Chapter 9

Epistemic Foundations of Game Theory

Giacomo Bonanno

Contents

Abstract This chapter provides an introduction to the so-called epistemic foundation program in game theory, whose aim is to characterize, for any game, the behavior of rational and intelligent players who know the structure of the game and the preferences of their opponents and who recognize each other's rationality and reasoning abilities. The analysis is carried out both semantically and syntactically, with a focus on the implications of common belief of rationality in strategic-form games and in dynamic games with perfect information.

Chapter 9 of the *Handbook of Logics for Knowledge and Belief*, H. van Ditmarsch, J.Y. Halpern, W. van der Hoek and B. Kooi (eds), College Publications, 2015, pp. 411–450.

9.1 Introduction

Game theory provides a formal language for the representation of interactive situations, that is, situations where several "entities" - called players - take actions that affect each other. The nature of the players varies depending on the context in which the game theoretic language is invoked: in evolutionary biology players are non-thinking living organisms; in computer science players are artificial agents; in behavioral game theory players are "ordinary" human beings, etc. Traditionally, however, game theory has focused on interaction among intelligent, sophisticated and rational individuals. The focus of this chapter is a relatively recent development in game theory, namely the so-called *epistemic foundation program.* The aim of this program is to characterize, for any game, the behavior of rational and intelligent players who know the structure of the game and the preferences of their opponents and who recognize each other's rationality and reasoning abilities. The two fundamental questions addressed in this literature are: (1) Under what circumstances can a player be said to be rational? and (2) What does 'mutual recognition' of rationality mean? Since the two main ingredients of the notion of rationality are beliefs and choice and the natural interpretation of 'mutual recognition' of rationality is in terms of common belief, it is clear that the tools of epistemic logic are the appropriate tools for this program.

It is useful to distinguish three related notions that have emerged in the analysis of games. The first notion is that of a solution concept, which is a map that associates with every game a set of strategy profiles that constitute a prediction of how the game will be played. Examples of solution concepts are Nash equilibrium, correlated equilibrium, perfect equilibrium, etc. The second notion is that of an algorithm that computes, for every game, a set of strategy profiles. The algorithm is often presented as an attempt to capture the steps in the reasoning process of the players. An example is the iterated deletion of dominated strategies. The third notion is that of an explicit epistemic hypothesis that describes the players' state of mind. An example is the hypothesis of common belief of rationality. Epistemic game theory is concerned with the third notion and seeks to provide an understanding of existing solution concepts in terms of explicit epistemic conditions, as well as a framework within which new solution concepts can be generated.

The chapter is organized as follows. In Sections 9.2 and 9.3 we begin with the semantic approach to rationality in simultaneous games with ordinal payoff. In Sections 9.4 and 9.5 we turn to the syntactic approach and explore the difference between common belief and common knowledge of rationality. In Section 9.6 we briefly discuss probabilistic beliefs and cardinal preferences. In Sections 9.7, 9.8 and 9.9 we turn to a semantic analysis of rationality in dynamic games with perfect information, based on dispositional belief revision (or subjective counterfactuals). Section 9.10 lists the most important contributions in the literature for the topics discussed in this chapter and gives references for additional solution concepts that could not be covered in this chapter because of space constraints.

9.2 Epistemic Models of Strategic-Form Games

Traditionally, game-theoretic analysis has been based on the assumption that the game under consideration is common knowledge among the players. Thus not only is it commonly known who the players are, what choices they have available and what the possible outcomes are, but also how each player ranks those outcomes. While it is certainly reasonable to postulate that a player knows his own preferences over the possible outcomes, it is much more demanding to assume that a player knows the preferences of his opponents. If those preferences are expressed as ordinal rankings of the outcomes, this assumption is less troublesome than in the case where preferences also incorporate attitudes to risk (that is, the payoff functions that represent those preferences are Bernoulli, or von Neumann Morgenstern, utility functions: see Section 9.6). We will thus begin by considering the case where preferences are expressed by *ordinal* rankings.

We first consider games where each player chooses in ignorance of the choices of the other players (as is the case, for example, in simultaneous games).

Definition 9.1
A *finite strategic-form game with ordinal payoffs* is a quintuple

$G = \left\langle \mathsf{Ag}, \{S_i\}_{i\in\mathsf{Ag}}, O, z, \{\succsim_i\}_{i\in\mathsf{Ag}} \right\rangle$

where

$\mathsf{Ag} = \{1, 2, \ldots, n\}$ is a finite set of *players*,

S_i is a finite set of *strategies* (or choices) of player $i \in \mathsf{Ag}$,

O is a finite set of *outcomes*,

$z : S \to O$ (where $S = S_1 \times ... \times S_n$) is a function that associates with every strategy profile $s = (s_1, ..., s_n) \in S$ an outcome $z(s) \in O$,

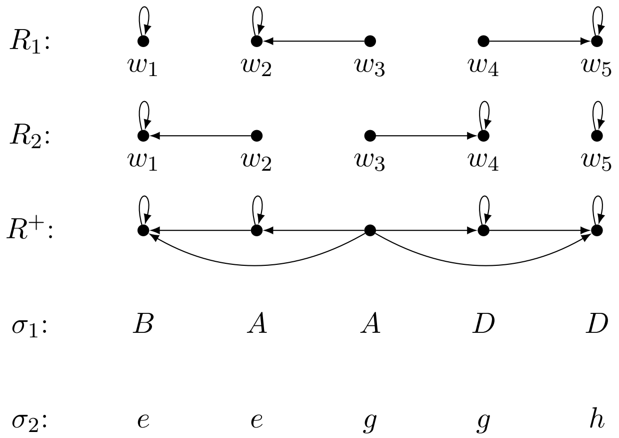
$\succsim_i$ is player i's *ranking* of O, that is, a binary relation on O which is complete (for all $o, o' \in O$, either $o \succsim_i o'$ or $o' \succsim_i o$) and transitive (for all $o, o', o'' \in O$, if $o \succsim_i o'$ and $o' \succsim_i o''$ then $o \succsim_i o''$). The interpretation of $o \succsim_i o'$ is that player i considers outcome o to be at least as good as outcome o'. The corresponding strict ordering, denoted by $\succ_i$, is defined by: $o \succ_i o'$ if and only if $o \succsim_i o'$ and not $o' \succsim_i o$. The interpretation of $o \succ_i o'$ is that player i strictly prefers outcome o to outcome o'.

Remark 9.1
Games are often represented in *reduced form*, which is obtained by replacing the triple $\langle O, z, \{\succsim_i\}_{i\in\mathsf{Ag}}\rangle$ with a set of *payoff functions* $\{\pi_i\}_{i\in\mathsf{Ag}}$ where $\pi_i : S \to \mathbb{R}$ is any real-valued function that satisfies the property that, $\forall s, s' \in S$, $\pi_i(s) \geq \pi_i(s')$ if and only if $z(s) \succsim_i z(s')$. In the following we will adopt this more succinct representation of strategic-form games. It is important to note, however, that (with the exception of Section 9.6) the payoff functions are taken to be purely ordinal and one could replace π_i with any other function obtained by composing π_i with an arbitrary strictly increasing function on the reals. $\dashv$

Part a of Figure 9.1 shows a two-player strategic-form game where the sets of strategies are $S_1 = \{A, B, C, D\}$ and $S_2 = \{e, f, g, h\}$. The game is represented as a table where the rows are labeled with the possible strategies of Player 1 and the columns with the possible strategies of Player 2. Each cell in the table corresponds to a strategy-profile, that is, an element of $S = S_1 \times S_2$; inside each cell the first number is the payoff of Player 1 and the second number is the payoff of Player 2; thus, for example, $\pi_1(A, e) = 6$ and $\pi_2(A, e) = 3$.

		Player 2			
		e	f	g	h
Player 1	A	$6,3$	$4,4$	$4,1$	$3,0$
	B	$5,4$	$6,3$	$0,2$	$5,1$
	C	$5,0$	$3,2$	$6,1$	$4,0$
	D	$2,0$	$2,3$	$3,3$	$6,1$

(a) A strategic-form game G



(b) An epistemic model of game G

Figure 9.1: A strategic-form game and an epistemic model of it

A strategic-form game provides only a partial description of an interactive situation, since it does not specify what choices the players make, nor what beliefs they have about their opponents' choices. A specification of these missing elements is obtained by introducing the notion of an epistemic model of a game, which represents a possible context in which the game is played. The players' beliefs are represented by means of a $\mathcal{KD}45$ Kripke frame $\langle W, \{R_i\}_{i \in \mathsf{Ag}} \rangle$, where W is a set of *states* (or possible worlds) and, for every player i, R_i is a binary relation on W which is serial ($\forall w \in W$, $R_i(w) \neq \varnothing$, where $R_i(w)$ denotes the set $\{w' \in W : wR_iw'\}$), transitive (if $w' \in R_i(w)$ then $R_i(w') \subseteq R_i(w)$) and euclidean (if $w' \in R_i(w)$ then $R_i(w) \subseteq R_i(w')$).[1] Given a state w, $R_i(w)$ is interpreted as

[1] In the game-theoretic literature, it is more common to view R_i as a function that associates with every state $w \in W$ a set of states $R_i(w) \subseteq W$ and to call such a function

the set of states that are doxastically accessible to player i at w, that is, the states that she considers possible according to her beliefs. The player, at a state w, is said to believe a formula φ if and only if φ is true at every state that she considers possible at w. Seriality of the accessibility relation R_i guarantees that the player's beliefs are consistent (it is not the case that she believes φ and also $\neg\varphi$), while transitivity corresponds to positive introspection (if the player believes φ then she believes that she believes φ) and Euclideaness corresponds to negative introspection (if the player does not believe φ then she believes that she does not believe φ). Note that erroneous beliefs are not ruled out: it is possible that a player believes φ even though φ is actually false.[2]

Definition 9.2
Given a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i\in\mathsf{Ag}} \right\rangle$ an *epistemic model of G* is a tuple $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{\sigma_i\}_{i\in\mathsf{Ag}}\rangle$ where $\langle W, \{R_i\}_{i\in\mathsf{Ag}}\rangle$ is a $\mathcal{KD}45$ Kripke frame and, for every player $i \in \mathsf{Ag}$, $\sigma_i : W \to S_i$ is a function that satisfies the following property: if $w' \in R_i(w)$ then $\sigma_i(w') = \sigma_i(w)$. ⊣

The interpretation of $\sigma_i(w) = s_i \in S_i$ is that, at state w, player i chooses strategy s_i and the requirement that if $w' \in R_i(w)$ then $\sigma_i(w') = \sigma_i(w)$ expresses the assumption that a player is always certain about what choice he himself makes. On the other hand, a player may be uncertain about the choices of the other players.

Remark 9.2
In an epistemic model of a game the function $\sigma : W \to S$ defined by $\sigma(w) = (\sigma_i(w))_{i\in\mathsf{Ag}}$ associates with every state a strategy profile. Given a state w and a player i, we will often denote $\sigma(w)$ by $(\sigma_i(w), \sigma_{-i}(w))$, where $\sigma_{-i}(w) \in S_{-i} = S_1 \times ... \times S_{i-1} \times S_{i+1} \times ... \times S_n$. Thus $\sigma_{-i}(w)$ is the strategy profile of the players other than i at state w. ⊣

Part b of Figure 9.1 shows an epistemic model for the game of Part a. The relations R_i $(i = 1, 2)$ are represented by arrows: for player i there is an arrow from state w to state w' if and only if $w' \in R_i(w)$. The relation R^+, which is discussed below, is the transitive closure of $R_1 \cup R_2$. [3]

a *possibility correspondence* or information correspondence. Of course, the two views (binary relation and possibility correspondence) are equivalent.

[2] Erroneous beliefs are ruled out if one imposes the restriction that R_i be reflexive ($w \in R_i(w), \forall w \in W$). If reflexivity is added to the above assumptions, then R_i gives rise to a partition of W and in such a case it is common to use the term 'knowledge' rather than 'belief'. In the game-theoretic literature, partitional structures tend to be more common than $\mathcal{KD}45$ frames.

[3] Thus in Figure 9.1 we have that

$$\begin{array}{rcl} R_1 & = & \{(w_1,w_1),(w_2,w_2),(w_3,w_2),(w_4,w_5),(w_5,w_5)\}, \\ R_2 & = & \{(w_1,w_1),(w_2,w_1),(w_3,w_4),(w_4,w_4),(w_5,w_5)\}, \text{and} \\ R^+ & = & \{(w_1,w_1),(w_2,w_1),(w_2,w_2),(w_3,w_1),(w_3,w_2), \\ & & (w_3,w_4),(w_3,w_5),(w_4,w_4),(w_4,w_5),(w_5,w_5)\}. \end{array}$$

Hence, for example, in terms of our notation, $R_1(w_3) = \{w_2\}$, $R_2(w_3) = \{w_4\}$ and $R^+(w_3) = \{w_1, w_2, w_4, w_5\}$.

In the game-theoretic literature individual beliefs and common belief are typically represented by means of semantic operators on events. Given a $\mathcal{KD}45$ Kripke frame $\langle W, \{R_i\}_{i \in \mathsf{Ag}} \rangle$, an *event* is any subset of W and one can associate with the doxastic accessibility relation R_i of player i a *semantic belief operator* $\mathbb{B}_i : 2^W \to 2^W$ and a *semantic common belief operator* $\mathbb{CB} : 2^W \to 2^W$ as follows:

$$\begin{aligned} \mathbb{B}_i E &= \{w \in W : R_i(w) \subseteq E\}, \text{ and} \\ \mathbb{CB} E &= \{w \in W : R^+(w) \subseteq E\} \end{aligned} \tag{9.1}$$

where R^+ is the transitive closure of $\bigcup_{i \in \mathsf{Ag}} R_i$.[4,5] $\mathbb{B}_i E$ is interpreted as the event that (that is, the set of states at which) player i believes event E and $\mathbb{CB}E$ as the event that E is commonly believed.[6]

The analysis of the consequences of common belief of rationality in strategic-form games was first developed in the game-theoretic literature from a semantic point of view. We will review the semantic approach in the next section and turn to the syntactic approach in Section 9.4.

9.3 Semantic Analysis of Common Belief of Rationality

A player's choice is considered to be rational if it is "optimal", given the player's beliefs about the choices of the other players. When beliefs are expressed probabilistically and payoffs are taken to be von Neumann-Morgenstern payoffs, a choice is optimal if it maximizes the player's expected payoff. We shall discuss the notion of expected payoff maximization in Section 9.6. In this section we will focus on the non-probabilistic beliefs represented by the qualitative Kripke frames introduced in Definition 9.2.

Within the context of an epistemic model of a game, a rather weak notion of rationality is the following.

[4] In the game-theoretic literature the transitive closure of the union of the accessibility relations is called the 'finest common coarsening'.

[5] The intuitive and prevalent definition of common belief is as follows. Let $\mathbb{B}_{all} E = \bigcap_{i \in \mathsf{Ag}} \mathbb{B}_i E$ denote the event that everybody believes E. Then the event that E is commonly believed is defined as the infinite intersection $\mathbb{CB} E = \mathbb{B}_{all} E \cap \mathbb{B}_{all} \mathbb{B}_{all} E \cap \mathbb{B}_{all} \mathbb{B}_{all} \mathbb{B}_{all} E \cap \dots$, that is, the event that everybody believes E and everybody believes that everybody believes E and everybody believes that everybody believes that everybody believes E, and so on. Let us call this the infinitary definition of common belief. It can be shown that, for every state w and every event E, $w \in \mathbb{CB}E$ according to the infinitary definition of $\mathbb{CB}$ if and only if $R^+(w) \subseteq E$.

