL: <http://www.lucs.lu.se/spinning/categories/decision/Schick/index.html>.
- [33] Amartya Sen. *Rationality and Freedom*, chapter Introduction: Rationality and Freedom, pages 3–64. Harvard University Press, Cambridge, MA, 2002).
- [34] Martin Shubik. *Game Theory in the Social Sciences*. The MIT Press, Cambridge, MA, 1982.
- [35] B. F. Skinner. Selection by consequences. *Science*, 213:501–504, 1981.
- [36] Brian Skyrms. The stag hunt. *Proceeding and Addresses of the American Philosophical Association*, 2001.
- [37] Elliott Sober. To give a surprise exam, use game theory. *Synthese*, 115:355–373, 1998.

[38] Miriam Solomon. *Social Empiricism*. The MIT Press, Cambridge, MA, 2001.

[39] Richar S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.