

Draft:  
On Concepts of Rationality in Games

Steven O. Kimbrough  
University of Pennsylvania  
Jon M. Huntsman Hall  
3730 Walnut Street  
Suite 500, Room 565  
Philadelphia, Pa 19104  
kimbrough (à) wharton.upenn.edu

Robert L. Axtell  
Senior Fellow  
Center on Social and Economic Dynamics  
The Brookings Institution  
1775 Massachusetts Avenue, NW  
Washington, DC 20036  
raxtell (à) brookings.edu

January 6, 2006

**Abstract**

Models in the classical theory of games and in neoclassical economics normally assume rationality in the sense that agents have complete and transitive preferences. The paper labels this *fundamental rationality* and distinguishes two other sorts of rationality pertinent to the study of strategic interaction: *individual economic rationality* (IER) and *effective rationality*. IER is, we observe, characteristic of the classical theory of games and neoclassical economics. Philosophers and others have discussed fundamental rationality (and its concomitant concept, the Nash equilibrium), which has a number of problems, limitations, and anomalies. This paper focuses instead on criticisms of IER and develops a critique of it. Our main complaint is that in general, and in very many cases of interest, there is no realistic, effective procedure by which agents can realize IER. As an alternative concept for explanation and prediction in strategic contexts we sketch a “type B” game setup for which effective procedures of play are available. We call this *effective rationality*. The paper offers suggestions for assessing effective rationality in type B games, with the aim of indicating something of how game theory (and economics) might be pursued eschewing IER as other than a benchmark or mere idealization.

# Contents

<b>1</b>	<b>Introduction and Background</b>	<b>1</b>
<b>2</b>	<b>Example Games</b>	<b>2</b>
<b>3</b>	<b>Fundamental Rationality</b>	<b>4</b>
<b>4</b>	<b>Type A Setup and Analysis</b>	<b>4</b>
<b>5</b>	<b>Type B Setup and Analysis</b>	<b>9</b>
<b>6</b>	<b>Discussion: Accessibility &amp; Games</b>	<b>10</b>
<b>7</b>	<b>Conclusion</b>	<b>13</b>

# 1 Introduction and Background

Interest in rationality, justification, warranted assertability, and related concepts is as old as philosophy. The interest has been sustained since the beginning of philosophy and remains vital in the present day (e.g., [5, 7, 21, 23]). There has also been sustained interest outside of philosophy. Decision theory and the decision sciences have attracted the attention of philosophers (e.g., [13, 16, 25]) and, voluminously, writers outside of philosophy (see, e.g., [14] for a textbook-level review). Rationality concepts are also at the center of game theory and economics, and indeed of all the social sciences. Here, too, philosophers have usefully weighed in (e.g., [6, 11, 12, 17, 18, 20, 22]) and economists have repaid the compliment (e.g., [18, 20, 27, 26]). A similar story can be told for the attentions of computer science and artificial intelligence to rationality and related concepts.

Given the enormous breadth of attention to rationality, it is hardly surprising to find multiple theories and even multiple senses of the word in active use. For example, the Wikipedia (<http://en.wikipedia.org/wiki/Rationality>, accessed 22 August 2005) is quite explicit in recognizing different senses of, and concepts for, the word rationality, some of which point usefully towards operational theories, e.g., “A logical argument is sometimes described as rational if it is logically valid,” and “In economics, sociology, and political science, a decision or situation is often called rational if it is in some sense optimal, and individuals or organizations are often called rational if they tend to act somehow optimally in pursuit of their goals,” and “Rationality is a central principle in artificial intelligence, where a rational agent is specifically defined as an agent which always chooses the action which maximises its expected performance, given all of the knowledge it currently possesses.”

While it may be possible to find an acceptable and useful general theory of rationality, our attention here is prefatory to that. We aim to explicate and clarify concepts of rationality that are pertinent to the theory of strategic behavior (game theory). That there should be a need to do this is perhaps surprising. After all, game theory is a widely and rigorously pursued, highly mathematicized field. Surprising or not, the fact is that game-theoretic rationality is a problematic and contested matter. The following passage from Gintis’s recent textbook on game theory is representative of the thoughts of very 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. [10, page xxiv] (emphasis in original)

If the concept of rationality as conceived by the established (“classical”) theory of games is flawed or mistaken, the consequences may be profound, for the theory of strategic behavior—in its established version or not—addresses phenomena that pervade the social and biological sciences. Less radically, one could claim that game-theoretic rationality is an ideal form, that people (and lesser beings) are empirically not ideally rational, that actual behavior only approximates ideal rationality because agents are limited and only boundedly rational, that the study of (ideal) rationality properly belongs to game theory, and that it is the business of psychology and behavioral game theory, broadly construed, to investigate the approximations to ideal rationality achieved by more limited beings. This leads to what might be called the approximation programme for the study of rationality in games.

Another tack that might be taken is to give up entirely the notion that rationality has much relevance at all to strategic behavior. This is the approach favored by Gintis and many others. Here is a representative passage.

*...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.

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. [10, pages xxv-xxvi]

Although we share the ambient disquiet with the notion of (ideal) rationality that is so central to (much of) received game theory, we are reluctant to follow Gintis and others in entirely abandoning rationality in the theory of strategic interaction. That reluctance is justified, at least supported, by the findings we report in this paper. Most of the paper is devoted to explicating three related concepts of rationality pertaining to strategic interaction (games). We conclude by offering evaluative comments on the three concepts. These comments serve to point towards fruitful areas for further investigation, philosophic and scientific. The upshot is to retain a concept of rationality as central to strategic interaction and to cast doubt on the approximation programme. The latter is *not* a challenge to behavioral game theory or to the psychological study of play in games. Instead, it offers a reinterpretation of this programme: behavior is aptly investigated as an approximation of the best available effective rationality.

## 2 Example Games

We will use four example games for the purpose of illustrating the points we wish to make. Our first game is the well-known Standard Prisoners’ Dilemma (SPD), presented in figure 1 in strategic form.

	$c_1$	$c_2$
$r_1$	(3, 3)	(0, 5)
$r_2$	(5, 0)	(1, 1)

Figure 1: Standard Prisoners' Dilemma (SPD). Player  $R$  chooses between strategies  $r_1$  and  $r_2$ . Player  $C$  chooses between  $c_1$  and  $c_2$ .

