Game Theory: A Language of Interaction

Adam Brandenburger

NYU Stern School of Business, NYU Tandon School of Engineering, NYU Shanghai, New York University, NY 10012, U.S.A.

Version 07/28/23

1. Introduction

Game theory is a mathematical language for investigating **interactions** in many different settings. The interacting entities (called the **players**) might be humans or other animals, sometimes treated as individuals and sometimes as groups. They might be microorganisms (West et al. [2006]) or even individual genes (Dawkins [1976]). In another kind of application of game theory, the interacting entities might be processors or other machines in a communication network (Halpern and Moses [1990]).

A roadmap for the reader: Sections 6, 8, and 14 on contain material that is often not included in an introduction to game theory. The reader can work through the other sections first and then come back to these more specialized sections.

2. Origins

The origins of game theory are found in an interest which mathematicians took, in the early 20th century, in analyzing Chess and other parlor games (Leonard [2012]). These are often called **games of strategy**, to distinguish them from **games of chance** (such as tossing dice or coins). Attempts to understand games of chance had helped motivate the 17th-century mathematicians Pascal and Fermat to develop probability theory (Devlin [2008]). In the early 20th century, mathematicians became interested in games of strategy. These are situations in which players face uncertainty not about how about Nature will act (e.g., about the physical conditions that will determine whether a coin lands heads or tails) but about how other players will act. Moreover, these other players may themselves be thinking about how the first player will act. This creates a kind of 'reflexivity' in a game of strategy which, not surprisingly, has been at the heart of game theory from its earliest days.

A good date to attach to the beginning of game theory is 1928, when the mathematician John von Neumann published a paper (von Neumann [1928]) that lays out much of the basic language of game theory and analyzes a particular rule of behavior in games.

3. Game Trees

The first model of an interaction that von Neumann proposed is the **game tree** (also called the **extensive form**). This is a model specifying which player moves first and what possible **moves** are available to that player; which player moves next and what moves are available to that player; and so on. Also specified is the **information** available to each player at each of the contingencies under which that player might need to choose a move. This information is made up of any information the player has about previous moves made by other players (and by himself, if we want to allow for limited memory) and about previous moves by Nature (i.e., chance moves).

Each possible sequence of moves by players, from the move chosen by the first player to the move chosen by the last player, brings us to a possible **terminal node** of the tree. Associated with each terminal node is an **outcome** of the game. (Depending on the game, two or more different terminal nodes might yield the same outcome.) Outcomes can take many forms — an amount of money to each player, the division of a physical product, or a more abstract

outcome such as Win or Lose. The final component of a game tree is the **payoff** to each player from each possible outcome. These are a player's evaluations of the different possible outcomes, and they might well incorporate various subjective and psychological factors that influence how a player values the objectively given outcomes.

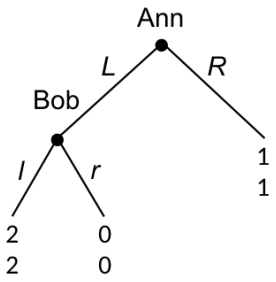

Figure 1a

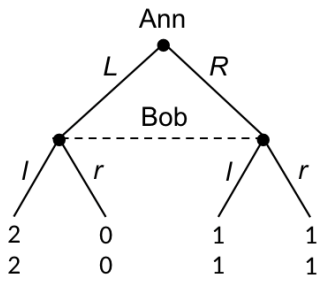

Figure 1b

Figures 1a and 1b depict two very simple game trees. Each involves two players, Ann and Bob. In each tree, Ann and Bob each get to choose between two moves: L or R for Ann, and l or r for Bob. In the tree of Figure 1b, the dotted line connecting the two nodes constitutes an **information set**, which indicates that when Bob gets to move, he does not know whether Ann previously chose L or R. Accordingly, this tree is said to exhibit **imperfect information**, while the tree of Figure 1a, where all information sets are trivial — i.e., contain only one node each — exhibits **perfect information**. We will come back to these two particular game trees, and look at other trees shortly.

4. Strategies and Game Matrices

A second key concept introduced by von Neumann [1928] is that of a player's strategy. A **strategy** for a particular player is a complete contingent plan for that player, specifying which of the available moves the player will make at each of his information sets in the tree. (Remember that information sets can be singleton nodes or sets of nodes.) The word "contingency" is important here. In Figure 1a, a strategy for Bob specifies whether he will move l or r at his (singleton node) information set, even though he will get to move only if Ann first moves L and not R.

Von Neumann used the concept of strategy to associate a **game matrix** (also called a **strategic form**) with any given game tree. To construct the matrix, we first list the set of possible strategies for each player in the given tree. Then, for each possible **profile** of strategies, one strategy per player, we trace out the induced path through the tree and arrive at a certain outcome. The resulting payoffs, one per player, yield the payoff profile associated with the chosen strategy profile. As a simple example, Figure 2 depicts the game matrix associated with the game tree of Figure 1a. Notice that in the matrix, if Ann plays R, then both players get a payoff of 1, regardless of Bob's strategy. This is because the choice R by Ann brings the game immediately to the right-most terminal node in Figure 1a. Bob's strategy does not affect the play of the game in this case.

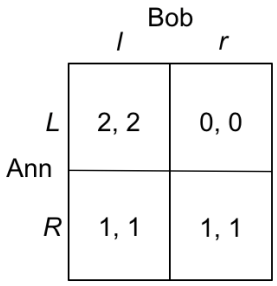

Figure 2

A second point to note is that the matrix of Figure 2 is also the game matrix associated with the game tree of Figure 1b. (Bob has two, not four, strategies in the tree of Figure 1b. He has a single information set, and a strategy for him specifies a move, l or r, at this single information set.)

For von Neumann, the game matrix was the starting point for the analysis of behavior. He proceeded by making certain assumptions about how players make decisions in a game setting and examined the consequences. We will look at von Neumann's approach and its successors a bit later.

An important point has emerged here. If two or more game trees can be mapped to the same game matrix (as we have just seen), this raises a question: Is the matrix really an adequate model of the situation being studied? Prima facie, information about which of several situations is being modeled may have been lost. How this issue is addressed brings out a very important distinction within game theory — one that has a long history and continues today.

5. A Priori and A Posteriori Reasoning in Game Theory

Some game theorists view the objective of the field as trying to pin down, through careful reasoning, the precise meaning of optimal or 'rational' behavior in a game setting. There is a natural appeal to this endeavor, which sees a game as posing a particularly deep kind of logic puzzle in which the best way for one player, Ann, to play a game depends on what she thinks is the best way for Bob to play, which, in turn, depends on what she thinks Bob thinks is best for her, and so on indefinitely. Perhaps, there is some clear way to resolve this infinite regress — a way that would then yield the 'rational' way to play a game.