[6] The operator $\mathbb{B}_i$ satisfies the following properties: $\forall E \subseteq W$, (i) Consistency: if $E \neq \varnothing$ then $\mathbb{B}_i E \neq \varnothing$, (because of seriality of R_i), (ii) Positive Introspection: $\mathbb{B}_i E \subseteq \mathbb{B}_i \mathbb{B}_i E$ (because of transitivity of R_i), (iii) Negative Introspection: $\neg \mathbb{B}_i E \subseteq \mathbb{B}_i \neg \mathbb{B}_i E$ (because of Euclideaness of R_i, where $\neg F$ denotes the complement of event F). Among the properties of the common belief operator $\mathbb{CB}$ we highlight one that we will use later, which is a consequence of transitivity of R^+: $\mathbb{CB} E \subseteq \mathbb{CB}\ \mathbb{CB} E$.

Definition 9.3
Fix a strategic-form game G and an epistemic model of G. At state w player i's strategy $s_i = \sigma_i(w)$ is *rational* if it is not the case that there is another strategy $s'_i \in S_i$ of player i which yields a higher payoff than s_i against *all* the strategy profiles of the other players that player i considers possible, that is, if

$$\{s'_i \in S_i : \pi_i\left(s'_i, \sigma_{-i}(w')\right) > \pi_i\left(\sigma_i(w), \sigma_{-i}(w')\right),\ \forall w' \in R_i(w)\} = \varnothing$$

[recall that, by Definition 9.2, the function $\sigma_i(\cdot)$ is constant on the set $R_i(w)$]. Equivalently, $s_i = \sigma_i(w)$ is rational at state w if, for every $s'_i \in S_i$, there exists a $w' \in R_i(w)$ such that $\sigma_i(w)$ is at least as good as s'_i against the strategy profile $\sigma_{-i}(w')$ of the other players, that is, $\pi_i\left(\sigma_i(w), \sigma_{-i}(w')\right) \geq \pi_i\left(s'_i, \sigma_{-i}(w')\right)$. ⊣

Given an epistemic model of a strategic-form game G, using Definition 9.3 one can compute the event that player i's choice is rational. Denote that event by RAT_i. Let $RAT = \bigcap_{i \in \mathsf{Ag}} RAT_i$. Then RAT is the event that (the set of states at which) the choice of every player is rational. One can then also compute the event $\mathbb{CB}RAT$, that is, the event that it is common belief among the players that every player's choice is rational. For example, in the epistemic model of Part *b* of Figure 9.1, $RAT_1 = \{w_2, w_3, w_4, w_5\}$ and $RAT_2 = \{w_1, w_2, w_3, w_4\}$, so that $RAT = \{w_2, w_3, w_4\}$. Hence $\mathbb{B}_1 RAT = \{w_2, w_3\}$, $\mathbb{B}_2 RAT = \{w_3, w_4\}$ and $\mathbb{CB}RAT = \varnothing$. Thus at state w_3 each player makes a rational choice and believes that also the other player makes a rational choice, but it is not common belief that both players are making rational choices (indeed we have that $\mathbb{B}_1\mathbb{B}_2 RAT = \mathbb{B}_2\mathbb{B}_1 RAT = \varnothing$, that is, neither player believes that the other player believes that both players are choosing rationally).

Remark 9.3
It follows from Definition 9.2 (in particular, from the requirement that a player always knows what choice he is making) that, for every player i, $\mathbb{B}_i RAT_i = RAT_i$, that is, the set of states where player i makes a rational choice coincides with the set of state where she believes that her own choice is rational.

The central question in the literature on the epistemic foundations of game theory is: What strategy profiles are compatible with common belief of rationality? The question can be restated as follows.

Problem 9.4
Given a strategic-form game G, determine the subset $\tilde{S}$ of the set of strategy profiles S that satisfies the following properties:

(A) given an arbitrary epistemic model of G, if w is a state at which there is common belief of rationality, then the strategy profile chosen at w belongs to $\tilde{S}$: if $w \in \mathbb{CB}RAT$ then $\sigma(w) \in \tilde{S}$, and

(B) for every $s \in \tilde{S}$, there exists an epistemic model of G and a state w such that $\sigma(w) = s$ and $w \in \mathbb{CB}RAT$. ⊣

A set $\tilde{S}$ of strategy profiles that satisfies the two properties of Problem 9.4 is said to *characterize* the notion of common belief of rationality in game G.

In order to obtain an answer to Problem 9.4 we introduce the notion of strictly dominated strategy and an algorithm known as the Iterated Deletion of Strictly Dominated Strategies.

Definition 9.4
Given a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$ we say that strategy $s_i \in S_i$ of player i is *strictly dominated in* G if there is another strategy $t_i \in S_i$ of player i such that – no matter what strategies the other players choose – player i prefers the outcome associated with t_i to the outcome associated with s_i, that is, if, for all $s_{-i} \in S_{-i}$, $\pi_i(t_i, s_{-i}) > \pi_i(s_i, s_{-i})$. $\dashv$

For example, in the game of Figure 9.1a, for Player 2 strategy h is strictly dominated (by g).

Let $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$ and $G' = \left\langle \mathsf{Ag}, \{S'_i, \pi'_i\}_{i \in \mathsf{Ag}} \right\rangle$ be two games. We say that G' is a *subgame* of G if for every player i, $S'_i \subseteq S_i$ (so that $S' \subseteq S$) and π'_i is the restriction of π_i to S' (that is, for every $s' \in S'$, $\pi'_i(s') = \pi_i(s')$).

Definition 9.5
The Iterated Deletion of Strictly Dominated Strategies (IDSDS) is the following procedure. Given a game $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$ let $\langle G^0, G^1, \ldots, G^m, \ldots \rangle$ be the sequence of subgames of G defined recursively as follows. For all $i \in \mathsf{Ag}$,

1. Let $S_i^0 = S_i$ and let $D_i^0 \subseteq S_i^0$ be the set of strategies of player i that are strictly dominated in $G^0 = G$;

2. For $m \geq 1$, let $S_i^m = S_i^{m-1} \backslash D_i^{m-1}$ and let G^m be the subgame of G with strategy sets S_i^m. Let $D_i^m \subseteq S_i^m$ be the set of strategies of player i that are strictly dominated in G^m.

Let $S_i^\infty = \bigcap\limits_{m \in \mathbb{N}} S_i^m$ (where $\mathbb{N}$ denotes the set of non-negative integers) and let G^∞ be the subgame of G with strategy sets S_i^∞. Let $S^\infty = S_1^\infty \times ... \times S_n^\infty$.[7] $\dashv$

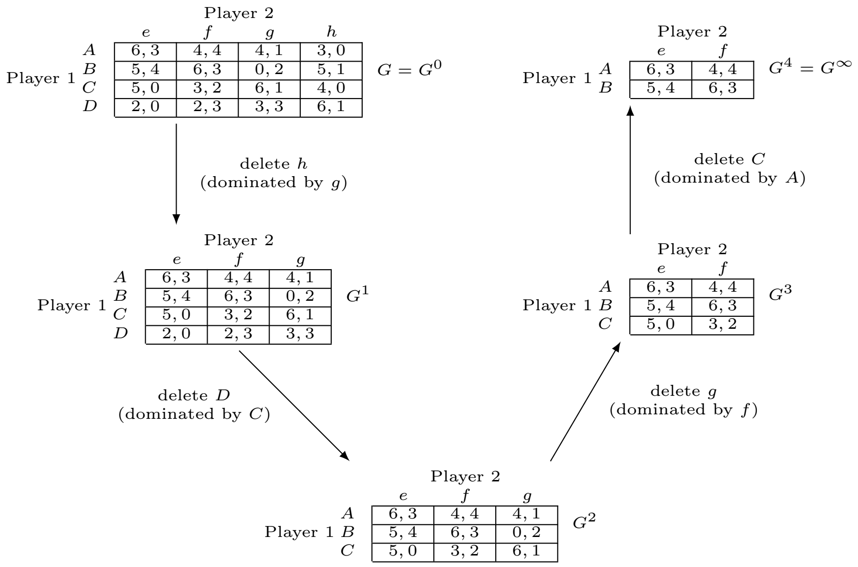
Figure 9.2 shows the application of the IDSDS procedure to the game of Figure 9.1a. In the initial game strategy h of Player 2 is strictly dominated by g; deleting h we obtain game G^1 where $S_1^1 = \{A, B, C, D\}$ and $S_2^1 = \{e, f, g\}$. In G^1 strategy D of Player 1 is strictly dominated by C; deleting D we obtain game G^2 where $S_1^2 = \{A, B, C\}$ and $S_2^2 = \{e, f, g\}$. In G^2 strategy g of Player 2 is strictly dominated by f; deleting g we obtain game G^3 where $S_1^3 = \{A, B, C\}$ and $S_2^3 = \{e, f\}$. In G^3 strategy C of Player 1 is strictly dominated by A; deleting C we obtain game G^4 where $S_1^4 = \{A, B\}$ and $S_2^4 = \{e, f\}$. In G^4 there are no strictly dominated strategies and, therefore, the procedure stops, so that $G^\infty = G^4$; thus $S_1^\infty = \{A, B\}$ and $S_2^\infty = \{e, f\}$.

The following proposition states that the answer to Problem 9.4 is provided by the output of the IDSDS procedure.

Proposition 9.1
Fix a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$ and let $S^\infty \subseteq S$ be the set of strategy profiles obtained by applying the IDSDS algorithm. Then:

(A) given an arbitrary epistemic model of G, if w is a state at which there is common belief of rationality, then the strategy profile chosen at w belongs to S^∞: if $w \in \mathbb{CB}RAT$ then $\sigma(w) \in S^\infty$, and

[7] Note that, since the strategy sets are finite, there exists an integer r such that $G^\infty = G^r = G^{r+k}$ for every $k \in \mathbb{N}$.



Player 2
Player 1
$G = G^0$
delete h (dominated by g)
G^1
delete D (dominated by C)
G^2
delete g (dominated by f)
G^3
delete C (dominated by A)
$G^4 = G^\infty$

$G = G^0$

	e	f	g	h
A	$6,3$	$4,4$	$4,1$	$3,0$
B	$5,4$	$6,3$	$0,2$	$5,1$
C	$5,0$	$3,2$	$6,1$	$4,0$
D	$2,0$	$2,3$	$3,3$	$6,1$

G^1

	e	f	g
A	$6,3$	$4,4$	$4,1$
B	$5,4$	$6,3$	$0,2$
C	$5,0$	$3,2$	$6,1$
D	$2,0$	$2,3$	$3,3$

G^2

	e	f	g
A	$6,3$	$4,4$	$4,1$
B	$5,4$	$6,3$	$0,2$
C	$5,0$	$3,2$	$6,1$

G^3

	e	f
A	$6,3$	$4,4$
B	$5,4$	$6,3$
C	$5,0$	$3,2$

$G^4 = G^\infty$

	e	f
A	$6,3$	$4,4$
B	$5,4$	$6,3$

Figure 9.2: Application of the IDSDS procedure to the game of Figure 9.1a

(B) for every $s \in S^\infty$, there exists an epistemic model of G and a state w such that $\sigma(w) = s$ and $w \in \mathbb{CB}RAT$. $\dashv$

Proof (A) Fix a game G, an epistemic model of it and a state w_0 and suppose that $w_0 \in \mathbb{CB}RAT$. We want to show that $\sigma(w_0) \in S^\infty$.

First we prove by induction that

$$\forall w \in R^+(w_0), \forall i \in \mathsf{Ag}, \forall m \geq 0,\ \sigma_i(w) \notin D_i^m \tag{9.2}$$

(recall that R^+ is the transitive closure of $\bigcup_{i \in \mathsf{Ag}} R_i$ and D_i^m is the set of strategies of player i that are strictly dominated in game G^m: see Definition 9.5).

1. Base step ($m = 0$). Fix an arbitrary $w \in R^+(w_0)$ and an arbitrary player i. If $\sigma_i(w) \in D_i^0$, then there is a strategy $\hat{s}_i \in S_i$ such that, for all $s_{-i} \in S_{-i}$, $\pi_i(\sigma_i(w), s_{-i}) < \pi_i(\hat{s}_i, s_{-i})$; thus, in particular, for all $w' \in R_i(w)$, $\pi_i(\sigma_i(w), \sigma_{-i}(w')) < \pi_i(\hat{s}_i, \sigma_{-i}(w'))$. Hence, by Definition 9.3, $w \notin RAT_i$ so that, since $RAT \subseteq RAT_i$, $w \notin RAT$, contradicting - since $w \in R^+(w_0)$ - the hypothesis that $w_0 \in \mathbb{CB}RAT$.

2. Inductive step: assume that (9.2) holds for all $k \leq m$; we want to show that it holds for $k = m + 1$. Suppose that $\forall w \in R^+(w_0), \forall i \in \mathsf{Ag}, \forall k \leq m$, $\sigma_i(w) \notin D_i^k$. Then (see Definition 9.5)

$$\forall w \in R^+(w_0), \sigma(w) \in S^{m+1}. \tag{9.3}$$

Fix an arbitrary $w \in R^+(w_0)$ and an arbitrary player i and suppose that $\sigma_i(w) \in D_i^{m+1}$. Then, by definition of D_i^{m+1} (see Definition 9.5) there is a strategy $\hat{s}_i \in S_i$ such that, for all $s_{-i} \in S_{-i}^{m+1}$, $\pi_i(\sigma_i(w), s_{-i}) < \pi_i(\hat{s}_i, s_{-i})$. By transitivity of R^+, since $w \in R^+(w_0)$, $R^+(w) \subseteq R^+(w_0)$. Thus, by (9.3) and the fact that

$R_i(w) \subseteq R^+(w)$, we have that $\pi_i(\sigma_i(w), \sigma_{-i}(w')) < \pi_i(\hat{s}_i, \sigma_{-i}(w'))$ for all $w' \in R_i(w)$, so that, by Definition 9.3, $w \notin RAT_i$, contradicting the hypothesis that $w_0 \in \mathbb{CB}RAT$.

Thus (9.2) holds and therefore, by Definition 9.5,

$$\forall w \in R^+(w_0), \forall i \in \mathsf{Ag},\ \sigma_i(w) \in S_i^\infty. \tag{9.4}$$

The proof is not yet complete, since it may be the case that $w_0 \notin R^+(w_0)$. Fix an arbitrary player i and an arbitrary $w \in R_i(w_0)$ (recall the assumption that R_i is serial). By definition of epistemic model (see Definition 9.2) $\sigma_i(w_0) = \sigma_i(w)$. By (9.4) $\sigma_i(w) \in S_i^\infty$. Thus $\sigma_i(w_0) \in S_i^\infty$ and hence $\sigma(w_0) \in S^\infty$.