The interpretation of a game in strategic form is straightforward. There are two players,<sup>1</sup> the row player,  $R$ , and the column player,  $C$ . The row player must choose one of its available strategies, either  $r_1$  or  $r_2$  in figure 1, and column must choose one of its strategies, either  $c_1$  or  $c_2$  in the present case. When, as in the Standard Prisoners' Dilemma game, there are two players each with two strategies, we say the game is a  $2 \times 2$  game: two players, two strategies each. For games in strategic form, we stipulate that each player picks its strategy without observing the other player's choice. Similarly, the players can make no enforceable agreement about which strategies to pick. Once the strategies are picked, the payoffs to the players are determined, as shown by the cells in the figure. If, for example, row chooses  $r_2$  and column chooses  $c_1$ , then row's payoff is 5 and column's is 0. In a  $2 \times 2$  game, there are four possible outcomes— $(r_1, c_1)$ ,  $(r_2, c_1)$ ,  $(r_1, c_2)$ ,  $(r_2, c_2)$ —and each outcome has a payoff vector specifying payoffs for each of the players. The payoffs for SPD are given as the entries in figure 1. For two-player games in strategic form, the convention is that payoff vector  $(x, y)$  gives the row player  $x$  and the column player  $y$ .

Our second example game is a *constant sum* game: the total payoff is the same for every outcome. This makes the game one of pure conflict.  $R$ 's gain is  $C$ 's loss, and vice versa. Because the game happens to appear on page 90 of Federic Schick's *Making Choices* [25, page 90], we'll call this  $2 \times 3$  game Schick90. See figure 2.

	$c_1$	$c_2$	$c_3$
$r_1$	(3, 7)	(9, 1)	(1, 9)
$r_2$	(5, 5)	(7, 3)	(6, 4)
$r_3$	(4, 6)	(2, 8)	(8, 2)

Figure 2: Schick90: A game of pure conflict

Our third example game, Standard Stag Hunt (SSH), is also well-known. It is presented in figure 3.

	$c_1$	$c_2$
$r_1$	(3, 3)	(0, 2)
$r_2$	(2, 0)	(1, 1)

Figure 3: Standard Stag Hunt (SSH). Player  $R$  chooses between strategies  $r_1$  and  $r_2$ . Player  $C$  chooses between  $c_1$  and  $c_2$ .

Our fourth game is an ancient one. We'll call this version of it *One-Two-Twenty*. Two players take turns placing either 1 or 2 tokens on a table, starting from an empty table. The player placing the twentieth token on the table wins the game.

<sup>1</sup>More general formulations are possible, but are not needed for present purposes.

### 3 Fundamental Rationality

We'll proceed with an illustration to hand: Standard Prisoners' Dilemma. The game has four possible payoff vectors:  $\Omega = \{(3, 3), (0, 5), (5, 0), (1, 1)\}$ . Player  $R$ 's *payoff* for a payoff vector  $(x, y)$  is  $x$  and because  $R$  prefers higher payoffs to lower payoffs,  $R$ , let us assume, has the following *preference ordering* on  $\Omega$ :  $(5, 0) \succ_R (3, 3) \succ_R (1, 1) \succ_R (0, 5)$ . If an agent  $a$  prefers  $x$  to  $y$ , we write  $x \succ_a y$ . If the agent is indifferent between  $x$  and  $y$ , we write  $x \sim_a y$ . If the agent prefers  $x$  to  $y$  or is indifferent between  $x$  and  $y$ , we write  $x \succeq_a y$ . We drop the subscript on the relation ( $\sim, \succ, \succeq$ ) if no ambiguity results. An agent is said to be *fundamentally rational* (with respect to a set of payoff vectors  $\Omega$ ) if the agent has a preference ordering,  $\succeq$ , on  $\Omega$  such that for every  $a, b \in \Omega$ :

1. The *totality* condition obtains:

$a \succ b$  or  $b \succ a$  or  $a \sim b$ , and

2. The *transitivity* condition obtains:

If  $a \succ b$  and  $b \succ c$  then  $a \succ c$ , and if  $a \sim b$  and  $b \sim c$ , then  $a \sim c$ .

Clearly,  $R$ 's assumed preference ordering on the outcomes qualifies as fundamentally rational in this sense. For the sake of the example, we also assume that  $R$ 's counter-player,  $C$ , has a different preference ordering on the outcomes:  $(0, 5) \succ_C (3, 3) \succ_C (1, 1) \succ_C (5, 0)$ . It too is fundamentally rational.

Given a set of payoff vectors, we say that an agent's choice for or decision of  $\omega \in \Omega$  is *rational* or *consistent* with respect to the rational preference ordering  $\succeq$  on  $\Omega$  if for all  $\omega' \in \Omega$ ,  $\omega \succeq \omega'$ . In short, a choice is rational or consistent with regard to a fundamentally rational preference ordering, if there is no better choice available. As Hausman remarks

Rational individuals rank available alternatives and *choose* what they most *prefer*. [12, page 18]

Additional, stronger conditions for basic rationality and the existence of a (subjective or objective) cardinal utility function are available and often useful. (Philosophers will find [12, chapter 2] useful. Binmore [3], Fudenberg and Tirole [9], and Kreps [15] are standard texts.) Present purposes, however, do not require them. Indeed, what we are calling fundamental rationality is often simply described as rationality in this literature (above, also [24, page 19]). Conversely, there have been many challenges to fundamental rationality, arguing that its conditions are too strong (see, e.g., [22] for a review of these objections). Again, present purposes allow us merely to note the objections and continue on. Our principal focus here is not on fundamental rationality.

Here now are two approaches—type A and type B—to analyzing these games.

### 4 Type A Setup and Analysis

For type A analysis of a game (or a type A game), we need to specify the following items as constituting the setup of the game:

1. The players.

For the examples to hand there are two players,  $R$  and  $C$  (think row or red or Robert, and column or cyan or Cynthia). In general there may be any finite number of players.