The most influential concept in game theory to have arisen from such **a priori reasoning** is Nash equilibrium (Nash [1951]). We will turn to this concept and some of its descendants shortly.

Another view of game theory is that it is a language suited to specifying various assumptions about how players reason and behave in games and to analyzing the implications of these assumptions. Data from outside of game theory itself — such as data from experimental game theory, cognitive psychology, animal behavior, and cognitive neuroscience — can be used to help select the right assumptions. This is more along the lines of **a posteriori reasoning** (from observations). This chapter tries to emphasize this linguistic way of thinking about the role of game theory, but it also covers important parts of the a priori approach.

The distinction between these two approaches to game theory can be seen in their respective treatments of the observation that the two trees in Figure 1 both map to the matrix in Figure 2.

The a priori approach turns this observation into an **invariance axiom**, which requires, as part of rationality, that a player choose the same way in any two game trees that map to the same matrix. In particular, a rational player will play the same way in the two game trees of Figure 1.

The underlying point of view is that it is the concept of strategy that is basic, and, therefore, it is the strategic form or matrix that is basic. Two trees that yield the same matrix differ only in inessential presentational effects. This line of argument, developed in detail by Dalkey [1953] and Thompson [1952], became the basis of influential modern work by Kohlberg and Mertens [1986] and subsequent papers.

The a posteriori approach sees an interesting **cognitive issue**: Do players, in fact, treat the trees in Figures 1a and 1b as equivalent, or do they think or behave differently in the two cases? Here, game theory plays the role of identifying possible distinctions among interactions that might otherwise be missed. Schotter, Weigelt, and Wilson [1994] and Cooper and Van Huyck [2003] investigate this area experimentally.

6. More on the Language of Game Trees

The standard definition of a game tree, as used in game theory today, comes from Kuhn [1950, 1953], and extends von Neumann's [1928] definition somewhat. A central distinction in Kuhn's papers is between games with **perfect** or **imperfect recall**. The tree in Figure 3 has imperfect recall.

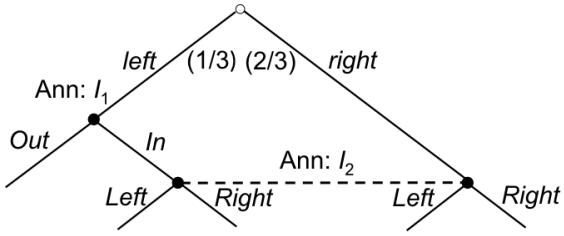

Figure 3

This tree begins with a move by Nature (indicated by the open node), with the probabilities of *left* vs. *right* given in parentheses. There is one player, Ann, who possesses the two information sets I_1 and I_2. At I_2, Ann has forgotten whether or not she put herself at this information set. The full definition of perfect vs. imperfect recall in Kuhn [1953] asks whether or not a player remembers: (i) everything that the player knew at previous information sets, and (ii) the player's own previous moves.

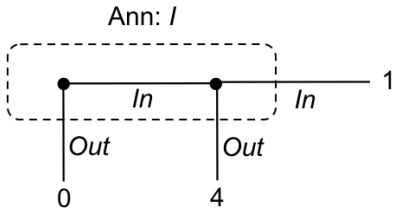

Figure 4

The language of game trees allows for additional phenomena. Extending the definition of a tree beyond Kuhn [1953], Isbell's [1957] definition allows for the tree in Figure 4 (which is from Piccione and Rubinstein [1997]).

This tree differs from the ones we have seen so far in that there is not a single root node, but an initial (non-singleton) information set. (In fact, there is just this one information set.) At

information set I, Ann has to choose between In and Out, but she does not know if this is the first or second time she is facing this choice. An interesting feature of this tree is that Ann can achieve a maximum payoff of 1 when choosing **deterministically** (she should choose In), but she can get a higher expected payoff by choosing **stochastically**. If Ann tosses a fair coin and chooses In or Out according to whether she gets Heads or Tails, her expected payoff is ½ × 0 + ¼ × 4 + ¼ × 1 = 5/4. One can show that a player can never gain by randomizing in a Kuhn tree (holding fixed what other players do). This example prompts another question about behavior: In which trees, in practice, do players choose to randomize, and in which trees do they not?

7. Behavior Under 'Complete Ignorance'

Having arrived at the matrix model of an interaction, von Neumann [1928] next investigated a particular rule for behavior. He assumed that each player must choose a strategy in a position of what he called "complete ignorance" concerning the strategies chosen by the other players and that, in this position, a player does not attempt to guess what other players will do but, instead, chooses 'safely.' This is his famous **maximin** (also known as **minimax**) decision criterion.

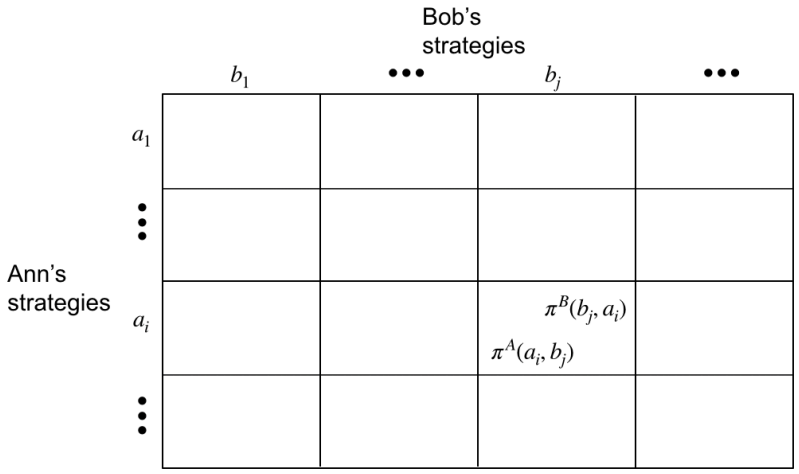

Figure 5

Let $\pi^A(a_i, b_j)$ be the payoff to Ann when she chooses strategy a_i and Bob chooses strategy b_j, and let $\pi^B(b_j, a_i)$ be the payoff to Bob, as depicted in Figure 5. Under the maximin criterion, Ann chooses a strategy to solve

$$\max_{a_i} \min_{b_j} \pi^A(a_i, b_j),$$

and Bob chooses a strategy to solve

$$\max_{b_j} \min_{a_i} \pi^B(b_j, a_i)\,.$$

In words, each player chooses a strategy that gives him or her the highest payoff possible, under the worst-case scenario about what strategy the other player chooses.