(B) Construct the following epistemic model of game G: $W = S^\infty$ and, for every player i and every $s \in S^\infty$ let $R_i(s) = \{s' \in S^\infty : s'_i = s_i\}$. Then R_i is an equivalence relation (hence serial, transitive and euclidean). For all $s \in S^\infty$, let $\sigma_i(s) = s_i$. Fix an arbitrary $s \in S^\infty$ and an arbitrary player i. By definition of S^∞, it is not the case that there exists an $\hat{s}_i \in S_i$ such that, for all $s_{-i} \in S_{-i}^\infty$, $\pi_i(s_i, s_{-i}) < \pi_i(\hat{s}_i, s_{-i})$. Thus, since - by construction - for all $s' \in R_i(s)$, $\sigma_{-i}(s') \in S_{-i}^\infty$, $s \in RAT_i$ (see Definition 9.3). Since i was chosen arbitrarily, $s \in RAT$; hence, since $s \in S^\infty$ was chosen arbitrarily, $RAT = S^\infty$. It follows that $s \in \mathbb{CB}RAT$ for every $s \in S^\infty$. $\dashv$

9.4 Syntactic Characterization of Common Belief of Rationality

We now turn to the syntactic analysis of rationality in strategic-form games. In order to be able to describe a game syntactically, the set of propositional variables (or atoms) At will be taken to include:

- Strategy symbols s_i^1, s_i^2, ... The intended interpretation of s_i^k is "player i chooses her k^{th} strategy s_i^k".[8]

- Atoms of the form $s_i^\ell \succeq_i s_i^k$, whose intended interpretation is "strategy s_i^ℓ of player i is at least as good, for player i, as her strategy s_i^k", and atoms of the form $s_i^\ell \succ_i s_i^k$, whose intended interpretation is "for player i strategy s_i^ℓ is better than strategy s_i^k".

Fix a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$ and let $S_i = \{s_i^1, s_i^2, ..., s_i^{m_i}\}$ (thus the cardinality of S_i is m_i). We denote by $\mathbf{KD45}_G$ the **KD45** multi-agent logic *without a common belief operator* that satisfies the following additional axioms: for all $i \in \mathsf{Ag}$ and for all $k, \ell = 1, ..., m_i$, with $k \neq \ell$,

[8] Thus, with slight abuse of notation, we use the symbol s_i^k to denote both an element of S_i, that is, a strategy of player i, and an element of At, that is, an atom whose intended interpretation is "player i chooses strategy s_i^k".

$$\begin{array}{ll}
\left(s_i^1 \vee s_i^2 \vee ... \vee s_i^{m_i}\right) & (\mathbf{G1}) \\
\neg(s_i^k \wedge s_i^\ell) & (\mathbf{G2}) \\
s_i^k \rightarrow B_i s_i^k & (\mathbf{G3}) \\
\left(s_i^k \succeq_i s_i^\ell\right) \vee \left(s_i^\ell \succeq_i s_i^k\right) & (\mathbf{G4}) \\
\left(s_i^\ell \succ_i s_i^k\right) \leftrightarrow \left(\left(s_i^\ell \succeq_i s_i^k\right) \wedge \neg\left(s_i^k \succeq_i s_i^\ell\right)\right) & (\mathbf{G5})
\end{array}$$

Axiom **G1** says that player i chooses at least one strategy, while axiom **G2** says that player i cannot choose more than one strategy. Thus **G1** and **G2** together imply that each player chooses exactly one strategy. Axiom **G3**, on the other hand, says that player i is conscious of his own choice: if he chooses strategy s_i^k then he believes that he chooses s_i^k. The remaining axioms state that the ordering of strategies is complete (**G4**) and that the corresponding strict ordering is defined as usual (**G5**).[9]

Proposition 9.2
The following is a theorem of logic $\mathbf{KD45}_G$: $B_i s_i^k \rightarrow s_i^k$. That is, every player has correct beliefs about her own choice of strategy.[10] $\dashv$

Proof In the following PL stands for 'Propositional Logic' and RK denotes the inference rule "from $\psi \rightarrow \chi$ infer $\Box\psi \rightarrow \Box\chi$", which is a derived rule of inference that applies to every modal operator $\Box$ that satisfies axiom **K** and the rule of Necessitation. Fix a player i and $k, \ell \in \{1, ..., m_i\}$ with $k \neq \ell$. Let φ_k denote the formula

$$(s_i^1 \vee ... \vee s_i^{m_i}) \wedge \neg s_i^1 \wedge ... \wedge \neg s_i^{k-1} \wedge \neg s_i^{k+1} \wedge ... \wedge \neg s_i^{m_i}.$$

1.	$\varphi_k \rightarrow s_i^k$	tautology
2.	$\neg(s_i^k \wedge s_i^\ell)$	axiom **G2** (for $\ell \neq k$)
3.	$s_i^k \rightarrow \neg s_i^\ell$	2, PL
4.	$B_i s_i^k \rightarrow B_i \neg s_i^\ell$	3, rule RK
5.	$B_i \neg s_i^\ell \rightarrow \neg B_i s_i^\ell$	axiom $\mathbf{D}_i$
6.	$s_i^\ell \rightarrow B_i s_i^\ell$	axiom **G3**
7.	$\neg B_i s_i^\ell \rightarrow \neg s_i^\ell$	6, PL
8.	$B_i s_i^k \rightarrow \neg s_i^\ell$	4, 5, 7, PL (for $\ell \neq k$)
9.	$s_i^1 \vee ... \vee s_i^{m_i}$	axiom **G1**
10.	$B_i s_i^k \rightarrow (s_i^1 \vee ... \vee s_i^{m_i})$	9, PL
11.	$B_i s_i^k \rightarrow \varphi_k$	8 (for every $\ell \neq k$), 10, PL
12.	$B_i s_i^k \rightarrow s_i^k$	1, 11, PL.

$\dashv$

Given a game G, let $\mathcal{F}_G$ denote the set of epistemic models of G (see Definition 9.2).

[9] We have not included the axiom corresponding to transitivity of the ordering, namely $\left(s_i^{k_1} \succeq_i s_i^{k_2}\right) \wedge \left(s_i^{k_2} \succeq_i s_i^{k_3}\right) \rightarrow \left(s_i^{k_1} \succeq_i s_i^{k_3}\right)$, because it is not needed in what follows.

[10] Note that, in general, logic $\mathbf{KD45}_G$ allows for incorrect beliefs. In particular, a player might have incorrect beliefs about the choices made by *other* players. By Proposition 9.2, however, a player cannot have mistaken beliefs about her own choice.

Definition 9.6
Given a game G and an epistemic model $F \in \mathcal{F}_G$ a *syntactic model of G based on F* is obtained by adding to F any propositional valuation $V : W \to (\mathsf{At} \to \{true, false\})$ that satisfies the following restrictions (we write $w \models p$ instead of $V(w)(p) = true$):

- $w \models s_i^h$ if and only if $\sigma_i(w) = s_i^h$,
- $w \models (s_i^k \succeq_i s_i^\ell)$ if and only if $\pi_i(s_i^k, \sigma_{-i}(w)) \geq \pi_i(s_i^\ell, \sigma_{-i}(w))$,
- $w \models s_i^k \succ_i s_i^\ell$ if and only if $\pi_i(s_i^k, \sigma_{-i}(w)) > \pi_i(s_i^\ell, \sigma_{-i}(w))$.

Thus, in a syntactic model of a game, at state w it is true that player i chooses strategy s_i^h if and only if the strategy of player i associated with w (in the semantic model on which the syntactic model is based) is s_i^h (that is, $\sigma_i(w) = s_i^h$) and it is true that strategy s_i^k is at least as good as (respectively, better than) strategy s_i^ℓ if and only if s_i^k in combination with $\sigma_{-i}(w)$ (the profile of strategies of players other than i associated with w) yields an outcome which player i considers at least as good as (respectively, better than) the outcome yielded by s_i^ℓ in combination with $\sigma_{-i}(w)$.

For example, a syntactic model of the game shown in Part a of Figure 9.1 based on the semantic model shown in Part b of Figure 9.1 satisfies the following formula at state w_1:

$$B \wedge e \wedge (A \succ_1 B) \wedge (A \succ_1 C) \wedge (A \succ_1 D) \wedge (B \succeq_1 C) \wedge (C \succeq_1 B) \wedge (B \succ_1 D)$$
$$\wedge (C \succ_1 D) \wedge (e \succ_2 f) \wedge (e \succ_2 g) \wedge (e \succ_2 h) \wedge (f \succ_2 g) \wedge (f \succ_2 h) \wedge (g \succ_2 h).$$

Remark 9.5
Let $\mathcal{M}_G$ denote the set of all syntactic models of game G. It is straightforward to verify that logic $\mathbf{KD45}_G$ is sound with respect to $\mathcal{M}_G$.[11] ⊣

We now provide an axiom that, for every game, characterizes the output of the IDSDS procedure (see Definition 9.5), namely the set of strategy profiles S^∞. The following axiom says that if player i chooses strategy s_i^k then it is not the case that she believes that a different strategy s_i^ℓ is better for her:

$$s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k). \qquad (\mathbf{WR})$$

[11] It follows from the following observations: (1) axioms **G1** and **G2** are valid in every syntactic model because, for every state w, there is a unique strategy $s_i^k \in S_i$ such that $\sigma_i(w) = s_i^k$ and, by the validation rules (see Definition 9.6), $w \models s_i^k$ if and only if $\sigma_i(w) = s_i^k$; (2) axiom **G3** is an immediate consequence of the fact (see Definition 9.2) that if $w' \in R_i(w)$ then $\sigma_i(w') = \sigma_i(w)$; (3) axioms **G4** and **G5** are valid because, for every state w, there is a unique profile of strategies $\sigma_{-i}(w)$ of the players other than i and the payoff function π_i of player i restricted to the set $S_i \times \{\sigma_{-i}(w)\}$ induces a complete (and transitive) ordering of S_i.

Proposition 9.3
Fix a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$. Then

(A) If $M = \langle W, \{R_i\}_{i \in \mathsf{Ag}}, \{\sigma_i\}_{i \in \mathsf{Ag}}, V \rangle$ is a syntactic model of G that validates axiom **WR**, then $\sigma(w) \in S^\infty$, for every state $w \in W$.

(B) There exists a syntactic model M of G that validates axiom **WR** and is such that (1) for every $s \in S^\infty$, there exists a state w such that $w \models s$, and (2) for every $s \in S$ and for every $w \in W$, if $w \models s$ then $\sigma(w) \in S^\infty$. $\dashv$

Proof (A) Fix a game and a syntactic model of it that validates axiom **WR**. Fix an arbitrary state w_0 and an arbitrary player i. By Axioms **G1** and **G2** (see Remark 9.5) $w_0 \models s_i^k$ for a unique strategy $s_i^k \in S_i$. Fix an arbitrary $s_i^\ell \in S_i$, with $s_i^\ell \neq s_i^k$. Since the model validates axiom **WR**, $w_0 \models \neg B_i(s_i^\ell \succ_i s_i^k)$, that is, there exists a $w_1 \in R_i(w_0)$, such that $w_1 \models \neg(s_i^\ell \succ_i s_i^k)$. Hence, by Definition 9.6, $\sigma_i(w_0) = s_i^k$ and $\pi_i(s_i^k, \sigma_{-i}(w_1)) \geq \pi_i(s_i^\ell, \sigma_{-i}(w_1))$, so that, by Definition 9.3, $w_0 \in RAT_i$. Since w_0 and i were chosen arbitrarily, $RAT = W$ and thus, $\mathbb{CB}RAT = W$, that is, for every $w \in W$, $w \in \mathbb{CB}RAT$. Hence, by Part A of Proposition 9.1, $\sigma(w) \in S^\infty$.

(B) Let F be the semantic epistemic model constructed in the proof of Part B of Proposition 9.1 and let M be a syntactic model based on F that satisfies the validation rules of Definition 9.6. First we show that M validates axiom **WR**. Recall that, in F, $W = S^\infty$, $s' \in R_i(s)$ if and only if $s_i = s'_i$ and σ is the identity function. Fix an arbitrary player i and an arbitrary state $\hat{s}$. We need to show that, for every $s_i^\ell \in S_i$, $\hat{s} \models \neg B_i(s_i^\ell \succ_i \hat{s}_i)$. Suppose that, for some $s_i^\ell \in S_i$, $\hat{s} \models B_i(s_i^\ell \succ_i \hat{s}_i)$, that is, for every $s' \in R_i(\hat{s})$, $s' \models (s_i^\ell \succ_i \hat{s}_i)$. Then, by Definition 9.6, for every $s' \in R_i(\hat{s})$, $\pi_i(s_i^\ell, s'_{-i}) > \pi_i(\hat{s}_i, s'_{-i})$, so that, by Definition 9.3, $\hat{s} \notin RAT_i$. But, as shown in the proof of Proposition 9.1, $RAT = S^\infty$ so that, since $RAT \subseteq RAT_i \subseteq W = S^\infty$, $RAT_i = S^\infty$, yielding a contradiction. Thus M validates axiom **WR**. Now fix an arbitrary $s \in S^\infty$. Then, by Definition 9.6, $s \models s$; thus (1) holds; conversely, let $s \models s$; then, by construction of F, $\sigma(s) = s$ and $s \in S^\infty$. Thus (2) holds. $\dashv$

Remark 9.6
Since, by Proposition 9.1, the set of strategy-profiles S^∞ characterizes the semantic notion of common belief of rationality, it follows from Proposition 9.3 that axiom **WR** provides a syntactic characterization of common belief or rationality in strategic-form games with ordinal payoffs. $\dashv$

Remark 9.7
Note that axiom **WR** provides a syntactic characterization of common belief of rationality in a logic that does *not* contain a common belief operator. However, since **WR** expresses the notion that player i chooses rationally, by the Necessitation rule every player believes that player i is rational [that is, from **WR** we obtain that, for every player $j \in \mathsf{Ag}$, $B_j\left(s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k)\right)$ is a theorem], and every player believes this [from $B_j\left(s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k)\right)$, by Necessitation, we get that $B_r B_j\left(s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k)\right)$ is a theorem, for every player $r \in \mathsf{Ag}$] and so on, so that - essentially - the rationality of every player's choice is commonly be-

lieved. Indeed, if one adds the common belief operator CB to the logic, then, by Necessitation, $CB\left(s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k)\right)$ becomes a theorem.[12] $\dashv$

Remark 9.8
There appears to be an important difference between the result of Section 9.3 and the result of this section: Proposition 9.1 gives a *local* result, while Proposition 9.3 provides a *global* one. For example, Part A of Proposition 9.1 states that if *at a state* there is common belief of rationality, then the strategy profile played *at that state* belongs to S^∞, while Part A of Proposition 9.3 states that in a syntactic model that validates axiom **WR** the strategy profile played *at every state* belongs to S^∞. As a matter of fact, the result of Section 9.3 is also "global" in nature. To see this, fix an epistemic model and a state w_0 and suppose that $w_0 \in \mathbb{CB}RAT$. By transitivity of R^+ (see Footnote 6) $\mathbb{CB}RAT \subseteq \mathbb{CB}\ \mathbb{CB}RAT$. Thus, for every $w \in R^+(w_0)$, $w \in \mathbb{CB}RAT$. Hence, by Proposition 9.1, $\sigma(w) \in S^\infty$. That is, if at a state there is common belief of rationality, then at that state, *as well as at all states reachable from it by the common belief relation* R^+, it is true that the strategy profile played belongs to S^∞. This is essentially a global result, since from the point of view of a state w_0, the "global" space is precisely the set $R^+(w_0)$. $\dashv$

9.5 Common Belief versus Common Knowledge

In the previous two sections we studied the implications of common *belief* of rationality in strategic-form games. What distinguishes belief from knowledge is that belief may be erroneous, while knowledge is veridical: if I know that φ then φ is true, while it is possible for me to believe that φ when φ is in fact false. In a game a player might have erroneous beliefs about the choices of the other players or about their beliefs. Perhaps one might be able to draw sharper conclusions about what the players will do in a game if one rules out erroneous beliefs. Thus a natural question to ask is: If we replace belief with knowledge, what can we infer from the hypothesis that there is *common knowledge* of rationality? Is the set of strategy profiles that are compatible with common knowledge of rationality a proper subset of S^∞? The answer is negative as can be seen from the epistemic model constructed in the proof of Part B of Proposition 9.1: that model is one where each accessibility relation is an equivalence relation and thus the underlying frame is an $\mathcal{S}5$ frame. Hence the set of strategy profiles that are compatible with common knowledge of rationality coincides with the set of strategy profiles that are compatible with common belief of rationality, namely S^∞. However, it is possible to obtain sharper predictions by replacing belief with knowledge and, at the same time, introducing a mild strengthening of the notion of rationality. Given a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i\in\mathsf{Ag}} \right\rangle$ we will now consider epistemic models of G of the form $\langle W, \{\sim_i\}_{i\in\mathsf{Ag}}, \{\sigma_i\}_{i\in\mathsf{Ag}}\rangle$ where

[12] Despite the fact that the intuitive definition of common belief involves an infinite conjunction (see Footnote 5), there is a finite axiomatization of common belief. For example, the following three axioms are sufficient (without any additional rule of inference: see Section 9.10 for a reference): (1) $CB\varphi \to B_i\varphi$, (2) $CB\varphi \to B_iCB\varphi$ and (3) $CB\left(\varphi \to B_1\varphi \wedge \cdots \wedge B_n\varphi\right) \to (B_1\varphi \wedge \cdots \wedge B_n\varphi \to CB\varphi)$.