2. The pure strategy sets,  $\Sigma^i$  for each player,  $i$ .

A strategy (for player  $i$ ) is a complete set of instructions for play of the game (by player  $i$ ). In the Standard Prisoners' Dilemma game,  $\Sigma^R = \{r_1, r_2\}$ , and  $\Sigma^C = \{c_1, c_2\}$ .  $\Sigma$  for a player is its set of *pure* strategies. When the game is presented in strategic form, as in figures 1, 2 and 3, the pure strategies for the row (column) player are the rows (columns) in the table. In addition to its pure strategies, each player also has *mixed strategies*. These are the probability-weighted combinations of the pure strategies. We denote mixed strategies with a tilde. For example,  $\Sigma^C$  denotes player  $C$ 's set of pure strategies and  $\tilde{\Sigma}^C$  denotes  $C$ 's mixed strategies. Since the pure strategies are a special case of probabilistic combination of pure strategies (one has a weight of 1, the others have 0),  $\tilde{\Sigma}^C$  denotes all of  $C$ 's strategies.

3. For each outcome, a payoff vector giving payoffs in that outcome for each player.

An *outcome* of a (type A) game is a strategy vector, giving the played strategy of each player.<sup>2</sup> Thus for Standard Prisoners' Dilemma, there are four possible outcomes:  $(r_1, c_1)$ ,  $(r_1, c_2)$ ,  $(r_2, c_1)$ , and  $(r_2, c_2)$ . The payoff vectors,  $\omega(\cdot)$ , for these outcomes are:  $\omega(r_1, c_1) = (3, 3)$ ,  $\omega(r_1, c_2) = (0, 5)$ ,  $\omega(r_2, c_1) = (5, 0)$ , and  $\omega(r_2, c_2) = (1, 1)$ . Our convention is that in payoff vector  $(x, y)$ ,  $R$  gets  $x$  and  $C$  gets  $y$ .

4. Rules of play for the game.

Standardly, "The rules of a game must tell us *who* can do *what* and *when* they can do it. They must also indicate who gets *how much* when the game is over." [3, page 25]. We are handling the *how much* aspect of a game separately, as the payoff vectors  $\omega$  (previous item).

In the general case of a  $2 \times 2$  game in strategic form, such as Standard Prisoners' Dilemma in figure 1, each player picks a strategy from its strategy set and this determines the outcome (and the payoffs). Players pick without observing each other's choices of strategy.

In addition to the game setup, type A games assume that the players each have fundamentally rational preference orders on the payoffs for the game.

Agents in games, however, do not get to choose payoff vectors, elements of  $\Omega$ , directly. Instead they get to choose strategies, elements of their  $\tilde{\Sigma}^i$ 's. Type A analysis of games is about rational choice among strategies, assuming all players are fundamentally rational regarding the payoff vectors,  $\Omega$ . We need to define type A rationality in terms of choice among strategies. It is natural to define this sort of rationality much as we defined fundamental rationality with respect to  $\Omega$ . The complication, of course, is that in a game the strategy choices of all the players must be taken into account. A further complication is that we must allow for play of *mixed strategies*. We need to discuss this last complication first.

Each agent has its  $\Sigma$ , a set of basic or *pure strategies*. To repeat: in the Stag Hunt game, for example,  $R$ 's pure strategies are  $r_1$  (hunt stag) and  $r_2$  (hunt hare), so we have  $\Sigma^R = \{r_1, r_2\}$ . In addition to the strategies in its  $\Sigma$ , an agent may also form a *mixed strategy* by probabilistic combination of its pure strategies. For

---

<sup>2</sup>Our terminology here is nonstandard. Usually, *outcome* is used for what we are calling the payoff vector. Also, see below for the distinction between the *play* of a game and the *outcome* of play.

the Standard Stag Hunt game, if  $R$  plays  $r_1$  with probability  $\frac{1}{2}$  and  $r_2$  with probability  $\frac{1}{2}$  we can write  $\tilde{r} = (r_1, \frac{1}{2}; r_2, \frac{1}{2})$ . I'll denote a mixed strategy by using this tilde notation, e.g.,  $\tilde{r}$  is a mixed strategy. Note that even if  $\Sigma$  is finite,  $|\tilde{\Sigma}|$ , the number of possible mixed strategies an agent can form from it is (uncountably) infinite.

Now some terminology. We need to distinguish the *play* of a game, from the *outcome* of (play of) a game, from the *payoff* resulting from the outcome of a game. The play of a game, *Play*, is the vector of strategies chosen by each player for playing the game. If every player plays a pure strategy, then the outcome of play, *Outcome*, is identical to the play. If at least one player plays a mixed strategy, however, chance must resolve the game into some pure strategy, and it is the resulting pure strategy that belongs to the outcome.

The following example illustrates this simple framework. Let the game be Standard Stag Hunt (SSH, figure 3). Suppose that  $Play = (\tilde{r}, c_2)$ . That is,  $R$  chooses to play the mixed strategy  $\tilde{r} = (r_1, \frac{1}{2}; r_2, \frac{1}{2})$ , defined above.  $C$  decides to play her pure strategy,  $c_2$ . Given  $Play$ , chance now resolves it by instantiating any mixed strategies. Let us say that the coin is flipped and  $R$ 's  $\tilde{r}$  gets resolved to  $r_1$ . Then the *Outcome* =  $(r_1, c_2)$ . Then the *Payoff* =  $(0, 2)$ , as indicated in figure 3. Note that  $R$ 's realized payoff is  $\frac{1}{2} \cdot 0 = 0$ . On average, however,  $R$ 's expected payoff (given that  $C$  plays  $c_2$ ) is  $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$ . We can denote this compactly by writing  $EPlay(R, (\tilde{r}, c_2)) = \frac{1}{2}$ . In words, the expected value,  $E$ , to  $R$  of the play  $(\tilde{r}, c_2)$  is  $\frac{1}{2}$ .

We are now in position to define rationality for type A games. Let  $\tilde{s}^i \in \tilde{\Sigma}^i$ , that is,  $\tilde{s}^i$  denotes a (possibly mixed) strategy available to agent  $i$ . Numbering the players from 1 to  $n$ , we can denote the play of a game by  $Play = (\tilde{s}^1, \tilde{s}^2, \dots, \tilde{s}^i, \dots, \tilde{s}^n)$ . We say that agent  $i$  is *individually economically rational* (IER) in the play of a game, if there is no strategy other than the one played by  $i$  that would yield a superior payoff for  $i$ , assuming the play is otherwise unchanged. Formally, given  $Play = (\tilde{s}^1, \tilde{s}^2, \dots, \tilde{s}^i, \dots, \tilde{s}^n)$ , agent  $i$  is IER if there is no  $\tilde{t}^i \in \tilde{\Sigma}^i$  such that  $EPlay(i, (\tilde{s}^1, \tilde{s}^2, \dots, \tilde{t}^i, \dots, \tilde{s}^n)) \succ EPlay(i, (\tilde{s}^1, \tilde{s}^2, \dots, \tilde{s}^i, \dots, \tilde{s}^n))$ . Put otherwise, given  $i$  played  $\tilde{s}^i$ , then  $i$  is IER if  $i$  has no better strategy, given how the other players played. Again,  $i$  is IER if the strategy  $i$  played is on average a *best response* to what the other players played.