Use of the maximin decision rule provides another setting in which a player may benefit from choosing stochastically rather than deterministically. If Ann thinks that Bob will learn her strategy, then she may be able to achieve a higher expected payoff by choosing among her strategies $a_1, \ldots, a_i, \ldots$ with certain (non-degenerate) probabilities. The best probabilities for Ann will depend on the game in question. (This effect operates only if Bob learns what probabilistic strategy Ann uses, not the realization of her strategy.) In game-theory language, underlying strategies ai are called **pure strategies**, while probabilistic choices over underlying strategies are called **mixed strategies**.

Historically, this is the point at which the reflexive nature of interactions in a game was first formally analyzed. In von Neumann [1928], we can find the following question (not quite in these words). Suppose that Ann follows the maximin criterion, possibly with the use of a mixed strategy. She attributes the same decision rule to Bob — i.e., she assumes that Bob follows the maximin criterion, too. Then, does it still make sense for Ann to follow this criterion? If not, this would not be a formal inconsistency, but it would leave open the question of whether or not there is a decision criterion that players in a game can follow, while assuming 'like-mindedness' on the part of other players.

The famous Minimax Theorem (von Neumann [1928]) proves that in **two-player zero-sum games**, that is, in games where

$$\pi^A(a_i, b_j) + \pi^B(b_j, a_i) = 0,$$

for all pairs of strategies a_i, b_j, the answer to the consistency question is yes. In such games, the maximin criterion is consistent in this sense. A final note: The requirement that the payoffs sum to 0 is just a normalization. They could just as well sum to some constant number in each cell in the matrix. The term **constant-sum games** is sometimes used for this reason.

8. Beyond Two-Player Zero-Sum Games: Cooperative Theory

When we move to the general case of N**-player non-zero-sum games**, we encounter a very important fork in the road of game theory. (In N-player non-zero-sum games, the payoffs can sum to different numbers in different cells in the matrix.)

One path was developed by von Neumann and Morgenstern in their foundational book (von Neumann and Morgenstern [1944]). Nash [1951] developed a different path. After a brief sketch of the first path, we will turn to the second path.

For games with several players, von Neumann and Morgenstern [1944] were interested in the possibilities for players to act not only individually, but also in a joint or collective manner. Von Neumann and Morgenstern developed what is called the **cooperative branch** of game theory, as opposed to the **non-cooperative branch** (the individualistic branch) to which we will return later. In cooperative game theory, each possible **coalition** (subset) of players is assumed to be in joint control of the strategies available to the players in the coalition, and to be able to choose any profile of these strategies freely. The total payoff — called the **value** — associated with such a coalition of players is assumed to be the maximin total payoff to these players. That is, it is the highest total payoff they can achieve under the worst-case assumption about the strategy profile that the complementary coalition chooses. (Similar to before, randomization over underlying strategy profiles is allowed.) The formal definition of a cooperative game, then, is a pair $(\mathcal{N}, v)$, where $\mathcal{N} = \{1,\ldots, N\}$ is the set of players, and v is

a map called the **characteristic function**, which gives, for each possible coalition S from $\mathcal{N}$, the total value $v(S)$ achievable by that coalition.

A characteristic function is **superadditive** if $v(S \cup T) \geq v(S) + v(T)$ whenever the coalitions S and T are disjoint. Superadditivity is a natural (though not inevitable) assumption to make. Von Neumann and Morgenstern [1944] proved the important theorem that any superadditive characteristic function can arise from a suitably defined game matrix, when the players behave in cooperative fashion. This says that, fundamentally, there is one game theory, based on game trees and their induced game matrices. What separates non-cooperative and cooperative game theory is whether it is assumed that players act only individualistically or whether it is assumed that they can also act in a joint fashion.

The field of cooperative game theory is extensive, but we will not delve into it further in this chapter. Owen [1995] is a standard reference.

9. Beyond Two-Player Zero-Sum Games: Nash Equilibrium

Nash [1951] developed a theory of individual behavior in $\mathcal{N}$-player non-zero-sum games. He made three assumptions. First, he assumed that associated with each game is a unique profile of (possibly mixed) strategies that specifies the unique 'rational' way for each player to play the game. Second, he assumed that this strategy profile is known to all of the players. Third, he assumed that each player chooses a strategy that is expected-payoff maximizing.

For Nash's theory, we shift to mixed strategies as the basic object, with pure strategies now viewed as degenerate mixed strategies (which put probability 1 on a particular underlying pure strategy). Let σ^n denote a mixed strategy for player n. Following Nash's first assumption, let $(\sigma^1, \ldots, \sigma^N)$ be a candidate profile for the unique rational way to play a given game. By Nash's second assumption, player n knows that the other players $m \neq n$ choose the mixed strategies σ^m. By his third assumption, the strategy σ^n must maximize player n's expected payoff, when the other players $m \neq n$ choose the mixed strategies σ^m. A **Nash equilibrium** is a mixed-strategy profile $(\sigma^1, \ldots, \sigma^N)$ where this condition is satisfied for every player n. Nash [1951] proved that every game with finitely many players, each with a finite pure strategy set, possesses at least one Nash equilibrium.

We see that Nash took von Neumann's idea of a consistent decision rule — a rule that still works when other players follow the same rule — and asked for a much stronger kind of consistency. He asked for a form of consistency under which no player wants to change his (mixed) strategy when informed not just of the decision rules that the other players follow (expected-payoff maximization in Nash's theory), but of the actual (mixed) strategies that the other players choose.

An immediate criticism of Nash's theory is that it does not address how players might come to know the actual (mixed) strategies chosen by other players, or, even more, that this circumstance seems quite special from an empirical perspective. Nash's theory is a clear instance of arguing from a priori assumptions in game theory. This said, Nash equilibrium can be better grounded in the setting of **learning in games**, where players play a given game multiple times and have a chance to adapt to one another's actual choices in the game. This chapter will not expand on this topic; see Fudenberg and Levine [1998] for an introduction.

10. Testing Nash Equilibrium: A Posteriori Validity

Not long after Nash developed his theory, Flood [1958] carried out some **game experiments** to test how good a prediction Nash equilibrium was. He found that a basic premise of Nash's theory — that players act individually — did not necessarily hold. Players had some tendency to cooperate, more in line with the von Neumann-Morgenstern theory.