$\langle W, \{\sim_i\}_{i \in \mathsf{Ag}} \rangle$ is an $\mathcal{S}5$ Kripke frame, that is, the accessibility relation $\sim_i$ of each player $i \in \mathsf{Ag}$ is an *equivalence* relation. Since we are dealing with $\mathcal{S}5$ frames, instead of belief we will speak of knowledge and denote the semantic operators for individual knowledge and common knowledge by $\mathbb{K}_i$ and $\mathbb{CK}$, respectively. Thus $\mathbb{K}_i : 2^W \to 2^W$ and $\mathbb{CK} : 2^W \to 2^W$ are given by:

$$\begin{aligned} \mathbb{K}_i E &= \{w \in W : \ \sim_i (w) \subseteq E\}, \text{ and} \\ \mathbb{CK}E &= \{w \in W : \ \sim^* (w) \subseteq E\} \end{aligned} \tag{9.5}$$

where, as before, $\sim_i (w) = \{w' \in W : w \sim_i w'\}$ and $\sim^*$ is the transitive closure of $\bigcup_{i \in \mathsf{Ag}} \sim_i$.[13] $\mathbb{K}_i E$ is interpreted as the event that (that is, the set of states at which) player i knows event E and $\mathbb{CK}E$ as the event that E is commonly known.

We now consider a stronger notion of rationality than the one given in Definition 9.3, which we will call *s-rationality* ('s' stands for 'strong').

Definition 9.7
Fix a strategic-form game G and an $\mathcal{S}5$ epistemic model of G. At state w player i's strategy $\sigma_i(w)$ is *s-rational* if it is not the case that there is another strategy $s_i' \in S_i$ which (1) yields *at least as high* a payoff as $\sigma_i(w)$ against *all* the strategy profiles of the other players that player i considers possible and (2) a higher payoff than $\sigma_i(w)$ against *at least one* strategy profile of the other players that player i considers possible, that is, if

there is no strategy $s_i' \in S_i$ such that
(1) $\pi_i\left(s_i', \sigma_{-i}(w')\right) \geq \pi_i\left(\sigma_i(w), \sigma_{-i}(w')\right)$, $\forall w' \in \ \sim_i (w)$, and
(2) $\pi_i\left(s_i', \sigma_{-i}(\tilde{w})\right) > \pi_i\left(\sigma_i(w), \sigma_{-i}(\tilde{w})\right)$, for some $\tilde{w} \in \ \sim_i (w)$.

[recall that, by Definition 9.2, the function $\sigma_i(\cdot)$ is constant on the set $\sim_i (w)$]. Equivalently, $\sigma_i(w)$ is s-rational at state w if, for every $s_i' \in S_i$, whenever there is a $w' \in \ \sim_i (w)$ such that $\pi_i\left(s_i', \sigma_{-i}(w')\right) > \pi_i\left(\sigma_i(w), \sigma_{-i}(w')\right)$ then there is another state $w'' \in \ \sim_i (w)$ such that $\pi_i\left(\sigma_i(w), \sigma_{-i}(w'')\right) > \pi_i\left(s_i', \sigma_{-i}(w'')\right)$. $\dashv$

Denote by $SRAT_i$ the event that (i.e. the set of states at which) player i's choice is s-rational and let $SRAT = \bigcap_{i \in \mathsf{Ag}} SRAT_i$. Then $SRAT$ is the event that the choice of every player is s-rational.

As we did in Section 9.3 for the weaker notion of rationality and for common belief, we will now determine, for every game G, the set of strategy profiles that are compatible with common knowledge of s-rationality. Also in this case, the answer is based on an iterated deletion procedure. However, unlike the IDSDS procedure given in Definition 9.5, the deletion procedure defined below operates not at the level of individual players' strategies but at the level of strategy profiles.

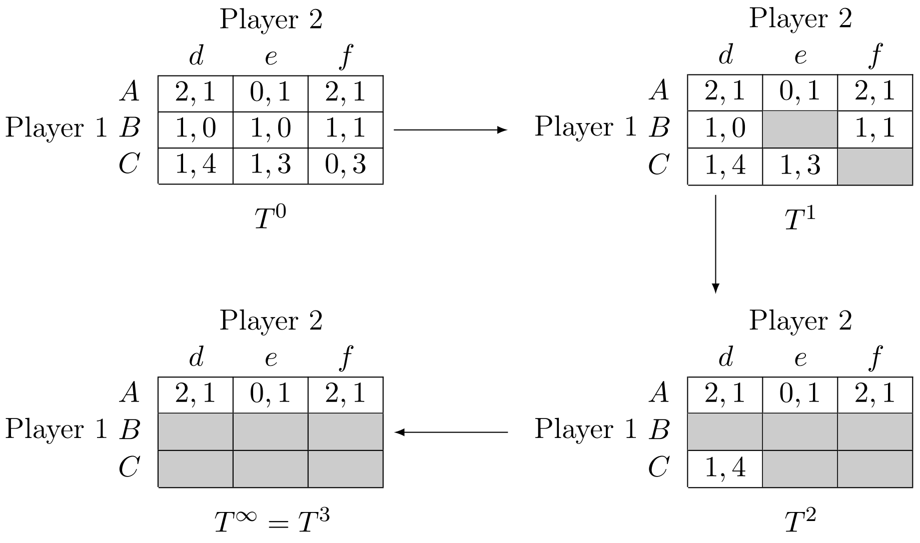
Definition 9.8
Given a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$, a subset of strategy profiles $X \subseteq S$ and a strategy profile $x \in X$, we say that x is *inferior*

[13] Thus, in addition to the properties listed in Footnote 6, the operator $\mathbb{K}_i$ satisfies the veridicality property $\mathbb{K}_i E \subseteq E, \forall E \subseteq W$ (because of reflexivity of $\sim_i$). Since reflexivity is inherited by $\sim^*$, also the common knowledge operator satisfies the veridicality property: $\mathbb{CK}E \subseteq E$.

relative to X if there exists a player i and a strategy $s_i \in S_i$ of player i (thus s_i need not belong to the projection of X onto S_i) such that:

1. $\pi_i(s_i, x_{-i}) > \pi_i(x_i, x_{-i})$, and
2. for all $s_{-i} \in S_{-i}$, if $(x_i, s_{-i}) \in X$ then $\pi_i(s_i, s_{-i}) \geq \pi_i(x_i, s_{-i})$.

The *Iterated Deletion of Inferior Profiles* (IDIP) is defined as follows. For $m \in \mathbb{N}$ define $T^m \subseteq S$ recursively as follows: $T^0 = S$ and, for $m \geq 1$, $T^m = T^{m-1} \backslash I^{m-1}$, where $I^{m-1} \subseteq T^{m-1}$ is the set of strategy profiles that are inferior relative to T^{m-1}. Let $T^\infty = \bigcap\limits_{m \in \mathbb{N}} T^m$.[14] $\dashv$



Player 2

		d	e	f
	A	$2,1$	$0,1$	$2,1$
Player 1	B	$1,0$	$1,0$	$1,1$
	C	$1,4$	$1,3$	$0,3$

T^0

Player 2

		d	e	f
	A	$2,1$	$0,1$	$2,1$
Player 1	B	$1,0$		$1,1$
	C	$1,4$	$1,3$	

T^1

Player 2

		d	e	f
	A	$2,1$	$0,1$	$2,1$
Player 1	B			
	C	$1,4$		

T^2

Player 2

		d	e	f
	A	$2,1$	$0,1$	$2,1$
Player 1	B			
	C			

$T^\infty = T^3$

Figure 9.3: Illustration of the IDIP procedure

The IDIP procedure is illustrated in Figure 9.3, where

T^0 $= \{(A,d),(A,e),(A,f),(B,d),(B,e),(B,f),(C,d),(C,e),\ (C,f)\}$.
This equals S.

I^0 $= \{(B,e),(C,f)\}$ (the elimination of (B,e) is done through Player 2 and strategy f, while the elimination of (C,f) is done through Player 1 and strategy B);

T^1 $= \{(A,d),(A,e),(A,f),(B,d),(B,f),(C,d),(C,e)\}$,

I^1 $= \{(B,d),(B,f),(C,e)\}$ (the elimination of (B,d) and (B,f) is now done through Player 1 and strategy A, while the elimination of (C,e) is done through Player 2 and strategy d);

T^2 $= \{(A,d),(A,e),(A,f),(C,d)\}$,

[14] Since the strategy sets are finite, there exists an integer r such that $T^\infty = T^r = T^{r+k}$ for every $k \in \mathbb{N}$.

$I^2 = \{(C,d)\}$ (the elimination of (C,d) is done through Player 1 and strategy A);

$T^3 = \{(A,d),(A,e),(A,f)\}$,

$I^3 = \varnothing$; thus

$T^\infty = T^3$.

The following Proposition is the counterpart to Proposition 9.1, when rationality is replaced with s-rationality, belief with knowledge and the IDSDS procedure with the IDIP procedure.

Proposition 9.4
Fix a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i\in\mathsf{Ag}} \right\rangle$ and let $T^\infty \subseteq S$ be the set of strategy profiles obtained by applying the IDIP procedure. Then:

(A) given an arbitrary $\mathcal{S}5$ epistemic model of G, if w is a state at which there is common knowledge of s-rationality, then the strategy profile chosen at w belongs to T^∞: if $w \in \mathbb{CK}SRAT$ then $\sigma(w) \in T^\infty$, and

(B) for every $s \in T^\infty$, there exists an $\mathcal{S}5$ epistemic model of G and a state w such that $\sigma(w) = s$ and $w \in \mathbb{CK}SRAT$. $\dashv$

Proof (A) Fix an $\mathcal{S}5$ epistemic model of G and a state w_0 and suppose that $w_0 \in \mathbb{CK}SRAT$. We want to show that $\sigma(w_0) \in T^\infty$.

First we prove by induction that

$$\forall w \in W \text{ such that } w_0 \sim^* w, \forall m \geq 0,\ \sigma(w) \notin I^m. \tag{9.6}$$

1. Base step ($m = 0$). Fix an arbitrary $w_1 \in W$ such that $w_0 \sim^* w_1$. If $\sigma(w_1) \in I^0$ (that is, $\sigma(w_1)$ is inferior relative to the entire set of strategy profiles S) then there exist a player i and a strategy $\hat{s}_i \in S_i$ such that, $\pi_i(\hat{s}_i, \sigma_{-i}(w_1)) > \pi_i(\sigma_i(w_1), \sigma_{-i}(w_1))$, and, for every $s_{-i} \in S_{-i}$, $\pi_i(\hat{s}_i, s_{-i}) \geq \pi_i(\sigma_i(w_1), s_{-i})$; thus, in particular, for all w' such that $w_1 \sim_i w'$, $\pi_i(\hat{s}_i, \sigma_{-i}(w')) \geq \pi_i(\sigma_i(w_1), \sigma_{-i}(w'))$. Furthermore, by reflexivity of $\sim_i$, $w_1 \sim_i w_1$. It follows from Definition 9.7 that $w_1 \notin SRAT_i$, so that, since $SRAT \subseteq SRAT_i$, $w_1 \notin SRAT$, contradicting the hypothesis that $w_0 \in \mathbb{CK}SRAT$ (since $w_0 \sim^* w_1$).

2. Inductive step: assume that (9.6) holds for all $k \leq m$; we want to show that it holds for $k = m+1$. Suppose that $\forall w \in W$ such that $w_0 \sim^* w$,$\forall k \leq m$, $\sigma(w) \notin I^k$. Then

$$\forall w \in W \text{ such that } w_0 \sim^* w, \sigma(w) \in T^{m+1}. \tag{9.7}$$

Fix an arbitrary $w_1 \in W$ such that $w_0 \sim^* w_1$ and suppose that $\sigma(w_1) \in I^{m+1}$, that is, $\sigma(w_1)$ is inferior relative to T^{m+1}. Then, by definition of I^{m+1}, there exist a player i and a strategy $\hat{s}_i \in S_i$ such that, $\pi_i(\hat{s}_i, \sigma_{-i}(w_1)) > \pi_i(\sigma_i(w_1), \sigma_{-i}(w_1))$ and, for every $s_{-i} \in S_{-i}$, if $(\hat{s}_i, s_{-i}) \in T^{m+1}$ then $\pi_i(\hat{s}_i, s_{-i}) \geq \pi_i(\sigma_i(w_1), s_{-i})$. By Definition 9.2, for every w such that $w \sim_i w_1$, $\sigma_i(w) = \sigma_i(w_1)$ and by (9.7), for every w such that $w_0 \sim^* w$, we have that $(\sigma_i(w), \sigma_{-i}(w)) \in T^{m+1}$. Thus, since $\sim_i(w_1) \subseteq\ \sim^*(w_1) \subseteq\ \sim^*(w_0)$, we have that, for every w such that $w \sim_i w_1$,

$(\sigma_i(w_1), \sigma_{-i}(w)) \in T^m$. By reflexivity of $\sim_i$, $w_1 \sim_i w_1$; hence, by Definition 9.7, $w_1 \notin SRAT_i$ and thus $w_1 \notin SRAT$ (since $SRAT \subseteq SRAT_i$). This, together with the fact that $w_0 \sim^* w_1$, contradicts the hypothesis that $w_0 \in \mathbb{CK}SRAT$.

Thus, we have shown by induction that, $\forall w \in W$ such that $w \sim^* w_0$, $\sigma(w) \in \bigcap\limits_{m \in \mathbb{N}} T^m = T^\infty$. It only remains to establish that $\sigma(w_0) \in T^\infty$, but this follows from reflexivity of $\sim^*$.