An agent that is individually economically rational (IER) is also said to be *consistent* with its fundamental preference ordering on  $\Omega$ . This is how remarks like the following should be interpreted.

When game theorists describe players as “rational”, they mean no more than that they make choices *consistently*. [3, page 309]

Individually economically rational agents are also said to maximize or optimize, given their fundamental preferences and the choices made by the counter-players in the game. In ordinary discourse, there is a sense of “Agent  $i$  is maximizing” that is equivalent to “Agent  $i$  is attempting to maximize” or “trying to maximize.” This is *not* the sense employed in the present context. To the contrary, to say that an agent maximizes or is individually economically rational or is consistent with its fundamental preferences on  $\Omega$  *means* that the agent actually succeeds in playing a strategy that is a best strategy, given the actual play by the other players. How the agent might know which strategy to play, given that the strategy choices by the other players are hidden, is an entirely separate matter—in the type A setup—for the sake of determining whether the agent is individually economically rational or not.

For a given play, if (and only if) all of the players are individually economically rational, we say that the play is a *Nash equilibrium* and that the outcome resulting from the play is *supported by* the Nash equilibrium

play. Put equivalently, in a Nash equilibrium no player *individually* has incentive to change its chosen strategy. It is not precluded by the Nash equilibrium concept that two or more players might together change their strategies in such a way that both (or all) are better off. The concept of a Nash equilibrium is tied essentially to that of the possibilities for individuals acting alone, one at a time. Note again that individually economically rational outcomes and Nash equilibrium outcomes are defined without reference to how they may be arrived at or discovered. While analysts of a finite game may discern all of the solutions and test for individual rationality and Nash equilibrium, players in the game may or may not have sufficient means to hand to make these discoveries. Finally, we say that a *solution to a type A game setup* is a Nash equilibrium.

With these notions to hand, we can now treat the example games from the type A perspective.

A type A story (or analysis) for the Prisoners' Dilemma game is especially straightforward. There is exactly one play that is individually economically rational for both players (and hence is a Nash equilibrium):  $(r_2, c_2)$ . Note that the associated payoff vector is  $(1, 1)$  and that both players would do better if the play were  $(r_1, r_1)$ . Hence the dilemma. Type A analysis predicts the Nash equilibrium as the play of this game.

There is also an attractive line of reasoning that explains how the players might reach the Nash equilibrium by reasoning individually from their knowledge of the game. Player  $R$  has two pure strategies. The second,  $r_2$ , is said to *dominate* the first because no matter which (mixed or pure) strategy the counter-player,  $C$ , plays,  $R$  gets an outcome he (we use "she" for  $C$ ) prefers more if he plays  $r_2$  rather than  $r_1$ . So, if  $R$  is to be consistent with his fundamentally rational preference ordering on his  $\Omega$ ,  $R$  must choose to play strategy  $r_2$ . A completely analogous story applies to  $C$ . She must play  $c_2$ , her dominant strategy, if she is to be consistent with her preference ordering over her  $\Omega$ . Game solved.

Our second example game is Schick90, figure 2. Each player for this game of pure conflict has a plausibly attractive decision rule for finding a strategy: play a *maximin* strategy, a strategy that maximizes its minimum possible payoff. Taking row's perspective, playing  $r_2$  guarantees  $R$  a payoff of at least 5. Since the minimum payoff for  $r_1$  is 1 and the minimum for  $r_3$  is 2,  $r_2$  is uniquely the strategy for  $R$  that maximizes the player's minimum payoff. Similar reasoning identifies  $c_1$  as  $C$ 's maximin strategy. Play of  $(r_2, c_1)$  is in consequence individually economically rational for both players and is a—indeed the—Nash equilibrium for this game. Game solved.

Note a subtle difference between Standard Prisoners' Dilemma (SPD) and Schick90. In SPD each player is able to reason by elimination of dominated strategies and determine a uniquely attractive strategy of play completely independent of how the counter-player will play. No matter what row does, column is better off playing  $c_2$ . The situation is different in Schick90. When both players are IER each gets a payoff of 5. The reasoning and justification for  $r_2$  (and  $c_1$ ) relies on the fact that these are "safety strategies". No matter what the counter-player does, these strategies guarantee to their players a maximal minimum. No other strategy for each player is guaranteed to produce more than 5. If, however, one player knows, or has a reasonable amount of evidence, that the counter-player will not play its maximin strategy, then the player might be able to do better with a different strategy. For example, if  $C$  is certain that  $R$  will play  $r_1$ , then  $C$  should play  $c_3$  for a payoff of 9. Generally and in distinction to Prisoners' Dilemma, which strategy yields the highest payoff for a player now depends on which strategy is played by the counter-player. Reasoning by an individual player leading to a Nash equilibrium play will normally require the assumption of IER play by the counter-player(s).

The type A analysis of Stag Hunt is straightforward, but a bit problematic. Plays  $(r_1, c_1)$  and  $(r_2, c_2)$  are both Nash equilibria. There is in addition a third play, involving *mixed strategies*. If an agent plays  $x$  with probability  $p$  and  $y$  with probability  $(1 - p)$  we write this mixed strategy as  $(x, p; y, 1 - p)$ . For the Stag Hunt game in figure 3 the play in mixed strategies,  $(\tilde{r} = (r_1, \frac{1}{2}; r_2, \frac{1}{2}), \tilde{c} = (c_1, \frac{1}{2}; c_2, \frac{1}{2}))$  is the third Nash equilibrium. Players playing at this equilibrium can expect a payoff of 1.5 each.

The Stag Hunt is problematic in two ways under type A analysis. First, which play will prevail and what will be the distribution of plays when the game is surveyed across many plays? The Nash equilibrium concept by itself cannot discriminate among the three Nash equilibria. It is possible, of course, to single out one or another of the equilibria as favored in virtue of properties it has in addition to being a Nash equilibrium. Game theorists have tended to favor  $(r_2, c_2)$  because it poses the least risk of a 0 payoff to any player, but there is not general agreement on this. There is a worry, moreover, that the properties so identified will not generalize to other games with multiple equilibria. Although principled selections may be made in specific cases, the problem of multiple equilibria for type A analysis remains recalcitrant.