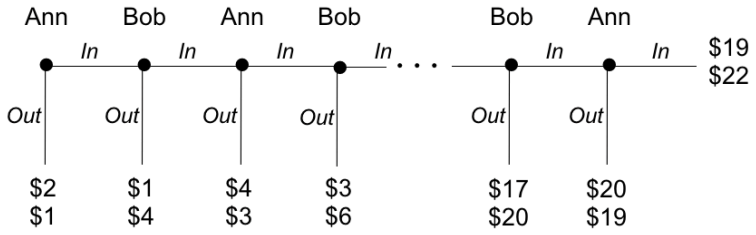

Figure 6

Within the modern experimental game theory literature, McKelvey and Palfrey [1992] is a seminal study of the empirical validity of Nash equilibrium. They used the so-called Centipede Game, created by Rosenthal [1982] and depicted in Figure 6. Here, Ann receives the top payoff at each terminal node, and Bob receives the bottom payoff. Ann has to decide, at each of her nodes, whether to choose Out and get a certain amount of money for sure, or choose In and, if Bob then chooses Out, get less, or if Bob then chooses In, be assured of more. The situation for Bob is analogous. At her final node, Ann clearly does better choosing Out over In. (All this is based on the assumption that the players care only about the monetary outcomes. If not, the game involves different considerations.)

Intuitively, both players will choose In for a while in Centipede, until Ann 'loses her nerve' and ends the game by choosing Out, or Bob does this first. This is indeed what is observed experimentally. Yet in any Nash equilibrium of this game, Ann chooses Out immediately. (If not, there is a player who moves immediately before the node at which the game ends. This player is not making an optimal choice in continuing to that node. This is essentially the proof for pure equilibrium. The argument can be extended to cover mixed equilibrium as well.)

Evidently, Nash equilibrium is not a good prediction in this game. The reason is apparent. Ann has no clear basis for correctly predicting Bob's strategy in Centipede (i.e., when he will choose Out). Likewise for Bob. In this game, the players' predictions seem likely to be based on guesses about each other's strategies. Moreover, when one of the players chooses Out at a certain node, we know that the other player's guess was wrong (or else the other player would have chosen Out one node earlier). We will come back in a while to game theory in which players make predictions, not necessarily correct, about one another's choices.

11. Testing Nash Equilibrium: A Priori Validity

A different issue with Nash equilibrium is that the concept can be argued not to go far enough in a priori theorizing about games. Further requirements should be included in the notion of rational behavior.

Go back to the game in Figure 1a and the associated matrix in Figure 2. This game has two Nash equilibria, one where Ann chooses L and Bob chooses l, and another where Ann chooses R and Bob chooses r. According to game theorists who work in the area called **refinement of Nash equilibrium**, the second Nash equilibrium does not properly reflect what the 'correct' definition of rationality should be. The argument is that Ann should not predict that Bob would play r if he actually got to move. At his node, Bob clearly does better choosing l over r. If Ann predicts this response, then she does better choosing L. The Nash equilibrium in which Ann chooses R and Bob chooses r does not hold up.

The general counterpart to this argument is the **backward-induction algorithm** (**BI**), which applies to any perfect-information tree. (Strictly speaking, it applies directly to any such tree with finitely many nodes, each of which has finitely many outgoing branches.) This concept first appears in von Neumann and Morgenstern [1944] (although it was likely known earlier). The general definition involves starting at final decision nodes — i.e., at decision nodes that lead only to terminal nodes of the tree. BI selects the best choice for the relevant player at each of these nodes and replaces the nodes with the resulting payoffs (to all the players). This process is repeated on the pruned tree, and it ends when it reaches the root of the tree.

Broadly speaking, the refinement program calls for restricting attention to Nash equilibria involving strategies that accord with BI. Some important aspects of this requirement need to be handled. The idea of BI needs to be extended to encompass imperfect-information as well as perfect-information trees. Also, it may be more natural to think in terms of the path through a tree that the BI strategies dictate, rather than the strategies themselves. Hillas and Kohlberg [2002] provide a comprehensive survey of the refinement program, covering these and many other issues.

A very important point to re-emphasize is that the refinement program is viewed — at least, by many of its practitioners — as definitely an a priori exercise. It is a journey to find a kind of theoretical baseline that involves the 'ultimate' meaning of rationality in games. Certainly, BI, just like Nash equilibrium, is not a good prediction in many games. The Centipede Game of Figure 6 makes this point. The unique BI path in this game is the same as the unique Nash-equilibrium path, that is, it is the path on which Ann chooses Out immediately.

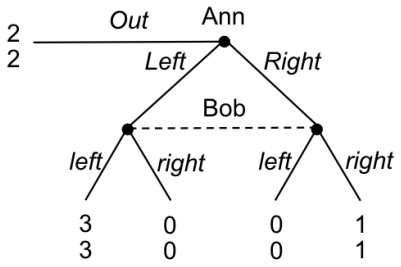

Figure 7

Another basic concept in the refinement program is complementary, at least in a rough sense, to BI. It is called **forward induction** (**FI**), and was introduced by Kohlberg and Mertens [1986]. Consider the game in Figure 7 (from Kohlberg and Mertens [1986]). There are two (pure) Nash equilibria, one in which Ann plays $Left$ and Bob plays $left$, and another in which Ann plays

Out and Bob plays $right$. FI rules out the second equilibrium via the following argument. When Bob gets to move, he knows that Ann has chosen either $Left$ or $Right$ (he does not know which) and that she has not chosen Out. Since Ann could have obtained a payoff of 2 by choosing Out, Bob should infer, the argument goes, that Ann has not chosen $Right$, which can give her a payoff of 1 at most. On this basis, Bob should choose $left$. Turning back to Ann, she can conclude that she should choose $Left$, anticipating that Bob will choose $left$, thereby giving her a payoff of 3, which is higher than she gets by choosing Out.

Thus, FI involves looking back in the tree and reasoning from what another player could have chosen but did not, while BI involves looking forward in the tree and reasoning from what another player would choose if given the opportunity. FI and BI are complementary ideas in this sense. We can ask about the exact nature of this complementarity. In particular, we can ask whether the two ideas can be implemented together in a game tree, or whether they somehow work against each other. This is the starting point for more-advanced research in the refinement program. See Kohlberg [1990] for a discussion of this point, and Govindan and Wilson [2009] for more recent progress in this area. The empirical (a posteriori) validity of FI has been investigated in the experimental games literature; see, e.g., Brandts, Cabrales, and Charness [2007].