(B) Construct the following epistemic model of game G: $W = T^\infty$ and, for every player i and every $s, s' \in T^\infty$ let $s \sim_i s'$ if and only if $s'_i = s_i$ Then $\sim_i$ is an equivalence relation and thus the frame is an $\mathcal{S}5$ frame. For all $s \in T^\infty$, let $\sigma(s) = s$. Fix an arbitrary $\tilde{s} \in T^\infty$ and an arbitrary player i. By definition of T^∞, it is not the case that there exists an $\hat{s}_i \in S_i$ such that $\pi_i(\hat{s}_i, \tilde{s}_{-i}) > \pi_i(\tilde{s}_i, \tilde{s}_{-i})$ and, for every $s'_{-i} \in S_{-i}$, if $(\hat{s}_i, s'_{-i}) \in T^\infty$ then $\pi_i(\hat{s}_i, s'_{-i}) \geq \pi_i(\tilde{s}_i, s'_{-i})$. Thus $\tilde{s} \in SRAT_i$; hence, since player i was chosen arbitrarily, $\tilde{s} \in SRAT$. Since $\tilde{s}$ was chosen arbitrarily, it follows that $SRAT = T^\infty$ and thus $\mathbb{CK}SRAT = T^\infty$. ⊣

We now turn to the syntactic analysis. Given a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i \in \mathsf{Ag}} \right\rangle$, let $\mathbf{S5}_G$ be the **S5** multi-agent logic *without a common knowledge operator* that satisfies axioms **G1-G5** of Section 9.4. Clearly, $\mathbf{S5}_G$ is an extension of $\mathbf{KD45}_G$. Let $\mathcal{M}_G^{S5}$ denote the set of all syntactic models of game G (see Definition 9.6) based on $\mathcal{S}5$ epistemic models of G. It is straightforward to verify that logic $\mathbf{S5}_G$ is sound with respect to $\mathcal{M}_G^{S5}$.

In parallel to the analysis of Section 9.4, we now provide an axiom that, for every game, characterizes the output of the IDIP procedure, namely the set of strategy profiles T^∞. The following axiom is a strengthening of axiom **WR** of Section 9.4: it says that if player i chooses strategy s_i^k then it is not the case that (1) she believes that a different strategy s_i^ℓ is at least as good for her as s_i^k and (2) she considers it possible that s_i^ℓ is better than s_i^k:

$$s_i^k \to \neg \left(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg (s_i^\ell \succ_i s_i^k) \right). \qquad (\mathbf{SR})$$

The following proposition confirms that axiom **SR** is a strengthening of axiom **WR**: the latter is derivable in the logic obtained by adding **SR** to $\mathbf{KD45}_G$.

Proposition 9.5
Axiom **WR** is a theorem of $\mathbf{KD45}_G + \mathbf{SR}$. ⊣

Proof

1.	$s_i^k \to \neg \left(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg (s_i^\ell \succ_i s_i^k) \right)$	**SR**
2.	$(s_i^\ell \succ_i s_i^k) \leftrightarrow (s_i^\ell \succeq_i s_i^k) \wedge \neg (s_i^k \succeq_i s_i^\ell)$	**G5**
3.	$(s_i^\ell \succ_i s_i^k) \to (s_i^\ell \succeq_i s_i^k)$	2, PL
4.	$B_i(s_i^\ell \succ_i s_i^k) \to B_i(s_i^\ell \succeq_i s_i^k)$	3, RK
5.	$B_i(s_i^\ell \succ_i s_i^k) \to \neg B_i \neg (s_i^\ell \succ_i s_i^k)$	Axiom $\mathbf{D}_i$
6.	$B_i(s_i^\ell \succ_i s_i^k) \to \left(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg (s_i^\ell \succ_i s_i^k) \right)$	4, 5, PL
7.	$\neg \left(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg (s_i^\ell \succ_i s_i^k) \right) \to \neg B_i(s_i^\ell \succ_i s_i^k)$	6, PL
9.	$s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k)$	1, 7, PL

⊣

The following proposition is the counterpart to Proposition 9.3: it shows that - when belief is replaced with knowledge - axiom **SR** provides a syntactic characterization of the output of the IDIP procedure (namely, the set of strategy-profiles T^∞) and thus, by Proposition 9.4, provides a syntactic characterization of common knowledge of s-rationality in strategic-form games with ordinal payoffs.

Proposition 9.6
Fix a strategic-form game with ordinal payoffs $G = \left\langle \mathsf{Ag}, \{S_i, \pi_i\}_{i\in\mathsf{Ag}} \right\rangle$. Then

(A) If $M = \langle W, \{\sim_i\}_{i\in\mathsf{Ag}}, \{\sigma_i\}_{i\in\mathsf{Ag}}, V\rangle$ is an $\mathcal{S}5$ syntactic model of G that validates axiom **SR**, then $\sigma(w) \in T^\infty$, for every state $w \in W$.

(B) There exists an $\mathcal{S}5$ syntactic model M of G that validates axiom **SR** and is such that (1) for every $s \in T^\infty$, there exists a state w in M such that $w \models s$, and (2) for every $s \in S$ and for every $w \in W$, if $w \models s$ then $\sigma(w) \in T^\infty$. $\dashv$

Proof
To stress the fact that we are dealing with $\mathcal{S}5$ models, we shall use the operator K_i (knowledge) instead of B_i (belief).

(A) Fix a game and an $\mathcal{S}5$ syntactic model of it that validates axiom **SR**. Fix an arbitrary state w_0 and an arbitrary player i. By Axioms **G1** and **G2** (see Remark 9.5) $w_0 \models s_i^k$ for a unique strategy $s_i^k \in S_i$. Fix an arbitrary $s_i^\ell \in S_i$, with $s_i^\ell \neq s_i^k$. Since the model validates axiom **SR**, $w_0 \models \neg\left(K_i(s_i^\ell \succeq_i s_i^k) \wedge \neg K_i \neg(s_i^\ell \succ_i s_i^k)\right)$, that is (since the formula $\neg\left(K_i(s_i^\ell \succeq_i s_i^k) \wedge \neg K_i \neg(s_i^\ell \succ_i s_i^k)\right)$ is propositionally equivalent to $\neg K_i \neg(s_i^\ell \succ_i s_i^k) \to \neg K_i(s_i^\ell \succeq_i s_i^k)$),

$$w_0 \models \neg K_i \neg(s_i^\ell \succ_i s_i^k) \to \neg K_i(s_i^\ell \succeq_i s_i^k). \tag{9.8}$$

If, for every w such that $w_0 \sim_i w$, $\pi_i(s_i^k, \sigma_{-i}(w)) \geq \pi_i(s_i^\ell, \sigma_{-i}(w))$, then, by Definition 9.7, $w \in SRAT_i$. If, on the other hand, there is a w_1 such that $w_0 \sim_i w_1$ and $\pi_i(s_i^\ell, \sigma_{-i}(w_1)) > \pi_i(s_i^k, \sigma_{-i}(w_1))$, then, by Definition 9.6, $w_1 \models (s_i^\ell \succ_i s_i^k)$ and thus $w_0 \models \neg K_i \neg(s_i^\ell \succ_i s_i^k)$. Hence, by (9.8), $w_0 \models \neg K_i(s_i^\ell \succeq_i s_i^k)$, that is, there exists a w_2 such that $w_0 \sim_i w_2$ and $w_2 \models \neg(s_i^\ell \succeq_i s_i^k)$, so that, by Axioms **G4** and **G5**, $w_2 \models s_i^k \succ_i s_i^\ell$; that is, by Definition 9.6, $\pi_i(s_i^k, \sigma_{-i}(w_2)) > \pi_i(s_i^\ell, \sigma_{-i}(w_2))$. Hence, by Definition 9.7, $w \in SRAT_i$. Since w_0 and i were chosen arbitrarily, it follows that $SRAT = W$ and thus $\mathbb{CK}SRAT = W$. Hence, by Proposition 9.4, $\sigma(w) \in T^\infty$ for every $w \in W$.

(B) Let F be the $\mathcal{S}5$ epistemic model constructed in the proof of Part B of Proposition 9.4 and let M be a syntactic model based on F that satisfies the validation rules of Definition 9.6. First we show that M validates axiom **SR.** Recall that in F, $W = T^\infty$, $s' \in \sim_i(s)$ if and only if $s_i = s'_i$ and σ is the identity function. Fix an arbitrary player i and an arbitrary state $\hat{s}$. We need to show that, for every $s_i^\ell \in S_i$, $\hat{s} \models \neg\left(K_i(s_i^\ell \succeq_i \hat{s}_i) \wedge \neg K_i \neg(s_i^\ell \succ_i \hat{s}_i)\right)$. Suppose that, for some $s_i^\ell \in S_i$, $\hat{s} \models \left(K_i(s_i^\ell \succeq_i \hat{s}_i) \wedge \neg K_i \neg(s_i^\ell \succ_i \hat{s}_i)\right)$, that is, for every s such that $\hat{s} \sim_i s$ (recall that $\hat{s} \sim_i s$ if and only if $\hat{s}_i = s_i$), $s \models s_i^\ell \succeq_i \hat{s}_i$ and there exists an $\tilde{s}$ such that $\hat{s} \sim_i \tilde{s}$ (that is, $\hat{s}_i = \tilde{s}_i$) and $\tilde{s} \models s_i^\ell \succ_i \hat{s}_i$. Then, by Definition 9.6, for all s such that $\hat{s} \sim_i s$, $\pi_i(s_i^\ell, s_{-i}) \geq \pi_i(\hat{s}_i, s_{-i})$ and $\hat{s} \sim_i \tilde{s}$ and $\pi_i(s_i^\ell, \tilde{s}_{-i}) > \pi_i(\hat{s}_i, \tilde{s}_{-i})$. Then by Definition 9.7, $\hat{s} \notin SRAT_i$. But, as shown in the proof of Proposition 9.4, $SRAT = T^\infty$ so that, since $SRAT \subseteq SRAT_i \subseteq W = T^\infty$, $SRAT_i = T^\infty$, yielding a contradiction. Thus M validates axiom **SR**. Now fix an arbitrary $s \in T^\infty$.

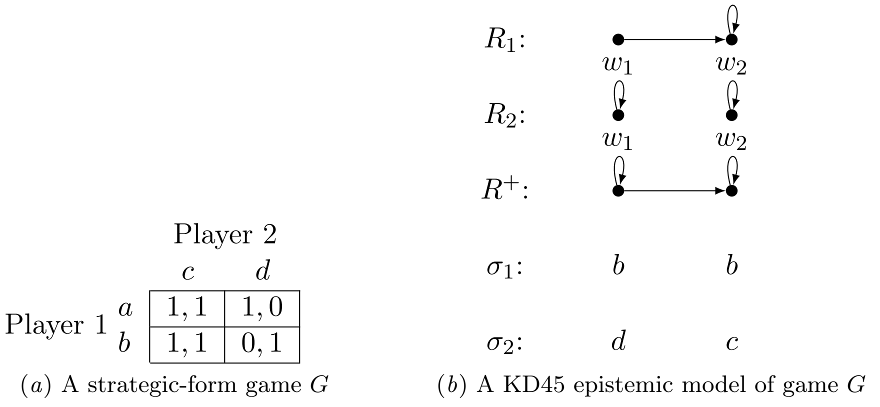
Then, by Definition 9.6, $s \models s$; thus (1) holds. Conversely, let $s \models s$; then, by construction of F, $\sigma(s) = s$ and $s \in T^{\infty}$. Thus (2) holds.

$\dashv$

As noted in Section 9.4 for the case of axiom **WR** (see Remark 9.7), axiom **SR** provides a syntactic characterization of common knowledge of s-rationality in a logic that does not include a common knowledge operator. However, since **SR** expresses the notion that player i chooses s-rationally, by the Necessitation rule every player knows that player i is s-rational and every player knows this, and so on, so that essentially the s-rationality of every player is commonly known. Indeed, if one adds the common knowledge operator CK to the logic, then, by Necessitation, $CK\left(s_i^k \to \neg\left(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k)\right)\right)$ becomes a theorem.

It is also worth repeating (see Remark 9.8), that the difference between the local character of Proposition 9.4 and the global character of Proposition 9.6 is only apparent: the characterization of Proposition 9.4 can in fact be viewed as a global characterization.

Remark 9.9
Note that neither Proposition 9.4 nor Proposition 9.6 is true if one replaces knowledge with belief, as illustrated in the game of Part a of Figure 9.4 and corresponding $\mathcal{KD}45$ frame of Part b. In the corresponding model we have that, according to the stronger notion of s-rationality (Definition 9.7), $SRAT = \{w_1, w_2\}$ so that $w_1 \in \mathbb{CB}SRAT$, despite the fact that $\sigma(w_1) = (b, d)$, which is an inferior strategy profile (relative to the entire game).[15] In other words, common *belief* of s-rationality is compatible with the players collectively choosing an inferior strategy profile. Thus, unlike the weaker notion expressed by axiom **WR**, with axiom **SR** there is a crucial difference between the implications of common *belief* and those of common *knowledge* of rationality. $\dashv$



	c	d
a	$1, 1$	$1, 0$
b	$1, 1$	$0, 1$

(a) A strategic-form game G

(b) A KD45 epistemic model of game G

Figure 9.4: A model with common *belief* of s-rationality at every state

[15] In the game of Figure 9.4 we have that, while $S^{\infty} = S = \{(a,c),(a,d),(b,c),(b,d)\}$, $T^{\infty} = \{(a,c),(b,c)\}$.

9.6 Probabilistic Beliefs and von Neumann Morgenstern Payoffs

So far we have assumed that each player has an *ordinal* ranking of the possible outcomes; furthermore, we restricted attention to *qualitative* beliefs, represented by Kripke frames. In such a framework one can express the fact that, say, Player 1 is uncertain as to whether Player 2 will choose strategy c or strategy d but one cannot express graded forms of beliefs, such as "Player 1 believes that it is twice as likely that Player 2 will play c rather than d". The preponderant approach in the game-theoretic literature is to endow players with probabilistic beliefs and to assume that the players' preferences can be represented by a Bernoulli (also called von Neumann-Morgenstern) utility function. In this section we briefly describe this approach.

		Player 2	
		c	d
Player 1	A	o_1	o_2
	B	o_3	o_4

Figure 9.5: A strategic-form game-frame

Consider the strategic-form game-frame shown in Figure 9.5 (a game-frame is a game without the players' ranking of the outcomes), where o_1, o_2, o_3 and o_4 are the possible outcomes, and suppose that Player 1 assigns subjective probability $\frac{1}{3}$ to the possibility that Player 2 will choose c and probability $\frac{2}{3}$ to Player 2 choosing d. What choice should Player 1 make? If he chooses A, then the outcome will be o_1 with probability $\frac{1}{3}$ and o_2 with probability $\frac{2}{3}$; on the other hand, choosing B will yield outcome o_3 with probability $\frac{1}{3}$ and o_4 with probability $\frac{2}{3}$. Thus comparing A to B amounts to comparing the lottery $\begin{pmatrix} o_1 & o_2 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$ to the lottery $\begin{pmatrix} o_3 & o_4 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$. An ordinal ranking of the set of basic outcomes $\{o_1, o_2, o_3, o_4\}$ is no longer sufficient to determine what is rational for Player 1 to do (given the hypothesized beliefs). Thus we need to modify the models that we have been using so far in two ways: we need to enrich our structures so that we can express probabilistic beliefs and we need to go beyond ordinal rankings of the outcomes.