The second problem for type A analysis presented by the Stag Hunt game is related to the first, and is perhaps but an aspect of it. This is the problem of specifying a procedure for arriving at an equilibrium. Sticking to just the two equilibria in pure strategies,  $r_1$  is  $R$ 's strategy for one of the equilibria, while  $r_2$  is the strategy for the other. But  $R$  can't by himself pick one of the equilibria. If he prefers  $(r_1, c_1)$  he can play  $r_1$ , but if  $C$  prefers  $(r_2, c_2)$  and she plays  $c_2$  then the play is not a Nash equilibrium. Further, although  $(\tilde{r}, \tilde{c})$  is a Nash equilibrium, plays combining mixed and pure strategies are not, e.g.,  $(\tilde{r}, c_2)$ . If there is no procedure or path of reasoning by which the players coordinate on a single equilibrium, how is it that play is at an equilibrium?

The type A story for One-Two-Twenty is a bit complex, but not especially difficult. We give it only in outline. A *state of the game*,  $e = (i, n)$ , is specified by which player,  $i$ , has the next play and how many stones,  $n$ , are on the board. Since there may be  $0, 1, \dots, 20$  stones on the board and at any time it may be either player's turn, there are 42 possible states. This number is reduced once we specify who goes first, but the details are not important for us. Let us say that  $C$  goes first. It is easy to see that there is a winning strategy for  $C$ . Let the value of state  $e$  to player  $i$ ,  $V^i(e)$ , be 1 if once the game is in that state player  $i$  can be guaranteed of a win. Clearly  $V^C(C, 19) = 1 = V^C(C, 18)$ , for  $C$  can add either 1 or 2 stones to produce 20 on the board and thereby win the game. Consequently  $V^C(R, 17) = 1$ , since  $R$  can then produce only states  $(C, 18)$  and  $(C, 19)$ , which have a value to  $C$  of 1. Continuing to reason backwards in a similar manner we find that  $V^C(R, 14) = 1 = V^C(R, 11) = \dots = V^C(R, 2)$ . Consequently,  $V^C(C, 0) = 1$ .  $C$  can begin the game by placing 2 stones on the table, thereby producing state  $(R, 2)$ . If  $R$  produces state  $(C, 3)$ , then  $C$  puts 2 stones down, producing state  $(R, 5)$ ; otherwise,  $R$  produces state  $(C, 4)$  and  $C$  puts 1 stone down, again producing state  $(R, 5)$ . Play continues in this fashion until  $C$  wins the game.  $C$ 's strategy, combined with *any* play by  $R$  is a Nash equilibrium. Note that if  $C$  deviates from this strategy, say by producing state  $(R, 6)$ , then there is a strategy available to  $R$  for winning the game. In the example,  $R$  would produce  $(C, 8)$  and be in position to win the game. Thus, in One-Two-Twenty rational (type A) play as specified by the Nash equilibrium accords well with what we would atheoretically expect rational players to do. The game favors whoever goes first. That player wins if the outcome is a Nash equilibrium.

## 5 Type B Setup and Analysis

The setup for, or description of, a type B game includes the following elements:

0. The supergames.

A type B game consists of one or more supergames. Each supergame comprises many (2 or more) subgames. In a simple case, the subgame (aka: stage game) would be Standard Prisoners' Dilemma and the supergame would be 25 rounds of play of the subgame between two fixed players.

1. The players.

Every game, including type B games, has at least 2 players.

2. The policy sets,  $\Pi^i$ , for each player,  $i$ .

A policy is a strategy (complete set of instructions) for playing a subgame. Policies are defined in such a way that a player may during the course of a supergame play under one policy for part of the supergame and under another policy for a different part. For example, if the supergame consists of 25 rounds of play of Standard Prisoners' Dilemma, a player might cooperate for rounds 1-11 and defect for rounds 12-25.

A main difference between games of type A and games of type B, is that in the former we view players as choosing among strategies, while in the latter we view players as choosing (directly) among policies. They choose strategies only indirectly, as emerging from their policy choices.

3. For each outcome of every atomic subgame, a payoff vector specifying payoffs for that outcome for each (involved) player.

A subgame is atomic if all of its outcomes are associated with elements of a relevant  $\Omega$ . A nonatomic subgame may be composed of atomic subgames or may have as payoffs the rewards from playing in other subgames.

4. The adaptation regime(s) used by the players,  $\rho$  (or indexed, e.g.,  $\rho_i$  if there is more than one).

Players play by following policies for play during a supergame. A particular action in a particular subgame is determined by the policy in effect for the player in question. The player's adaptation regime determines which policy will be in effect at any given time. Necessary for games of type B, a game description of type A lacks this element entirely.

5. Rules of play for the (super)game(s).

As in type A game descriptions, the rules of play govern the sequencing and other conditions under which the players make their decisions.

The key differences, then, between a type B game setup (description, model) and one of type A are that (i) type B games are always supergames, consisting of multiple subgames, (ii) players directly choose policies rather than strategies, with the policies in turn determining play in subgames,<sup>3</sup> and (iii) players have adaptation regimes which produce their choices of policies. The salient feature of type B game setups is that players try policies, receive feedback from play of subgames, invoke their adaptation regimes, and either

---

<sup>3</sup>Policies need not be deterministic. They may involve randomized decisions. No special notation—such as use of the tilde in type A games,  $\tilde{s}$ —will be needed.

try new policies or continue on, depending on direction from their adaption regimes. This is a process of adapting (and perhaps learning) in policy space.