12. Back to the Beginning: What the Players Know

So far, the method we have followed is to write down a game model and then go through some reasoning process (maximin, Nash, backward induction, forward induction) that can be attributed to the players. This raises a key issue. To see what this is, go back again to the game in Figure 1a. Backward induction says that Ann plays L and Bob plays l. The reasoning process invoked is that Ann anticipates that Bob, if he gets to move, will make the higher-payoff move of l vs. r. But this assumes that Ann knows Bob's payoffs.

Take a longer game tree, such as Centipede (Figure 6). The same reasoning process applied to this game assumes not only that Ann knows Bob's payoffs, but also that Ann knows Bob knows her payoffs, and so on for longer chains of "Ann knows Bob knows …," up to the length of the tree. A useful definition here is that of **common knowledge** (Aumann [1976], Friedell [1967], Lewis [1969]). A statement S (e.g., a specification of the players' payoffs) is said to be common knowledge if all players know S, all players know that all players know S, and so on indefinitely. The preceding assumption can then be stated succinctly, as the requirement that the game tree, payoffs included, is common knowledge between Ann and Bob. (For a tree of given length, common knowledge is overkill. But it is easiest to state an assumption that applies simultaneously to trees of different lengths.)

Von Neumann and Morgenstern [1944] were well aware of these considerations. They distinguished between games of **complete** and **incomplete information**, according to whether the game is or is not known to all the players. (They did not have the concept of common knowledge, and we would now restate their distinction in terms of common knowledge, not just knowledge.) They made clear that their analysis applied to complete-information games, and they left open the analysis of incomplete information.

Not all of game theory attributes reasoning to the players. A very important example of this other kind of game theory is **evolutionary game theory** (see Weibull [1997] for an introduction). The central concept here is that of an evolutionary stable strategy (ESS), proposed by Maynard Smith and Price [1973]. An ESS is a particular kind of Nash equilibrium, but defined for biological settings in which different players (organisms) are assumed to be genetically

encoded with different strategies, and payoffs measure reproductive success. Players are 'mindless' entities that do not reason or even choose strategies.

13. Games of Incomplete Information

The treatment of games of incomplete information began with Harsanyi [1967-8]. In fact, he made the bold move of proposing that such games be treated, instead, as games of **complete but imperfect information**. Let's apply Harsanyi's proposal to an example.

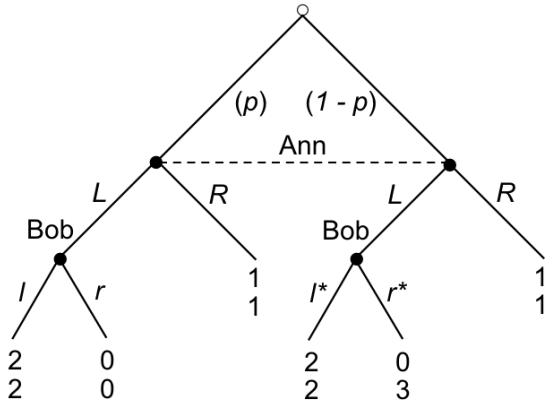

Figure 8

Figure 8 depicts a game tree in which Nature moves first (open node) to determine Bob's payoffs. If Nature chooses the branch labeled with probability p, then Bob gets a payoff of 0 when Ann chooses L and he chooses r. This is as in our first tree in Figure 1a. But if Nature chooses the branch labeled with probability $1 - p$, then Bob gets a payoff of 3 when Ann chooses L and he chooses r. (Technically, we have to use a different label — we use r^* — for Bob's choice r this time, since it is made at a different information set in the tree in Figure 8.) The key is that when Ann gets to move L or R, she does not know which are Bob's payoffs. If Ann thinks it sufficiently likely that Bob has the original payoffs (i.e., if p is high enough), she may decide to choose L (as in our earlier analysis). But if she thinks it sufficiently likely that Bob has the new payoffs (i.e., if p is low enough), she may decide to choose R on the basis that if she chooses L, then Bob will choose r^*.

We can also envisage more complicated games of incomplete information in which the players' higher-order knowledge matters. In some games, it might matter not only what Ann thinks Bob's payoffs are, but also what Ann thinks Bob thinks her payoffs are, or even what Ann thinks Bob thinks she thinks his payoffs are, and so on. To build a general language for incomplete-information games, we need to examine more deeply what is involved in the Harsanyi reduction to complete but imperfect-information games and then determine whether such a reduction is always possible. Early papers on these matters include Armbruster and Böge [1979], Böge and Eisele [1979], Mertens and Zamir [1985], and Brandenburger and Dekel [1993]; see Siniscalchi [2008] for a survey.

14. Back to the Beginning Again: Epistemic Game Theory

From the early days of the field analyzing Chess and other parlor games, game theorists have always been interested in the idea of consciously strategizing players who, in deciding how to

play, try to put themselves in the heads of the other players to infer how they play. Perhaps, Ann even tries to put herself in Bob's head to infer how he might put himself in her head to try to reach a better prediction about how Bob will play. This process might even continue to higher levels. It is the reflexive process we mentioned at the beginning of the chapter.

In one of the first papers in game theory, Morgenstern [1928] wrote about the famous battle of wits between Sherlock Holmes and Professor Moriarty in Conan Doyle's story *The Final Problem*, in exactly these "I think you think … I choose strategy s" terms. Yet, it is a curiosity in the history of game theory that, despite abiding intuitions about the central role of such reasoning in games, some of the principal methods of analysis that have been proposed — in particular, maximin and Nash equilibrium — have sidestepped this reasoning. In the case of maximin, a player does not attempt to predict the strategy chosen by another player, but adopts a best worst-case decision criterion, called "protective" by von Neumann and Morgenstern [1944]. In the case of Nash equilibrium, the prediction problem is assumed solved under Nash's [1951] assumption that all the players know the actual strategies chosen.

In the modern era, Bernheim [1984] and Pearce [1984] made the bold move of writing down the implications of very basic "I think you think … I choose strategy s" reasoning in a game matrix. Say that Ann is **rational** if she chooses a strategy that is expected-payoff maximizing for her, under some probability distribution she puts on Bob's set of strategies. Say that Ann is rational and knows that Bob is rational if the previous condition holds, and, moreover, she puts positive probability only on strategies for Bob that are expected payoff-maximizing for him, under some probability distribution he puts on Ann's set of strategies. Continuing this way, Bernheim [1984] and Pearce [1984] define the condition of **common knowledge of rationality**.