Definition 9.9
A *probabilistic frame* is a tuple $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{p_i\}_{i\in\mathsf{Ag}}\rangle$ where $\langle W, \{R_i\}_{i\in\mathsf{Ag}}\rangle$ is a $\mathcal{KD}45$ Kripke frame and, for every agent $i \in \mathsf{Ag}$, $p_i : W \to \Delta(W)$ (where $\Delta(W)$ denotes the set of probability measures over W) is a function that satisfies the following properties (we use the notation $p_{i,w}$ instead of $p_i(w)$):[16] $\forall w, w' \in W$,
1. $supp(p_{i,w}) = R_i(w)$, and
2. if $w' \in R_i(w)$ then $p_{i,w'} = p_{i,w}$. ⊣

[16] If μ is a probability measure over W, we denote by $supp(\mu)$ the support of μ, that is, the set of states to which μ assigns positive probability.

Thus $p_{i,w} \in \Delta(W)$ is agent i's subjective probability measure at state w. Condition 1 says that the agent assigns positive probability to all and only the states that she considers possible (according to her accessibility relation R_i) and Condition 2 says that the agent knows her own probabilistic beliefs, since she has the same probability measure at every state that she considers possible.

The semantic belief operator $\mathbb{B}_i : 2^W \to 2^W$ of player i (obtained from the doxastic accessibility relation R_i) is defined as in Section 9.2 (see 9.1) and so is the common belief operator $\mathbb{CB} : 2^W \to 2^W$. In this context, the interpretation of $\mathbb{B}_i E$ is "the event that player i assigns probability 1 to event E".

As noted above, the ordinal ranking of the set of outcomes O that we have postulated so far is not sufficient to determine whether one lottery is better than another. Traditionally, game theorists have assumed that every player has a complete ranking of all the lotteries over the set of basic outcomes O. The *theory of expected utility*, developed by the founders of game theory, namely John von Neumann and Oscar Morgenstern, provides a list of "rationality" or "consistency" axioms for how lotteries should be ranked and yields the following representation theorem. Given a finite set O of *basic outcomes*, we denote by $\Delta(O)$ the set of probability distributions or *lotteries* over O. A *von Neumann-Morgenstern ranking* of $\Delta(O)$ is a binary relation $\succsim^{vnm}$ on $\Delta(O)$ that satisfies a number of properties, known as the von Neumann-Morgenstern axioms or expected utility axioms.[17] If $L, L' \in \Delta(O)$, the interpretation of $L \succsim^{vnm} L'$ is that lottery L is considered to be at least as good as lottery L'.

Theorem 9.10
[von Neumann and Morgenstern (1944)]. Let $O = \{o_1, ..., o_m\}$ be a set of basic outcomes and $\succsim^{vnm}$ a von Neumann-Morgenstern ranking of $\Delta(O)$. Then there exists a function $U : O \to \mathbb{R}$, called a *Bernoulli* (or *von Neumann-Morgenstern*) *utility function* such that, given any two lotteries $L = \begin{pmatrix} o_1 & \cdots & o_m \\ p_1 & \cdots & p_m \end{pmatrix}$ and $L' = \begin{pmatrix} o_1 & \cdots & o_m \\ q_1 & \cdots & q_m \end{pmatrix}$, $L \succsim^{vnm} L'$ if and only if $\sum_{j=1}^{m} U(o_j)p_j \geq \sum_{j=1}^{m} U(o_j)q_j$. The number $\sum_{j=1}^{m} U(o_j)p_j$ is called the *expected utility of lottery L*.
Furthermore, if $U : O \to \mathbb{R}$ is a Bernoulli utility function that represents the ranking $\succsim^{vnm}$, then, for every pair of real numbers $a, b \in \mathbb{R}$ with $a > 0$, the function $V : O \to \mathbb{R}$ defined by $V(o) = aU(o) + b$ is also a Bernoulli utility function that represents $\succsim^{vnm}$. $\dashv$

Definition 9.10
A *finite strategic-form game with cardinal (or von Neumann Morgenstern) payoffs* is a quintuple $G = \left\langle \mathsf{Ag}, \{S_i\}_{i \in \mathsf{Ag}}, O, z, \{\succsim_i^{vnm}\}_{i \in \mathsf{Ag}} \right\rangle$, where Ag, S_i, O and z are as in Definition 9.1 and, for every player $i \in N$, $\succsim_i^{vnm}$ is a von Neumann-Morgenstern ranking of $\Delta(O)$. Such games are often represented in *reduced form* by replacing the triple $\left\langle O, z, \{\succsim_i^{vnm}\}_{i \in \mathsf{Ag}} \right\rangle$ with a set of *cardinal payoff functions* $\{\pi_i\}_{i \in \mathsf{Ag}}$ with $\pi_i : S \to \mathbb{R}$ defined by $\pi_i(s) = U_i(z(s))$, where $U_i : O \to \mathbb{R}$ is a Bernoulli

[17] Because of space limitations we shall not list those axioms. The interested reader is referred to Kreps (1988).

utility function that represents the ranking $\succsim_i^{vnm}$ (whose existence is guaranteed by Theorem 9.10). $\dashv$

Going back to the above example based on Figure 9.5, where Player 1 assigns subjective probability $\frac{1}{3}$ to Player 2 choosing c and probability $\frac{2}{3}$ to Player 2 choosing d, if Player 1 has a von Neumann-Morgenstern ranking $\succsim_1^{vnm}$ of $\Delta(\{o_1, o_2, o_3, o_4\})$, then it is rational for him to choose A if and only if $\frac{1}{3}U_1(o_1) + \frac{2}{3}U_1(o_2) \geq \frac{1}{3}U_1(o_3) + \frac{2}{3}U_1(o_4)$, where U_1 is a Bernoulli utility function that represents $\succsim_1^{vnm}$.

It is worth stressing that the move from games where players have ordinal rankings of the basic outcomes to games where they have von Neumann-Morgenstern rankings of lotteries (over basic outcomes) is not an innocuous move. The reason is not only that much more is assumed about each individual player's preferences, but also that - since the game is implicitly assumed to be common knowledge among the players - each player is assumed to know the cardinal rankings of his opponents (how they rank all possible lotteries, what their attitude to risk is, etc.).

The definition of an epistemic model of a game (Definition 9.2) can be straightforwardly extended to games with von Neumann-Morgenstern payoffs.

Definition 9.11
Given a strategic-form game with von Neumann Morgenstern payoffs G of the form $G = \left\langle \mathsf{Ag}, \{S_i\}_{i\in\mathsf{Ag}}, \{\pi_i\}_{i\in\mathsf{Ag}} \right\rangle$, an *epistemic-probabilistic model* of G is a tuple $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{p_i\}_{i\in\mathsf{Ag}}, \{\sigma_i\}_{i\in\mathsf{Ag}}\rangle$ where $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{p_i\}_{i\in\mathsf{Ag}}\rangle$ is a probabilistic frame (see Definition 9.9) and $\sigma_i : W \to S_i$ is - as before - a function that associates, with every state, a strategy of player i, satisfying the property that if $w' \in R_i(w)$ then $\sigma_i(w') = \sigma_i(w)$. $\dashv$

As before, given a state w and a player i, we denote by $\sigma_{-i}(w)$ the strategy profile of the players other than i at state w. The definition of rationality (Definition 9.3) can now be sharpened, as follows.

Definition 9.12
Fix a strategic-form game with von Neumann Morgenstern payoffs G and an epistemic-probabilistic model of G. At state w player i's strategy $s_i = \sigma_i(w)$ is *rational* if it maximizes player i's payoff, given his beliefs at w, that is, if

$$\textstyle\sum_{x\in R_i(w)} p_{i,w}(x)\ \pi_i(s_i, \sigma_{-i}(x)) \geq \sum_{x\in R_i(w)} p_{i,w}(x)\ \pi_i(s_i', \sigma_{-i}(x)), \quad \forall s_i' \in S_i.$$

[Recall that, by Definition 9.11, the function $\sigma_i(\cdot)$ is constant on the set $R_i(w)$].$\dashv$

What are the implications of common belief of rationality in this framework? It turns out that a result similar to Proposition 9.1 holds in this case too: common belief of rationality is characterized by a strengthening of the IDSDS procedure (Definition 9.5).[18] Because of space limitations we omit the details. Similarly, a

[18] The modified procedure allows the deletion of pure strategies that are strictly dominated by a mixed strategy, that is, by a probability distribution over the set of pure strategies. This is because, as shown by Pearce (1984), a pure strategy s is strictly dominated by another, possibly mixed, strategy if and only if there is no (probabilistic) belief concerning the strategies chosen by the opponents that makes s a best reply, that is, there is no belief that makes s a rational choice.

result along the lines of Proposition 9.4 holds in this case too for a strengthening of the IDIP procedure (see Stalnaker (1994)).

9.7 Dynamic Games with Perfect Information

So far we have restricted attention to strategic-form games, where the players make their choices simultaneously or in ignorance of the other players' choices. We now turn to dynamic games, where players make choices sequentially, having some information about the moves previously made by their opponents. If information is partial, the game is said to have *imperfect information*, while the case of full information is referred to as *perfect information*. Because of space limitations we shall restrict attention to perfect-information games.

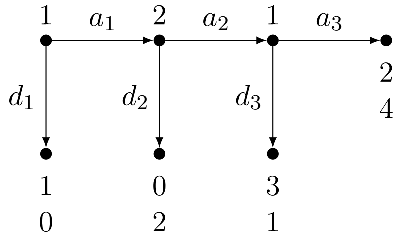


Figure 9.6: A dynamic game with perfect information

An example of a dynamic game with perfect information is shown in Figure 9.6 in the form of a tree. Each node in the tree represents a history of prior moves and is labeled with the player whose turn it is to move. For example, at history a_1a_2 it is Player 1's turn to move (after his initial choice of a_1 followed by Player 2's choice of a_2) and he has to choose between two actions: a_3 and d_3. The terminal histories (the leaves of the tree) represent the possible outcomes and each player i is assumed to have an ordinal preference relation $\succsim_i$ over the set of terminal histories (in Figure 9.6 the players' preferences over the terminal histories have been represented by means of ordinal utility functions, as explained below).

The formal definition of a perfect-information game is as follows. If A is a set, we denote by A^* the set of finite sequences in A. If $h = \langle a_1, ..., a_k \rangle \in A^*$ and $1 \leq j \leq k$, the sequence $\langle a_1, ..., a_j \rangle$ is called a *prefix* of h. If $h = \langle a_1, ..., a_k \rangle \in A^*$ and $a \in A$, we denote the sequence $\langle a_1, ..., a_k, a \rangle \in A^*$ by ha.

Definition 9.13
A *finite extensive game with perfect information and ordinal payoffs* is a tuple $\left\langle A, H, \mathsf{Ag}, \iota, \{\succsim_i\}_{i \in \mathsf{Ag}} \right\rangle$ whose elements are:

- A finite set of actions A.
- A finite set of histories $H \subseteq A^*$ which is closed under prefixes (that is, if $h \in H$ and $h' \in A^*$ is a prefix of h, then $h' \in H$). The null history $\langle\rangle$,

denoted by $\emptyset$, is an element of H and is a prefix of every history. A history $h \in H$ such that, for every $a \in A$, $ha \notin H$, is called a *terminal history.* The set of terminal histories is denoted by Z. $D = H \backslash Z$ denotes the set of non-terminal or *decision* histories. For every decision history $h \in D$, we denote by $A(h)$ the set of actions available at h, that is, $A(h) = \{a \in A : ha \in H\}$.

- A finite set Ag of players.
- A function $\iota : D \to \mathsf{Ag}$ that assigns a player to each decision history. Thus $\iota(h)$ is the player who moves at history h. For every $i \in \mathsf{Ag}$, let $D_i = \iota^{-1}(i)$ be the set of histories assigned to player i.
- For every player $i \in \mathsf{Ag}$, $\succsim_i$ is an ordinal ranking of the set Z of terminal histories. $\dashv$

The ordinal ranking of player i is normally represented by means of an ordinal *utility (*or *payoff) function* $U_i : Z \to \mathbb{R}$ satisfying the property that $U_i(z) \geq U_i(z')$ if and only if $z \succsim_i z'$. In the game of Figure 9.6, associated with every terminal history is a pair of numbers: the top number is the utility of Player 1 and the bottom number is the utility of Player 2.

Histories will be denoted more succinctly by listing the corresponding actions, without angled brackets and without commas; thus instead of writing for instance $\langle \emptyset, a_1, a_2, a_3, a_4 \rangle$ we simply write $a_1a_2a_3a_4$.

In their seminal book, von Neumann and Morgenstern (1944) showed that a dynamic game can be reduced to a strategic-form game by defining strategies as complete, contingent plans of action. In the case of perfect-information games a *strategy* for a player is a function that associates with every decision history assigned to that player one of the choices available there. For example, a possible strategy of Player 1 in the game of Figure 9.6 is (d_1, d_3). A profile of strategies (one for each player) determines a unique path from the null history (the root of the tree) to a terminal history (a leaf of the tree). Figure 9.7 shows the strategic-form corresponding to the extensive form of Figure 9.6.

		Player 2	
		a_2	d_2
Player 1	a_1a_3	$2,4$	$0,2$
	a_1d_3	$3,1$	$0,2$
	d_1a_3	$1,0$	$1,0$
	d_1d_3	$1,0$	$1,0$

Figure 9.7: The strategic-form of the game of Figure 9.6

How should a model of a dynamic game be constructed? One approach in the literature has been to consider models of the corresponding strategic-form (the type of models considered in Section 9.2). However, there are several conceptual issues that arise in this context. The interpretation of $s_i = \sigma_i(w)$ is that at state w player i "chooses" strategy s_i. Now consider a model of the game of Figure

9.6 and a state w where $\sigma_1(w) = (d_1, a_3)$. What does it mean to say that Player 1 "chooses" strategy (d_1, a_3)? The first part of the strategy, namely d_1, can be interpreted as a description of Player 1's actual choice to play d_1, but the second part of the strategy, namely a_3, has no such interpretation: if Player 1 in fact plays d_1 then he knows that he will not have to make any further choices and thus it is not clear what it means for him to "choose" to play a_3 in a situation that is made impossible by his decision to play d_1.[19] Thus it does not seem to make sense to interpret $\sigma_1(w) = (d_1, a_3)$ as 'at state w Player 1 chooses (d_1, a_3)'. Perhaps the correct interpretation is in terms of a more complex sentence such as 'Player 1 chooses to play d_1 and if - contrary to this - he were to play a_1 and Player 2 were to follow with a_2, then Player 1 would play a_3'. Thus while in a simultaneous game the association of a strategy of player i to a state can be interpreted as a description of player i's actual behavior at that state, in the case of dynamic games this interpretation is no longer valid, since one would end up describing not only the actual behavior of player i at that state but also his counterfactual behavior. Methodologically, this is not satisfactory: if it is considered to be necessary to specify what a player would do in situations that do not occur in the state under consideration, then one should model the counterfactual explicitly. But why should it be necessary to specify at state w (where Player 1 is playing d_1) what he would do at the counterfactual history a_1a_2? Perhaps what matters is not so much what Player 1 would actually do there but what Player 2 believes that Player 1 would do: after all, Player 2 might not know that Player 1 has decided to play d_1 and needs to consider what to do in the eventuality that Player 1 actually ends up playing a_1. So, perhaps, the strategy of Player 1 is to be interpreted as having two components: (1) a description of Player 1's behavior and (2) a conjecture in the mind of Player 2 about what Player 1 would do. If this is the correct interpretation, then one could - from a methodological point of view - object that it would be preferable to disentangle the two components and model them explicitly.[20]

An alternative - although less common - approach in the literature dispenses with strategies and considers models of games where (1) states are described in terms of players' *actual behavior* and (2) players' conjectures concerning the actions of their opponents (as well as their own actions) in various hypothetical situations are modeled by means of a generalization of the Kripke frames considered so far. The generalization is obtained by encoding not only the initial beliefs of the players (at each state) but also their *dispositions to revise those beliefs* under various hypothesis. These structures are reviewed in the next section.