Our example games can be used to illustrate type B game setups. Standard Prisoners' Dilemma (SPD) first. The supergame for this example consists of iterated play of SPD as a stage game. After each round of play the supergame halts with probability 0.02; otherwise another round is played. Players  $R$  and  $C$  have identical consideration sets of policies for play,  $\Pi = \{\text{ALWAYS DEFECT}, \text{ALWAYS COOPERATE}, \text{TIT FOR TAT}\}$ . Under the ALWAYS DEFECT policy, if the player is  $R$  he plays  $r_2$  whenever he has the policy in force and if the player is  $C$  she plays  $c_2$  whenever she has the policy in force. Similarly, they play  $r_1$  and  $c_1$  if ALWAYS COOPERATE is in force. Finally, under TIT FOR TAT, the player playing it cooperates ( $r_1$  or  $c_1$ , depending on the player) in the first play for which the policy is in force. After that, so long as the policy is used, the player mimics the play (cooperative or not) of the counter-player in the previous round of play. Players independently pick policies and play them for a number of rounds of play. Each player keeps track of the performance of, the returns from, play with its policies, and uses this information when selecting a new policy for play. What happens in the supergame—e.g., whether one policy comes to dominate play—depends on the details of the adaptation regimes involved, the policies, and perhaps on chance events. It is, however, discoverable by computation, if not by mathematical analysis.

Our example model for Standard Stag Hunt is a *gridscape* model. Players are arrayed on a regular network. Think of a chessboard, but one that “wraps” around so that each cell has 8 neighbors. This is a model for a simple society. Agents are either stag hunters or hare hunters. Agents in parallel play all of their 8 neighbors and count up their points. Each agent then looks to its neighbors and sees if any have obtained more points by using a different strategy. If so, the agent adopts the strategy of a neighbor whose achieved points is highest and play continues. As in the SPD example, what happens depends on the details of the setup, but is discoverable.

For either One-Two-Twenty or Schick90 (both of which are games of pure opposition) imagine again that two players  $R$  and  $C$  play the game repeatedly and that in the case of One-Two-Twenty  $C$  is given the first move each time the game is played. Neither  $R$  nor  $C$ , let us assume, have the capacity to analyze the game as we did above and to figure out an optimal strategy. Are there ways that simpler agents (simpler than us) might figure out the game? Clearly yes. An agent with a bit of memory and an elementary ability to reason backwards could surely learn by experience in iterated play of the game and achieve optimal, or at least high quality, play. In the case of One-Two-Twenty, the value of later states could be learned first by trial and error, followed by the value of neighboring earlier states. Eventually the entire game could be learned. How well and how quickly players might learn in this way depends greatly on the details, which are again open to investigation. We note that the general approach has been validated by the success of computer programs in learning to play difficult games. Fogel's discussion of his checker programs [8] is very readable and noteworthy in this regard.

## 6 Discussion: Accessibility & Games

A few more terminological stipulations will facilitate the discussion.

Suppose that algorithm (or procedure or rule)  $\alpha$  accepts inputs  $\beta$  and produces  $\gamma$ . Let us then say that  $\gamma$  is *accessible from  $\beta$  via  $\alpha$* . If needed, it is possible to give a more formal, rigorous definition of accessibility,

but that requirement is not to hand.<sup>4</sup> Examples can carry the burden of clarification: (1)  $\gamma$  is  $\neg P$ ,  $\beta$  is  $P \rightarrow Q, \neg Q$ , and  $\alpha$  is *modus tollens*. (2)  $\gamma$  is 27,  $\beta$  is  $x = 3$ , and  $\alpha = x^3$ . We deliberately leave open what sorts of things  $\gamma$  and  $\beta$  may be (e.g., numbers, statements, formulas, etc.).  $\alpha$  is correspondingly open; it is any procedure—deterministic or randomized—for producing  $\gamma$  from  $\beta$ .  $\gamma$  itself may be definite—*Rain tomorrow*—or probabilistic—*Chance of rain tomorrow greater than 0.7*. Let us also say that  $\gamma$  is *accessible from  $\beta$*  if there is some (not necessarily specified) algorithm  $\alpha$  such that  $\gamma$  is accessible from  $\beta$  via  $\alpha$ .

Under what conditions is something *not* accessible? Let us say that  $\gamma$  is *practicably accessible from  $\beta$*  via  $\alpha$  if  $\gamma$  is accessible without undue cost or delay from  $\beta$  via  $\alpha$ . Quite clearly, much that is accessible (in principle) is not practicably accessible.

One way for absolute inaccessibility to occur is when applying  $\alpha$  to  $\beta$  yields more than one result. Then we have to say that any single result is not accessible from  $\beta$  via  $\alpha$ . Equations with more than one root (solution) will perhaps be the most familiar example. Suppose that  $y$ , the position of an object, equals  $t^2$  in minutes, with noon today set as  $t = 0$ . At what time is  $y$  at 16? The equation allows two answers: 4 minutes before noon and 4 minutes after noon. So here we can say that  $\gamma$  is  *$t = 4$  minutes before noon or  $t = 4$  minutes after noon* is accessible from the equation via mathematical solution. It would be incorrect to say that  $\gamma$  is  *$t = 4$  minutes before noon* is accessible from the equation via mathematical solution, even though  $\gamma = t$  is *4 minutes before noon* is consistent with the equation via mathematical solution. Accessibility resembles visibility under a microscope. A set (organelle) may be accessible (visible under the microscope) without its elements (components) being accessible (visible).

Back now to games. Let  $\beta$  be a type A game with the assumption of fundamental rationality for all players. Let  $\alpha$  be a procedure that selects player  $R$ 's strategy in every Nash equilibrium. Then in general neither  $\gamma$  as  *$R$  plays strategy  $r_1$*  nor  $\gamma$  as  *$R$  does not play strategy  $r_1$*  is accessible (from  $\beta$  via  $\alpha$ ). The Stag Hunt game illustrates the point. In some equilibria  $R$  plays  $r_1$ , in some  $r_2$ , and in some a mixture. The Nash equilibrium concept does not reveal which equilibrium will obtain. Because player  $R$  has more than one strategy involved in the equilibria, only a set (larger than 1) is accessible.

Practicable accessibility is also an issue for type A games. Under Zermelo's theorem (<http://www.ams.org/featurecolumn/archive/games3.html>, accessed 17 December 2005) any (type A) game that is finite, played by two players under perfect information (each player knows all the moves so far from the other player), and is strictly competitive is such that either the first player can force a win or a draw or the second player can force a win or a draw. Finding such an equilibrium, however, is another matter. Chess, checkers and many other board games qualify under the theorem, but their sizes and complexities preclude completion of the analysis. The equilibria of such a game are practicably inaccessible via the Nash equilibrium procedure (here, Zermelo's backwards induction procedure).