Previously in this chapter, we have generally put word "rationality" in single quotes, to indicate that it is more of a label than a defined term. In this section, however, the word has a precise definition. Moreover, the definition is deliberately a subjective internal one, in that it asks only for consistency (in the sense of Savage [1954]) among Ann's payoffs, probabilities, and behavior. Ann acts to maximize her expected payoff under her own subjective probabilities. She is not required to be objectively correct in any sense about the strategy that Bob chooses.

A strategy s^n for player n is **dominated** if there is another (possibly mixed) strategy σ^n for player n such that, with respect to each profile $(s^1, \ldots, s^{n-1}, s^{n+1}, \ldots, s^N)$ of strategies for players other than n, the strategy σ^n yields player n a strictly higher expected payoff than does s^n. With this definition, we can define a strategy elimination process on a game matrix, which begins by removing all dominated strategies for all players, to arrive at a sub-matrix. Then, in the submatrix, all dominated strategies for all players are removed, to arrive at a sub-matrix of the sub-matrix. This process continues until no further removals are possible. The surviving strategies are called the **iteratively undominated strategies** in the game.

Bernheim [1984] and Pearce [1984] argued that the assumption of common knowledge of rationality in a game implies that each player will choose an iteratively undominated strategy. This is quite intuitive, given one key mathematical step. This is the equivalence between the conditions that a strategy is undominated and that it is expected-payoff maximizing for some probability distribution on the strategy profiles of the other players (Arrow, Barankin, and Blackwell [1953]). Also, the focus that Bernheim and Pearce put on processes of iterated removal of 'bad' (as variously defined) strategies from a game had some precedents, Gale [1953] in particular. But the important novelty of their work is that it connects a particular **epistemic assumption** about the players in a game — here, the assumption of common

knowledge of rationality — to the players' behavior. The word "epistemic" means "of or relating to cognition or knowledge" (see http://en.wiktionary.org).

The argument we just went through raises a number of issues. First, the epistemic condition in question is better called **rationality and common belief of rationality** than common knowledge of rationality. This is because the epistemic modality involved is probability (as in Ann "puts positive probability only on strategies …"), which describes a subjective belief state of a player, not the objective certainty of knowledge. A second issue is the treatment of **correlation vs. independence**, which arises once we consider a game with three players, say: Ann, Bob, and Charlie. The question is whether Charlie forms a probability assessment that treats Ann's choice of strategy and Bob's choice of strategy as independent or — as some have argued to be more appropriate — as possibly correlated. Brandenburger and Friedenberg [2008] develop an epistemic theory of correlation in games. A very important third issue is the extension of this whole line of analysis to game trees, where there is not only the matter of the players' beliefs, beliefs about beliefs, etc., but also the **revision of beliefs** in light of what players observe in the tree about other players' moves as play proceeds. Battigalli and Siniscalchi [1999, 2002] make fundamental advances in this area. A fourth issue is finding epistemic foundations for **iterated weak dominance** (the definition earlier is for **iterated strong dominance**); see Brandenburger, Friedenberg, and Keisler [2008].

The investigation of these and other issues marks the beginning of what is now called **epistemic game theory** (**EGT**). Not surprisingly, the underlying mathematical apparatus is the same as that built to formalize the Harsanyi [1967-8] reduction described in the previous section. Whether the uncertainty is over the structure (payoffs, say) of the game, as in incomplete-information game theory, or over the strategies chosen in the game, as in EGT, there is the same need to build hierarchical models describing beliefs, beliefs about beliefs, etc. over what is uncertain.

EGT is a good example of research in game theory that aims to extend the expressive power of game theory as a language. With the benefit of the language of hierarchies of beliefs, notions such as rationality, belief, belief about rationality, belief about belief, etc., can now be given precise meanings, and the implications of different epistemic assumptions on games can be worked out. See Brandenburger [2014] for an introduction to the field.

15. Epistemics Meets Experiments

Work in experimental game theory has already been mentioned in this chapter. Nagel [1995] and Stahl and Wilson [1995] are pioneering experimental papers on levels of reasoning in games. There is a fundamental **identification problem** in this area. The strategies consistent with rationality and $(m+1)$th-order belief of rationality, are a subset of the strategies consistent with rationality and mth-order belief of rationality, for any $m = 0,1,2,\ldots$. This is just a re-statement of the basic argument from the previous section that these epistemic conditions correspond to iterated removal of 'bad' (dominated) strategies. This simple observation implies the following identification problem: A player might choose a strategy consistent with a high number of levels of reasoning, but might do so without necessarily reasoning to this level.

Early experimental work in this area (Nagel [1995], Stahl and Wilson [1995], and subsequent papers) overcame the identification problem by making specific assumptions about the players' beliefs and behavior. More recent work by Kneeland [2015] achieves identification making only basic epistemic assumptions about the players. This is possible because of a novel experimental design in Kneeland [2015], which involves ring games.

Figure 9 depicts a **ring game** in schematic form. This is an N-player game in which payoffs depend on strategies in a particular way. Player 1's payoffs depend on the strategy he chooses and on the strategy player 2 chooses. His payoffs do not depend on the strategies chosen by any of the players 3, 4, …, N. Player 2's payoffs depend on the strategy she chooses and on the strategy player 3 chooses, and on no other strategies chosen. This dependency pattern continues until we reach player N, whose payoffs depend on the strategy she chooses and on the strategy player 1 chooses.

The experimenter now examines player 1's behavior as the payoffs to various players are manipulated. Suppose that manipulating player 2's payoffs changes player 1's behavior. Then, we can conclude that player 1 satisfies at least rationality and 1st-order belief of rationality. If manipulating player 3's payoffs changes player 1's behavior, we can conclude that player 1 satisfies at least rationality and 2nd-order belief of rationality. The first manipulation that does not affect player 1's behavior identifies the order of belief of rationality satisfied by player 1.