[19] For this reason, some authors, instead of using strategies, use the weaker notion of "plan of action" introduced in Rubinstein (1991). A plan of action for a player only contains choices that are not ruled out by his earlier choices. For example, the possible plans of action for Player 1 in the game of Figure 9.6 are d_1, (a_1, a_3) and (a_1, d_3). However, most of the issues raised below apply also to plans of action. The reason for this is that a choice of player i at a later decision history of his may be counterfactual at a state because of the choices of *other* players (which prevent that history from being reached).

[20] For a more in-depth discussion of these issues see (Bonanno, 2014).

9.8 The Semantics of Belief Revision

A $\mathcal{KD}45$ Kripke frame $\langle W, \{R_i\}_{i\in\mathsf{Ag}}\rangle$ represents the actual beliefs of the agents at every state w. In order to capture the agents' disposition to revise their beliefs under various hypotheses, we need to consider extensions of those frames.

Definition 9.14
A *belief revision frame* is a triple $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{\mathcal{E}_i, f_i\}_{i\in\mathsf{Ag}}\rangle$, where the pair consisting of $\langle W, \{R_i\}_{i\in\mathsf{Ag}}\rangle$ is a $\mathcal{KD}45$ Kripke frame and, for every agent $i \in \mathsf{Ag}$, $\mathcal{E}_i \subseteq 2^W \backslash \varnothing$ is a set of admissible hypotheses (or potential items of information) and $f_i : W \times \mathcal{E}_i \to 2^W$ is a function that satisfies the following properties: $\forall w \in W$, $\forall E, F \in \mathcal{E}_i$,

1. $f_i(w,E) \neq \varnothing$,
2. $f_i(w,E) \subseteq E$,
3. if $R_i(w) \cap E \neq \varnothing$ then $f_i(w,E) = R_i(w) \cap E$,
4. if $E \subseteq F$ and $f_i(w,F) \cap E \neq \varnothing$ then $f_i(w,E) = f_i(w,F) \cap E$.

$\dashv$

The event $f_i(w,E)$ is interpreted as the set of states that player i would consider possible, at state w, under the supposition that (or if informed that) E is true. Condition 1 requires these suppositional beliefs to be consistent. Condition 2 requires that, under the supposition that E is true, E be indeed considered true. Condition 3 says that if E is compatible with the initial beliefs (given by $R_i(w)$) then the suppositional beliefs coincide with the initial beliefs conditioned on event E.[21] Condition 4 is an extension of Condition 3: if E implies F and E is compatible (not with player i's prior beliefs but) with the *posterior* beliefs that player i would have if she supposed (or learned) that F were the case (let's call these her posterior F-beliefs), then her beliefs under the supposition (or information) that E must coincide with her posterior F-beliefs conditioned on event E.

Thus the function f_i can be used to model the full epistemic attitude of player i at every state w: her prior (or initial) beliefs are given by the set $R_i(w)$ and, for every event E, the set $f_i(w,E)$ captures how she is disposed to revise those beliefs under the supposition that E is true. In particular, the function f_i tells us how player i would revise her prior beliefs if she learned information that contradicted those beliefs.

Remark 9.11
If $\mathcal{E}_i = 2^W \backslash \varnothing$ then Conditions 1-4 of Definition 9.14 imply that, for every $w \in W$, there exists a "plausibility" relation Q_i^w on W which is complete ($\forall w_1, w_2 \in W$, either $w_1 Q_i^w w_2$ or $w_2 Q_i^w w_1$ or both) and transitive ($\forall w_1, w_2, w_3 \in W$, if $w_1 Q_i^w w_2$ and $w_2 Q_i^w w_3$ then $w_1 Q_i^w w_3$) and such that, for every $E \subseteq W$ with $E \neq \varnothing$, $f_i(w,E) = \{x \in E : x Q_i^w y,\ \forall y \in E\}$. The interpretation of $x Q_i^w y$ is that - at state w and according to player i - state x is at least as plausible as state y. Thus $f_i(w,E)$ is the set of most plausible states in E (according to player i at

[21] Note that it follows from Condition 3 and seriality of R_i that, for every $w \in W$, $f_i(w,W) = R_i(w)$, so that one could simplify the definition of a belief revision frame by dropping the relations R_i and recovering the initial beliefs at state w from the set $f_i(w,W)$. We have chosen not to do so in order to maintain continuity in the exposition.

state w). If $\mathcal{E}_i \neq 2^W \backslash \varnothing$ then Conditions 1-4 in Definition 9.14 are necessary but not sufficient for the existence of such a plausibility relation. The existence of a plausibility relation that rationalizes the function $f_i(w, \cdot) : \mathcal{E}_i \to 2^W$ is necessary and sufficient for the belief revision policy encoded in $f_i(w, \cdot)$ to be compatible with the syntactic theory of belief revision introduced in Alchourrón, Gärdenfors, and Makinson (1985), known as the AGM theory.

One can associate with each function f_i a conditional belief operator $\overline{\mathbb{B}}_i : 2^W \times \mathcal{E}_i \to 2^W$ as follows, with $F \in 2^W$ and $E \in \mathcal{E}_i$:

$$\overline{\mathbb{B}}_i(F|E) = \{w \in W : f_i(w, E) \subseteq F\}. \tag{9.9}$$

Possible interpretations of the event $\overline{\mathbb{B}}_i(F|E)$ are "according to player i, if E were the case, then F would be true" or "if informed that E, player i would believe that F" or "under the supposition that E, player i would believe that F".

The unconditional belief operator $\mathbb{B}_i : 2^W \to 2^W$ remains as defined in Section 9.5 and represents the initial beliefs of agent i.[22] Similarly, the common belief operator $\mathbb{CB}$ remains as defined in Section 9.5 and captures what is *initially* common belief among the agents.

9.9 Common Belief of Rationality in Perfect-Information Games

We can now return to dynamic games with perfect information. First we define an algorithm, known as *backward induction*, which is meant to capture the "rational" way of playing these games and explore the possibility of providing an epistemic foundation for it.

The backward induction algorithm starts at the end of the game and proceeds backwards towards the root:

1. Start at a decision history h whose immediate successors are only terminal histories (e.g. history a_1a_2 in the game of Figure 9.6) and select a choice that maximizes the utility of player $\iota(h)$ (in the example of Figure 9.6, at a_1a_2 Player 1's optimal choice is d_3 (since it gives her a payoff of 3 rather than 2, which is the payoff that she would get if she played a_3). Delete the immediate successors of history h (that is, turn h into a terminal history) and assign to h the payoff vector associated with the selected choice.

2. Repeat Step 1 until all the decision histories have been exhausted.

For example, the choices selected by the backward-induction algorithm in the game of Figure 9.6 are d_3, d_2 and d_1.[23]

[22] Note that, for every event F, $\mathbb{B}_iF = \overline{\mathbb{B}}_i(F|W)$.

[23] The backward induction algorithm may yield more than one solution. Multiplicity arises if there is at least one player who has more than one payoff-maximizing choice at a decision history of his.

A question that has been studied extensively in the literature is whether *initial* common belief of rationality can provide an epistemic justification for the backward-induction solution. In order to answer this question we need to introduce the notion of an epistemic model of a perfect-information game.

Definition 9.15
Given a dynamic game with perfect information and ordinal payoffs $\Gamma = \left\langle A, H, \mathsf{Ag}, \iota, \{\succsim_i\}_{i \in \mathsf{Ag}} \right\rangle$, an *epistemic model* of Γ is a tuple $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{\mathcal{E}_i, f_i\}_{i\in\mathsf{Ag}}, \zeta\rangle$ where $\langle W, \{R_i\}_{i\in\mathsf{Ag}}, \{\mathcal{E}_i, f_i\}_{i\in\mathsf{Ag}}\rangle$ is a belief revision frame (Definition 9.14) and $\zeta : W \to Z$ is a function that associates with every state a terminal history and satisfies the following property: $\forall w, w' \in W, \forall i \in \mathsf{Ag}, \forall h \in H, \forall a \in A$,

$$\begin{array}{l}\text{If } h \text{ is a decision history of player } i,\ a \text{ an action at } h\\ \text{and } ha \text{ a prefix of } \zeta(w) \text{ then, } \forall w' \in R_i(w),\\ \text{if } h \text{ is a prefix of } \zeta(w') \text{ then } ha \text{ is a prefix of } \zeta(w').\end{array} \tag{9.10}$$

$\dashv$

The function ζ describes the *actual behavior* of the players at any given state. Thus we are not associating a strategy profile with a state but a sequence of actions leading from the null history to a terminal history. Condition (9.10) states that if at a state the play of the game reaches decision history h of player i, where she actually takes action a, then either player i initially believes that history h will not be reached or, if she considers it possible that history h will indeed be reached, then she has correct beliefs about what action she will take (namely a) if h is reached.

Condition (9.10) can be stated more succinctly in terms of events. If E and F are two events, we denote by $E \to F$ the event $\neg E \cup F$. Thus $E \to F$ captures the material conditional. Given a history h in the game, we denote by $[h]$ the event that h is reached, that is, $[h] = \{w \in W : h \text{ is a prefix of } \zeta(w)\}$. Recall that D_i denotes the set of decision histories of player i and $A(h)$ the set of choices available at h. Then (9.10) can be stated as follows:[24]

$$\begin{array}{l}\forall h \in D_i, \forall a \in A(h),\\ [ha] \subseteq \mathbb{B}_i([h] \to [ha]).\end{array} \tag{9.11}$$

In words: if, at a state, player i takes action a at her decision history h, then she believes that if h is reached then she takes action a.

Condition (9.11) rules out the possibility that a player may be uncertain about her own choice of action at decision histories of hers that are not ruled out by her initial beliefs. In general, a corresponding condition might not hold for *revised* beliefs. That is, suppose that at state w player i erroneously believes that her decision history h will not be reached ($w \in [h]$ but $w \in B_i\neg[h]$); suppose also that a is the action that she will choose at h ($w \in [ha]$). It may be the case that, according to her revised beliefs on the supposition that h is reached, she believes

[24] Note that, if at state w player i believes that history h will *not* be reached ($\forall w' \in R_i(w)$, $w' \notin [h]$) then $R_i(w) \subseteq \neg[h] \subseteq [h] \to [ha]$, so that $w \in \mathbb{B}_i\left([h] \to [ha]\right)$ and therefore (9.11) is satisfied even if $w \in [ha]$.

that she takes an action b different from the action that she actually takes, namely a. In order to rule this out we need to impose the following strengthening of (9.11):[25]

$$\begin{aligned}&\forall h \in D_i,\ \forall a \in A(h),\\ &[ha] \subseteq \overline{\mathbb{B}}_i([ha]|[h]).\end{aligned} \tag{9.12}$$

How can rationality be captured in the models that we are considering? Various definitions of rationality have been suggested in the literature, most notably *material rationality* and *substantive rationality* . The former notion is weaker in that a player can be found to be irrational only at decision histories of hers that are actually reached. The latter notion, on the other hand, is more stringent since a player can be judged to be irrational at a decision history h of hers even if she correctly believes that h will not be reached. We will focus on the weaker notion of material rationality. As before, we shall define a player's rationality as a proposition, that is, an event. Recall that Z denotes the set of terminal histories and $u_i : Z \to \mathbb{R}$ is player i's ordinal utility function (representing her preferences over the set Z). Define $\pi_i : W \to \mathbb{R}$ by $\pi_i(w) = u_i(\zeta(w))$. For every $x \in \mathbb{R}$, let $[\pi_i \leq x]$ be the event that player i's payoff is not greater than x, that is, $[\pi_i \leq x] = \{w \in W : \pi_i(w) \leq x\}$ and, similarly, let $[\pi_i > x] = \{w \in W : \pi_i(w) > x\}$. Then we say that player i is materially rational at a state if, for every decision history h of hers that is actually reached at that state and for every real number x, it is not the case that she believes that – under the supposition that h is reached – (1) her payoff from her actual choice would not be greater than x and (2) it would be greater than x if she were to take an action different from the one that she is actually taking (at that history in that state).[26]

Formally this can be stated as follows (recall that D_i denotes the set of decision

[25] (9.12) is implied by (9.11) whenever player i's initial beliefs do not rule out h. That is, if $w \in \neg\mathbb{B}_i\neg[h]$ (equivalently, $R_i(w) \cap [h] \neq \varnothing$) then, for every $a \in A(h)$,

$$\text{if } w \in [ha] \text{ then } w \in \overline{\mathbb{B}}_i([ha]|[h]). \qquad \text{(F1)}$$

In fact, by Condition 3 of Definition 9.14 (since, by hypothesis, $R_i(w) \cap [h] \neq \varnothing$),

$$f_i(w, [h]) = R_i(w) \cap [h]. \qquad \text{(F2)}$$

Let $a \in A(h)$ be such that $w \in [ha]$. Then, by (9.11), $w \in \mathbb{B}_i([h] \to [ha])$, that is, $R_i(w) \subseteq \neg[h] \cup [ha]$. Thus $R_i(w) \cap [h] \subseteq (\neg[h] \cap [h]) \cup ([ha] \cap [h]) = \varnothing \cup [ha] = [ha]$ (since $[ha] \subseteq [h]$) and therefore, by (F2), $f_i(w, [h]) \subseteq [ha]$, that is, $w \in \overline{\mathbb{B}}_i([ha]|[h])$.

[26] This definition is a "local " definition in that it only considers, for every decision history of player i, a change in player i's choice at that decision history and not also at later decision histories of hers. One could make the definition of rationality more stringent by simultaneously considering changes in the choices at a decision history and subsequent decision histories of the same player (if any).

histories of player i and $A(h)$ the set of actions available at h):

$$\begin{array}{ll} \text{Player } i \text{ is } \textit{materially rational} \text{ at } w \in W \text{ if, } \forall h \in D_i, \forall a \in A(h) & \\ \text{if } ha \text{ is a prefix of } \zeta(w) \text{ then, } \forall b \in A(h),\ \forall x \in \mathbb{R}, & (9.13) \\ \overline{\mathbb{B}}_i([\pi_i \le x]\,|[ha]) \to \neg\overline{\mathbb{B}}_i([\pi_i > x]\,|[hb]). & \end{array}$$

Note that, in general, we cannot replace the antecedent $\overline{\mathbb{B}}_i([\pi_i \le x]\,|[ha])$ with $\mathbb{B}_i([ha] \to [\pi_i \le x])$, because at state w player i might initially believe that h will not be reached, in which case it would be trivially true that $w \in \mathbb{B}_i([ha] \to [\pi_i \le x])$; however, if decision history h is actually reached at w then player i will be surprised and will have to revise her beliefs. Thus her rationality is judged on the basis of her *revised* beliefs. Note, however, that if $w \in \neg\mathbb{B}_i\neg[h]$, that is, if at w she does not rule out the possibility that h will be reached and $a \in A(h)$ is the action that she actually takes at history h at state w ($w \in [ha]$), then, for every event F, $w \in \mathbb{B}_i([ha] \to F)$ if and only if $w \in \overline{\mathbb{B}}_i(F|[ha])$.[27]

Note also that, according to (9.13), a player is trivially rational at any state at which she does not take any actions.