Some but not all type A games are not practicably accessible because of computational complexity. Some but not all are not accessible at all because of multiple equilibria. The situation is quite different for type B games. By hypothesis, the players each have adaptation regimes that select policies for play and these produce outcomes in atomic subgames. Convergence to a stable outcome or even stable distribution of outcomes may or may not occur, let alone convergence to an equilibrium. Because conditions of play may be stochastic, different runs of play may produce different results. We may think of  $\gamma$  in the context of type B games ( $\beta$ ) as a stream of outcomes and  $\alpha$  as the adaptation regimes assigned to the several players.

---

<sup>4</sup>Morton in [19] describes a stronger concept in the same spirit: manageability. We would subscribe to the requirement for a stronger concept, such as described by Morton, but that is not required for the argument to hand.

Conceived this way, something, some  $\gamma$ , is always accessible, practicably accessible, in type B games. They are designed to be that way.

At bottom, the distinction between type A and type B games is largely one of perspective or stance, of attitude we bring to the subject. In a type A model we are mainly interested in the Nash equilibria and are less concerned with accessibility issues.<sup>5</sup> The theory for type A games may be called *equilibrium game theory*. In a type B model we endow the players with policy spaces and adaptation regimes, which they deploy in conducting the game and from which their strategies emerge. We may call the theory for these games *effective game theory*. It investigates how game results emerge from the interplay of policy spaces and adaptation regimes via effective procedures (whether successful or not). Whatever rationality type B agents have is a *effective rationality*; play,  $\gamma$ , is produced by a policy in force,  $\alpha$ , typically relying on a history of play,  $\beta$ .<sup>6</sup>

The perspective of effect game theory is, we submit, likely to be apt under a number of conditions, including these:

1. When fundamental rationality cannot be assumed.

Note that economic rationality is easily violated if agents have to rely on realistic perceptual mechanisms. Letting  $a \sim b$  be interpreted as “ $a$  is not perceptually distinguishable from  $b$ ,” we can easily have  $a \sim b$ ,  $b \sim c$  and  $a \succ c$ . The relation  $a \succ c$  might be used for such perceptual modalities as “is at least as hot as” and perhaps even “is worth at least as much as.”

Effective game theory (type B) representations may or may not assume fundamental rationality. In either case, the games will play out, agents will adapt and an evaluatable pattern of play will result.

2. When a type B game is a natural model of the system under examination, regardless of whether a type A representation is useful.

The one-shot Prisoners’ Dilemma is more naturally modeled as a type A game, while the indefinitely repeated version is usefully addressed from the perspective of effective game theory.

3. If at least some players are epistemically limited (nonhuman animals surely count) and cannot plausibly be assumed to be individually economically rational.

4. If the underlying game is too complex for meaningful type A (equilibrium) analysis.

Equilibrium analysis for tic-tac-toe and One-Two-Twenty, constructivist analysis for chess and checkers.

5. If the underlying game has multiple equilibria after excluding implausible equilibria.

The Folk Theorem (a real theorem; see standard texts such as [3, 15] for proofs and discussion) tells us that in indefinitely repeated games the equilibria proliferate.

6. If a procedure of play is sought by which the players can operate and with which explanations, predictions, and interventions may be made.

---

<sup>5</sup>By “Nash equilibria” we mean Nash equilibria or refinements thereof. Also, Brams’s theory of moves [4] is an interesting ... move, but is one we postpone to another venue.

<sup>6</sup>These terms—*effective rationality* and *effective game theory*—are neologisms. Perhaps *constructive rationality* would do as well. The name used is not very important, except to avoid confusing the reader. We defer the question of what relationship effective rationality has to similar concepts, notably Herbert Simon’s *procedural rationality*, which he distinguishes from what he calls *substantive rationality* [28], and Richard Thaler’s psychologized *quasi-rationality* [32].

There are special cases, such as when elimination of dominated alternatives produces a unique outcome, in which procedures are available for producing equilibria in type A games. With complexity or repetition, however, such cases are rare.

In an effective (type B) game setup, we model players as agents with specific, albeit quite limited, powers of thought. Agents have a cognitive apparatus,  $(\Pi, \rho)$ , consisting of a consideration set of policies,  $\Pi$ , and an adaptation regime,  $\rho$ . Under equilibrium analysis there is only one criterion of evaluation: What are the individually rational outcomes? Such parsimony is not available to us in the case of effective games. Instead, the list of criteria is best left open. Here is a list that serves to begin a fruitful discussion.

1. Performance against self.

Does the apparatus do well against itself? If all agents use the apparatus, do the agents as a group prosper relatively well in the ambient environment?

2. Performance against others.

Does the apparatus do well playing against other regimes that do well against themselves?

3. Exploitability.

Is the apparatus catastrophically exploitable? Does it have weaknesses that may be discovered by another apparatus?

4. Robustness.

Is the apparatus robust under perturbations of its parameters? Does the apparatus perform well against a field of others?

5. Learnability.

Can the apparatus be parameterized in such a way that an agent can (easily) learn profitable, well-performing settings?

6. Computational cost.

Is the apparatus computationally tractable? Is it simple or does it require excessive computational resources from the agent?

7. Informational requirements.

Does the apparatus rely on plausibly available information? Or does it require information not likely to be available in the actual system being modeled?

## 7 Conclusion

What we have described and labeled as fundamental rationality is in fact foundational for classical (what we have called type A or equilibrium) game theory. The theory presumes it and the literature typically speaks of it as simply rationality itself. Questions have been raised and responses have been made regarding fundamental rationality (see, e.g., [12, chapter 1]). That discussion continues. Our main observation is that

while individual economic rationality (IER) presumes fundamental rationality, effective rationality need not. Agents that violate fundamental rationality may still engage in strategic interaction and explore the returns from a collection of policies in a policy space. Whether we judge them to be rational or not will depend on our theory of effective rationality and on how these agents perform in relevant contexts.

What we have described and labeled as individual economic rationality is also central for type A or equilibrium game theory. Existing theory largely presumes it. We frame game theory's core concept—the Nash equilibrium—as definable in terms of IER. The play of a game is a Nash equilibrium if and only if every player is IER. This framing serves the useful purpose of drawing our attention to IER directly.