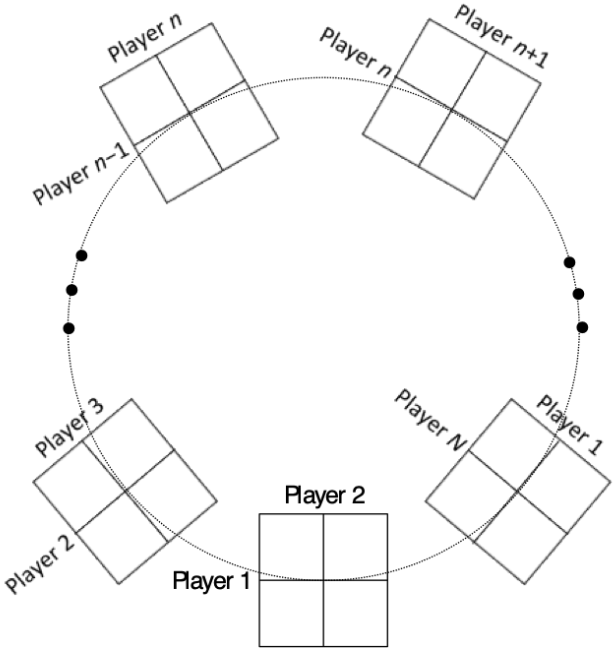

Figure 9

16. From Rationality to Cognition

The term cognition refers to the process of thinking. Although we have already defined rationality, when using this term now, let's have in mind a process of thoughtfully and well-chosen (e.g., undominated) action. Then, we can say that we can have **cognition without rationality**, although we cannot have rationality without cognition.

This simple point alerts us to the risk of drawing an unwarranted inference from experimental epistemics. Suppose that we run a ring-game experiment, as in the previous section, and conclude that player 1 is rational but does not satisfy 1st-order belief of rationality. (The case of a higher-order belief of rationality is more interesting, but let's keep to this case for simplicity.) The temptation is to infer that the player thinks one level (he thoughtfully chooses his strategy), but he does not think two levels (he does not think about how player 2 thoughtfully chooses her strategy). This is an unwarranted inference. It may be that player 1

does think about how player 2 thoughtfully chooses her strategy, but he does not think that she is rational as defined. Perhaps he thinks that she chooses a dominated strategy because she has some separate (unknown) reason for doing so. This question of how to infer levels of cognition, not just levels of rationality, from behavior in games is a current frontier; see, e.g., Friedenberg and Kneeland [2023] and Jin [2021].

Outside game theory, levels of cognition have been much studied in cognitive psychology, animal behavior, and cognitive neuroscience. Here, the term used is **Theory of Mind** (**ToM**), due to Premack and Woodruff [1978] and defined as the ability to think about another person's intentions, beliefs, and desires (Sabbagh and Bowman [2018]). A typical ToM experiment in **cognitive psychology** has a subject read a short story describing a social situation and then asks the subject questions about the story (Kinderman, Dunbar, and Bentall [1998] and Stiller and Dunbar [2007]). The questions differ in terms of the number of levels of "Ann thinks Bob thinks Charlie thinks …" that they contain (Ann, Bob, and Charlie are characters in the story). A subject's ToM ability is defined as the maximum number of such levels that a question can contain and still elicit a correct answer from the subject. There would seem to be good opportunities for cross-fertilization between game-based and narrative-based studies of levels of cognition.

McCabe et al. [2001], Gallagher et al. [2002], and Rilling et al. [2004] established a very important link between **cognitive neuroscience** and game theory. These papers used neuroimaging to establish that when people play a game with a human (but not with a computer) counterpart, regions of their brain known to be active in ToM processing are similarly activated. This is direct empirical evidence that justifies building the language of game theory in terms of "I think you think …" ingredients.

As game theory today joins the cognitive sciences, more effort in the field is shifting towards a posteriori theory building and away from a priori theory building. This is a healthy and exciting development.

Acknowledgements
I am indebted to Rena Henderson, Jessy Hsieh, Ye (Wendy) Jin, Terri Kneeland, Yixu Lin, Paula Miret, Rosemarie Nagel, Kai Steverson, and to attendees at the 2015 Shanghai Neuroeconomics Collective Summer School, NYU Shanghai, July 2015, for important input. A a referee and an editor made valuable suggestions. Financial support from the NYU Stern School of Business, NYU Shanghai, and J.P. Valles is gratefully acknowledged.

References
Armbruster, W., and W. Böge, "Bayesian Game Theory," in Moeschlin, O, and D. Pallaschke (eds.), *Game Theory and Related Topics*, North-Holland, 1979, 17-28.

Arrow, K., E. Barankin, and D. Blackwell, "Admissible Points of Convex Sets," in Kuhn H., and A. Tucker (eds.), *Contributions to the Theory of Games*, Vol. II, Princeton University Press, 1953, 87-91.

Aumann, R., "Agreeing to Disagree," *Annals of Statistics*, 4, 1976, 1236-1239.

Battigalli, P., and M. Siniscalchi, "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games," *Journal of Economic Theory*, 88, 1999, 188-230.

Battigalli, P., and M. Siniscalchi, "Strong Belief and Forward-Induction Reasoning," *Journal of Economic Theory*, 106, 2002, 356-391.

Bernheim, D., “Rationalizable Strategic Behavior,” *Econometrica*, 52, 1984, 1007-1028.

Böge, W., and T. Eisele, “On Solutions of Bayesian Games,” *International Journal of Game Theory*, 8, 1979, 193-215.

Brandenburger, A., *The Language of Game Theory: Putting Epistemics into the Mathematics of Games*, World-Scientific, 2014.

Brandenburger, A., and E. Dekel, “Hierarchies of Beliefs and Common Knowledge,” *Journal of Economic Theory*, 59, 1993, 189-198.

Brandenburger, A., and A. Friedenberg, “Intrinsic Correlation in Games,” *Journal of Economic Theory*, 141, 2008, 28-67.

Brandenburger, A., A. Friedenberg, and H.J. Keisler, “Admissibility in Games,” *Econometrica*, 76, 2008, 307-352.

Brandts, J., A. Cabrales, and G. Charness, “Forward Induction and Entry Deterrence: An Introduction,” *Economic Theory*, 33, 2007, 183-209.

Cooper, D., and J. Van Huyck “Evidence on the Equivalence of the Strategic and Extensive Form Representation of Games,” *Journal of Economic Theory*, 110, 2003, 290-308.

Dalkey, N., “Equivalence of Information Patterns and Essentially Determinate Games,” in Kuhn, H., and A. Tucker (eds.), *Contributions to the Theory of Games*, Volume II, Princeton University Press, Princeton, 1953, 217-244.

Dawkins, R., *The Selfish Gene*, Oxford University Press, 1976.

Devlin, K., *The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern*, Basic Books, 2008.

Flood, M., “Some Experimental Games,” *Management Science*, 5, 1958, 5-26.

Friedell, M., “On the Structure of Shared Awareness,” Paper #27, Center for Research on Social Organization, University of Michigan, 1967.

Friedenberg, A., and T. Kneeland, “Is Bounded Reasoning about Rationality Driven By Limited Ability?”, 2023, available at amandafriedenberg.org.