Having introduced informally the concept of accessibility (in the present context), we noted that in very many games (those with multiple Nash equilibria) there is no effective means by which IER is universally accessible.<sup>7</sup> Further, many important games are so complex (chess, checkers, etc.) that it is utterly unrealistic to think that computational resources might be deployed to find IER strategies for the players. These are, we think, fundamental impediments to the use of game theory as an explanatory device for real agents, human or not. While it is common to worry about the heroic epistemic powers attributed to players by game theory, our objections are more basic. If theory requires enormous cognitive powers in its models, then the models may well approximate cognitively weaker agents. Approximating the inaccessible is an entirely different matter.

If, for the great majority of interesting games, equilibrium rationality (universal IER) is an implausible model, must we follow Gintis's move and give up on rationality? The third rationality concept we described and labeled—effective rationality—offers an alternative. Given that we want to have a scientific theory—supporting description, explanation, prediction, and intervention—of strategic interaction by real agents, the theory will be incomplete without an account of how the agents choose their plays. Equilibrium game theory is largely incomplete in this sense. The gist of the argument we are making is that equilibrium game theory is fundamentally uncompletable in the same sense, or at least that there is good reason to think so.<sup>8</sup>

What effective game theory offers is the prospect of descriptively accurate accounts of play by agents of all kinds in strategic contexts. In addition, broadly procedural or effectively constructive adaptation regimes, not present in any known agent, may be investigated with the aim of prescribing more optimal behavior. If players can't actually optimize, effective models can help them try to optimize. We have been able only to hint at an adequate development of the concept of effective rationality. Much work of this sort, with very encouraging results, has been and is being done. All of the usual disciplines, including philosophy (e.g., [1, 2, 29, 31, 30]) are well represented. It is this kind of work that must ultimately articulate and validate the concept of effective rationality.

## References

- [1] Christina Bicchieri. *Rationality and Coordination*. Cambridge University Press, New York, NY, 1993.

---

<sup>7</sup>Rationalizability, as an alternative to IER, we think does not answer to this problem either, since in general the number rationalizable outcomes is large [3, page 484] and [9, pages 48–50]. This, along with correlated equilibria and other refinements of the Nash equilibrium/IER, require discussion that is beyond the scope of this short paper.

<sup>8</sup>There is the possibility that concepts could be found for uniquely selecting an equilibrium in the majority of games. That prospect has been on the table for 20 years. It does not appear credible at present. Also, it does very little to address the problem of practicable accessibility.

- [2] Christina Bicchieri, Richard Jeffrey, and Brian Skyrms, editors. *The Dynamics of Norms*. Cambridge University Press, New York, NY, 1997.
- [3] Ken Binmore. *Fun and Games: A Text on Game Theory*. D.H. Heath and Company, Lexington, MA, 1992.
- [4] Steven J. Brams. *Theory of Moves*. Cambridge University Press, Cambridge, United Kingdom, 1994.
- [5] Harold I. Brown. *Rationality*. Routledge, New York, NY, 1988.
- [6] Nancy Cartwright. The limits of exact science, from economics to physics. *Perspectives on Science*, 7(3):318–336, Fall 1999.
- [7] Donald Davidson. *Problems of Rationality*. Oxford University Press, New York, NY, 2004.
- [8] David B. Fogel. *Blondie24: Playing at the Edge of AI*. Morgan Kaufmann, San Francisco, CA, 2002.
- [9] Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, Cambridge, MA, 1991.
- [10] Herbert Gintis. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. Princeton University Press, Princeton, NJ, 2000.
- [11] Gilbert Harman. *Reasoning, Meaning and Mind*. Clarendon Press, Oxford, United Kingdom, 1999.
- [12] Daniel M. Hausman. *The Inexact and Separate Science of Economics*. Cambridge University Press, Cambridge, United Kingdom, 1992.
- [13] Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, IL, second edition, 1983.
- [14] Paul R. Kleindorfer, Howard C. Kunreuther, and Paul J.H. Schoemaker. *Decision Sciences: An Integrative Perspective*. Cambridge University Press, Cambridge, United Kingdom, 1993.
- [15] David M. Kreps. *A Course in Microeconomic Theory*. Princeton University Press, Princeton, NJ, 1990.
- [16] Isaac Levi. *Hard Choices : Decision Making under Unresolved Conflict*. Cambridge University Press, Cambridge, United Kingdom, 1990.
- [17] David Lewis. *Convention: A Philosophical Study*. Basil Blackwell, Oxford, United Kingdom, 1969/1986.
- [18] Uskali Mäki, editor. *Fact and Fiction in Economics*. Cambridge University Press, Cambridge, United Kingdom, 2002.
- [19] Adam Morton. Game theory and knowledge by simulation. In Martin Davies and Tony Stone, editors, *Mental Simulation*, pages 235–246. Blackwell, Oxford, UK, 1995.
- [20] Paul K. Moser, editor. *Rationality in Action*. Cambridge University Press, Cambridge, 1990.
- [21] Robert Nozick. *The Nature of Rationality*. Princeton University Press, Princeton, NJ, 1993.
- [22] Hilary Putnam. *The Collapse of the Fact/Value Dichotomy and other Essays*, chapter On the Rationality of Preferences, pages 79–95. Harvard University Press, Cambridge, MA, 2002.
- [23] Nicholas Rescher. *Rationality: A Philosophical Inquiry into the Nature and the Rationale of Reason*. Clarendon Press, Oxford, United Kingdom, 1988.

- [24] Alvin E. Roth and Marilda A. Oliveira Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press, Cambridge, United Kingdom, 1990.
- [25] Frederic Schick. *Making Choices: A Recasting of Decision Theory*. Cambridge University Press, Cambridge, United Kingdom, 1997.
- [26] Amartya Sen. *Rationality and Freedom*, chapter Introduction: Rationality and Freedom, pages 3–64. Harvard University Press, Cambridge, MA, 2002.
- [27] Amartya K. Sen. Rational fools: A critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs*, 6:317–344, 1977.
- [28] Herbert Simon. From substantive to procedural rationality. In Spiro J. Latsis, editor, *Method and appraisal in economics*. Cambridge University Press, New York, NY, 1976.
- [29] Brian Skyrms. *Evolution of the Social Contract*. Cambridge University Press, Cambridge, UK, 1996.
- [30] Brian Skyrms. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, Cambridge, UK, 2004.
- [31] Brian Skyrms and Robin Pemantle. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences*, 97(16):9340–9346, August 1, 2000.
- [32] Richard H. Thaler. *Quasi Rational Economics*. Russell Sage Foundation, New York, NY, 1991.