Fudenberg, D., and D. Levine, *The Theory of Learning in Games, MIT Press*, 1998.

Gale, D., “A Theory of *N*-Person Games with Perfect Information,” *Proceedings of the National Academy of Sciences*, 39, 1953, 496-501.

Gallagher, H., A. Jack, A. Roepstorff and C. Frith, “Imaging the Intentional Stance in a Competitive Game,” *NeuroImage*, 16, 2002, 814-821.

Govindan, H., and R. Wilson, “On Forward Induction,” *Econometrica*, 77, 2009, 1-28.

Halpern, J., and Y. Moses, “Knowledge and Common Knowledge in a Distributed Environment,” *Journal of the ACM*, 37, 1990, 549-587.

Harsanyi, J., “Games with Incomplete Information Played by ‘Bayesian’ Players, I-III,” *Management Science*, 14, 1967-8, 159-182, 320-334, 486-502.

Hillas, J., and E. Kohlberg, “Foundations of Strategic Equilibrium,” in Aumann, R., and S. Hart (eds.), *Handbook of Game Theory*, Volume 3, Elsevier, 2002, 1597-1663.

Isbell, J., “Finitary Games,” in Drescher, M., A. Tucker, and P. Wolfe (eds.), *Contributions to the Theory of Games*, Volume III, Princeton University Press, 1957.

Jin, Y., “Does Level-k Behavior Imply Level-k Thinking?” *Experimental Economics*, 24, 2021, 330-353.

Kinderman, P., R. Dunbar, and R. Bentall, “Theory-of-Mind Deficits and Causal Attributions,” *British Journal of Psychology*, 89, 1998,191-204.

Kneeland, T., “Identifying Higher-Order Rationality,” *Econometrica*, 83, 2015, 2065-2079.

Kohlberg, E., “Refinement of Nash Equilibrium: The Main Ideas,” in Ichiishi, T., A. Neyman, and Y. Tauman (eds.), *Game Theory and Applications*, Academic Press, 1990.

Kohlberg, E., and J.-F. Mertens, “On the Strategic Stability of Equilibria,” *Econometrica*, 54, 1986, 1003-1038.

Kuhn, H., “Extensive Games,” *Proceedings of the National Academy of Sciences*, 36, 1950, 570-576.

Kuhn, H., “Extensive Games and the Problem of Information,” in Kuhn, H., and A. Tucker (eds.), *Contributions to the Theory of Games*, Volume II, Princeton University Press, 1953, 193-216.

Leonard, R., *Von Neumann, Morgenstern, and the Creation of Game Theory: From Chess to Social Science*, 1900–1960, Cambridge University Press, 2012.

Lewis, D., *Convention: A Philosophical Study*, Harvard University Press 1969.

Maynard Smith, J., and G. Price, “The Logic of Animal Conflict,” *Nature*, 246, 1973, 15-18.

McCabe, K., D. Houser, L. Ryan, V. Smith and T. Trouard, “A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange,” *Proceedings of the National Academy of Sciences*, 98, 2001, 11832-11835.

McKelvey, R., and T. Palfrey, “An Experimental Study of the Centipede Game,” *Econometrica*, 60, 1992, 803-836.

Mertens, J-F., and S. Zamir, “Formulation of Bayesian Analysis for Games with Incomplete Information,” *International Journal of Game Theory*, 14, 1985, 1-29.

Morgenstern, O., *Wirtschaftsprognose, eine Untersuchung ihrer Voraussetzungen und Mödglichkeiten*, Springer Verlag, 1928.

Nagel, R., “Unraveling in Guessing Games: An Experimental Study,” *American Economic Review*, 85, 1995, 1313-1326.

Nash, J., “Non-cooperative Games,” *Annals of Mathematics*, 54, 1951, 286-295.

Owen G., *Game Theory*, 3rd edition, Emerald, 1995.

Pearce, D., “Rational Strategic Behavior and the Problem of Perfection,” *Econometrica*, 52, 1984, 1029-1050.

Piccione, M., and A. Rubinstein, “On the Interpretation of Decision Problems with Imperfect Recall,” *Games and Economic Behavior*, 20, 1997, 3-24.

Premack, D., and G. Woodruff, “Does the Chimpanzee Have a Theory of Mind?” *Behavioral and Brain Sciences*, 4, 1978, 515-526.

Rilling, J., A. Sanfey, J. Aronson, L. Nystrom and J. Cohen, “The Neural Correlates of Theory of Mind Within Interpersonal Interactions,” *NeuroImage*, 22, 2004, 1694-1703.

Rosenthal, R., “Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox,” *Journal of Economic Theory*, 25, 1982, 92-100.

Sabbagh, M., and L. Bowman, “Theory of Mind,” in Wixted, J., and S. Ghetti (eds.), *The Stevens’ Handbook of Experimental Psychology, Volume 4, Developmental and Social Psychology*, Wiley, 2018, 249-287.

Savage, L., *The Foundations of Statistics*, Wiley, 1954.

Schotter, A., K. Weigelt, and C. Wilson, “A Laboratory Investigation of Multiperson Rationality and Presentation Effects,” *Games and Economic Behavior*, 6, 1994, 445-468.

Siniscalchi, M., “Epistemic Game Theory: Beliefs and Types,” in Blume, L., and S. Durlauf (eds.), *The New Palgrave Dictionary of Economics*, 2nd edition, Palgrave McMillan, 2008.

Stahl, D. and P. Wilson, “On Players’ Models of Other Players: Theory and Experimental Evidence,” *Games and Economic Behavior*, 10, 1995, 218-254.

Stiller, J., and R. Dunbar, “Perspective-Taking and Memory Capacity Predict Social Network Size,” *Social Networks*, 29, 2007, 93-104.

Thompson, F., “Equivalence of Games in Extensive Form,” Research Memorandum RM-759, The RAND Corporation, 1952.

Von Neumann, J., “Zur Theorie der Gesellschaftsspiele,” *Mathematische Annalen*, 100, 1928, 295-320. English translation by Bargman, S., “On the Theory of Games of Strategy,” in Tucker, A., and R.D. Luce (eds.), *Contributions to the Theory of Games*, Volume IV, Princeton University Press, 1955, 13-42.

Von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1944.

Weibull, J., *Evolutionary Game Theory*, MIT Press, 1997.

West, S., A. Griffin, A. Gardner, and S. Diggle, “Social Evolution Theory for Microorganisms,” *Nature Reviews Microbiology*, 4, 2006, 597-607.