Does initial common belief that all the players are materially rational (according to 9.13) imply backward induction in perfect-information games? The answer is negative.[28] To see this, consider the perfect-information game shown in Figure 9.6 and the model of it shown in Figure 9.8.[29]

First of all, note that the common belief relation R^+ is obtained by adding to R_2 the pair (w_2, w_2); thus, in particular, $R^+(w_2) = \{w_2, w_3\}$. We want to show

[27] Proof. Suppose that $w \in [ha] \cap \neg\mathbb{B}_i\neg[h]$. As shown in Footnote 25 (see (F2)),

$$R_i(w) \cap [h] = f_i(w,[h]). \quad \text{(G1)}$$

Since $[ha] \subseteq [h]$,

$$R_i(w) \cap [h] \cap [ha] = R_i(w) \cap [ha]. \quad \text{(G2)}$$

As shown in Footnote 25, $f_i(w,[h]) \subseteq [ha]$ and, by Condition 1 of Definition 9.14, $f_i(w,[h]) \neq \varnothing$. Thus $f_i(w,[h]) \cap [ha] = f_i(w,[h]) \neq \varnothing$. Hence, by Condition 4 of Definition 9.14

$$f_i(w,[h]) \cap [ha] = f_i(w,[ha]). \quad \text{(G3)}$$

By intersecting both sides of (G1) with $[ha]$ and using (G2) and (G3) we get that $R_i(w) \cap [ha] = f_i(w,[ha])$.

[28] In fact, common belief of material rationality does not even imply a Nash equilibrium outcome. A Nash equilibrium is a strategy profile satisfying the property that no player can increase her payoff by unilaterally changing her strategy. A Nash equilibrium outcome is a terminal history associated with a Nash equilibrium. Note that a backward-induction solution of a perfect-information game can be expressed as a strategy profile and is always a Nash equilibrium.

[29] In Figure 9.8 we have only represented parts of the functions f_1 and f_2, namely the following: $f_1(w_3,\{w_1,w_2,w_4\}) = \{w_4\}$ and $f_2(w_2,\{w_1,w_2,w_4\}) = f_2(w_3,\{w_1,w_2,w_4\}) = \{w_1\}$. Note that $[a_1] = \{w_1,w_2,w_4\}$).

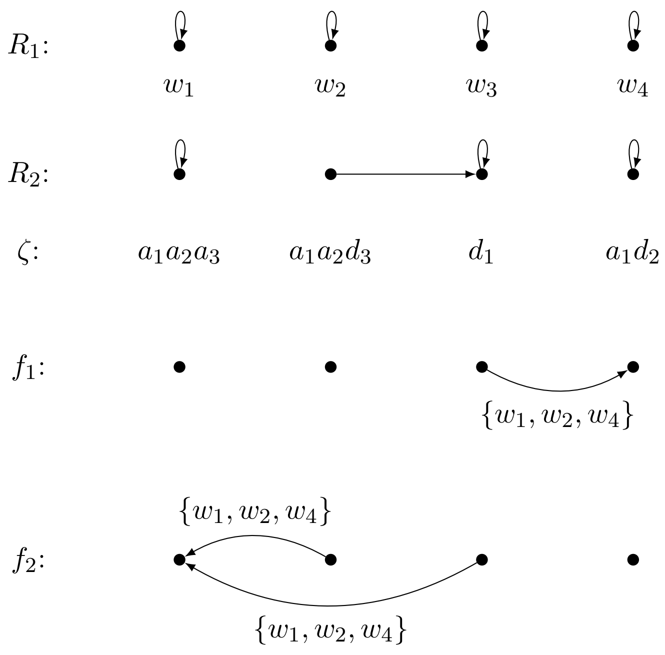


Figure 9.8: A (partial) model of the game of Figure 9.6

that both players are materially rational at both states w_2 and w_3, so that at state w_2 it is initially common belief that both players are materially rational, despite that fact that the play of the game at w_2 is $a_1a_2d_3$, which is not the backward-induction play. Clearly, Player 1 is rational at state w_2 (since he obtains his largest possible payoff); he is also rational at state w_3 because he knows that he plays d_1, obtaining a payoff of 1, and believes that if he were to play a_1 then Player 2 would respond with d_2 and give him a payoff of zero: this belief is encoded in $f_1(w_3, [a_1]) = \{w_4\}$, where $[a_1] = \{w_1, w_2, w_4\}$, and $\zeta(w_4) = a_1d_2$. Player 2 is trivially rational at state w_3 since she does not take any actions there. Now consider state w_2. Player 2 initially erroneously believes that Player 1 will end the game by playing d_1: $R_2(w_2) = \{w_3\}$ and $\zeta(w_3) = d_1$. However, at state w_2, Player 1 is in fact playing a_1 and thus Player 2 will be surprised. Her initial disposition to revise her beliefs on the supposition that Player 1 plays a_1 is such that she would believe that she herself would play a_2 and Player 1 would follow with a_3, thus giving her the largest possible payoff: this belief is encoded in $f_2(w_2, [a_1]) = \{w_1\}$ (recall that $[a_1] = \{w_1, w_2, w_4\}$) and $\zeta(w_1) = a_1a_2a_3$. Hence she is rational at state w_2, according to (9.13).

In order to obtain the backward-induction solution, one needs to go beyond common initial belief of material rationality. Proposals in the literature include the notions of epistemic independence, strong belief, stable belief and substantive rationality. Space limitations prevent us from discussing these topics.

It is worth stressing that *in the models considered above, strategies do not play* *any role*: states are described in terms of the players' actual behavior along a play of the game. One could view a player's strategy as her (conditional) beliefs about what she would do under the supposition that each of her decision histories is reached. However, the models considered so far do not guarantee that a player's revised beliefs select a unique action at each of her decision histories. One could impose such a restriction on the players' dispositions to revise their beliefs.[30] However, in this setup strategies would then be cognitive constructs rather than objective counterfactuals about what a player would actually do at each of her decision histories.

9.10 Notes

In this section we point to the main references in the areas reviewed in this chapter, as well as references for related topics.

The birth of game theory The beginning of game theory is normally associated with the publication, in 1944, of the book *Theory of games and economic behavior* by von Neumann and Morgenstern (1944), although Cournot (1838) provided an analysis of simultaneous games among firms as early as 1838. Cournot's analysis of competition was later elaborated on by Bertrand (1883), von Stackelberg (1934) and by Hotelling (1929). Other notable precursors of the book by von Neumann and Morgenstern are an article by Zermelo (1913) (where he proved that in the game of chess either White has a strategy that guarantees him a win, or Black has a strategy that guarantees her a win, or both players have a strategy that guarantees a draw) and an article by von Neumann (1928) (where he proved the existence of a value in every finite zero-sum game). For a brief history of the first forty years of the development of game theory see the paper by Aumann (1987b).

The birth of the epistemic foundation program The origins of the literature on the epistemic foundations of solution concepts in non-cooperative games can be traced to two seminal papers by Bernheim (1984) and by Pearce (1984), both published in 1984. The purpose of these two articles was to capture the notion of "common recognition of rationality" in games. The analysis, however, was not developed explicitly in terms of epistemic notions: the idea of common belief of rationality was captured indirectly through the notion of rationalizability, which is an iterative procedure of elimination of strategies that are never a best response.

Another pioneering contribution was that by Aumann (1987a), providing an epistemic characterization of the notion of correlated equilibrium in terms of common knowledge of rationality when the players' beliefs share a common prior.

Extensive surveys of the literature on the epistemic foundation program are provided by Battigalli and Bonanno (1999), Dekel and Gul (1997) and by Perea (2012).

[30] The relevant restriction is as follows: $\forall h \in D_i, \forall a, b \in A(h), \forall w, w', w'' \in W$, if $w', w'' \in f_i(w, [h])$ and ha is a prefix of $\zeta(w')$ and hb is a prefix of $\zeta(w'')$ then $a = b$.

Epistemic models of strategic-form games There are two types of epistemic models of strategic-form games used in the game-theoretic literature: the "state-space" models and the "hierarchy of beliefs" models. The qualitative Kripke models considered in Sections 9.2 and 9.3 and their probabilistic counterparts considered in Section 9.6 are known in the game-theoretic literature as state-space models. Although, in the philosophy literature, Kripke frames date back to the work of Kripke (1963), in game theory state-space models first appeared in the work of Aumann (1976). Aumann (1987a) used a state-space model to obtain a characterization of the notion of correlated equilibrium using $\mathcal{S}5$ frames. Stalnaker (1994, 1996) provided the first systematic analysis of solution concepts in terms of $\mathcal{KD}45$ epistemic models of games.

The alternative approach in the game-theoretic literature uses the probabilistic hierarchy-of-belief models and type spaces that where introduced in the seminal papers of Harsanyi (1968), which started the literature on incomplete-information games. The first epistemic characterization of common belief of rationality in strategic-form games using these structures was provided by Tan and Werlang (1988). They showed that the (probabilistic version of) the iterative elimination of strictly dominated strategies identifies the strategy profiles that are compatible with common belief of rationality. The state-space formulation of this result is due to Stalnaker (1994), but it was implicit in Brandenburger and Dekel (1987). All these characterizations were for games with von Neumann-Morgenstern payoffs and for probabilistic beliefs. The stronger iterative elimination procedure (the stronger version of the IDIP algorithm given in Definition 9.8) and corresponding epistemic characterization is due to Stalnaker (1994) (with a correction by Bonanno and Nehring (1998)). The qualitative characterizations of Propositions 9.1 and 9.4 are based on work by Bonanno (2008).

Epistemic foundations of other strategic-form solution concepts Because of space limitations, we have restricted attention to the epistemic foundations of only some solution concepts. In the literature, epistemic conditions have been studied for additional solution concepts, such as for correlated equilibrium by Aumann (1987a) and Barelli (2009), for Nash equilibrium by Aumann and Brandenburger (1995), Bach and Tsakas (2014), Barelli (2009), Perea (2007b), and by Polak (1999), and for iterated admissibility by Brandenburger (1992), Barelli and Galanis (2013) Brandenburger, Friedenberg, and Keisler (2008), Samuelson (1992), and by Stahl (1995). Surveys of the literature are given by Battigalli and Bonanno (1999), Dekel and Gul (1997) and by Perea (2012).

The use of logic in the analysis of games The literature on the epistemic foundation program is predominantly based on the semantic approach. The first to use formal logic in the analysis of games were Bacharach (1987) (who used first-order logic to investigate the notion of Nash equilibrium in strategic- form games) and Bonanno (1991) (who used propositional logic to investigate the notion of backward-induction in dynamic games with perfect information). There is now a sizeable literature that analyzes games using logic, in particular epistemic logic (see, for example, work by Board (2004), Bonanno (2001), Clausing (2003, 2004), de Bruin (2010) and by van Benthem (2011)). The analysis of Sections 9.4 and

9.5 is based on a paper by Bonanno (2008). The three axioms given in footnote 12 that provide a finite axiomatisation for common belief are also taken from a paper by Bonanno (1996).

Epistemic foundations of backward induction The issue of whether the backward-induction algorithm can be given an epistemic foundation has given rise to a large literature. The seminal paper was by Ben-Porath (1997). There are two strands in this literature. One group of papers uses epistemic models where states are described in terms of strategies (see, for example, work by Aumann (1995, 1998), Balkenborg and Winter (1997), Battigalli and Siniscalchi (2002) ,Halpern (2001), and by Stalnaker (1998)). A second group of papers (by, for example, Baltag, Smets, and Zvesper (2009), Battigalli, Di-Tillio, and Samet (2013), and Samet (1996)) uses the "behavioral" models discussed in Section 9.9, which were introduced by Samet (1996). There is a bewildering collection of claims in the literature concerning the implications of rationality in dynamic games with perfect information: Aumann (1995) proves that common *knowledge* of rationality implies the backward induction solution, Ben-Porath (1997) and Stalnaker (1998) prove that common *belief / certainty* of rationality is *not* sufficient for backward induction, Samet (1996) proves that what is needed for backward induction is common *hypothesis* of rationality, Feinberg (2005) shows that common *confidence* of rationality logically contradicts the knowledge implied by the structure of the game, etc. The sources of this wide variety of results are partly clarified in two recent surveys of this literature, by Brandenburger (2007) and by Perea (2007a).

It is worth noting that the models of dynamic games considered in Section 9.9 are not the only possibility. Instead of modeling the epistemic states of the players in terms of their prior beliefs and prior disposition to revise those beliefs in a static framework, one can model the actual beliefs that the players hold at the time at which they make their choices. In such a framework the players' initial belief revision policies (or dispositions to revise their initial beliefs) can be dispensed with: the analysis can be carried out entirely in terms of the actual beliefs at the time of choice. This alternative approach is put forward by Bonanno (2013), where an epistemic characterization of backward induction is provided that does not rely on (objective or subjective) counterfactuals.

Epistemic foundations of other extensive-form solution concepts Because of space constraints, for extensive-form games we have restricted attention to the epistemic foundations of only one solution concept, namely backward induction. In the literature, epistemic conditions have been studied for additional solution concepts, such as extensive-form rationalizability (by Battigalli (1997) and Pearce (1984)), forward induction (by Battigalli and Siniscalchi (2002)), and perfect Bayesian equilibrium (by Bonanno (2011)). An account of part of this literature can be found in a paper by Perea (2012).

Belief revision The semantics for belief revision described in Section 9.8 has its roots in the well-known AGM theory which was introduced by Alchourrón et al. (1985). The AGM theory is a syntactic theory, whose semantic counterpart was first explored by Grove (1988). There is a vast literature on AGM belief revision.

For a recent overview see the special issue of the *Journal of Philosophical Logic* on *25 Years of AGM Theory* (Volume 40 (2), April 2012). The conditions under which there is a precise correspondence between the subjective counterfactual functions f_i described in Section 9.8 and the syntactic AGM theory are explored by Bonanno (2009).

Acknowledgements I am grateful to Elias Tsakas for helpful comments and suggestions.

References

Alchourrón, C., P. Gärdenfors, and D. Makinson (1985). On the logic of theory change: partial meet contraction and revision functions. *The Journal of Symbolic Logic 50*, 510–530.

Aumann, R. (1976). Agreeing to disagree. *The Annals of Statistics 4*, 1236–1239.

Aumann, R. (1987a). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica 55*, 1–18.

Aumann, R. (1987b). Game theory. In J. Eatwell, M. Milgate, and P. Newman (Eds.), *The New Palgrave, a dictionary of economics*, Volume 2, pp. 460–482. London: Macmillan.

Aumann, R. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior 8*, 6–19.

Aumann, R. (1998). On the centipede game. *Games and Economic Behavior 23*, 97–105.

Aumann, R. and A. Brandenburger (1995). Epistemic conditions for Nash equilibrium. *Econometrica 63*, 1161–1180.

Bach, C. and E. Tsakas (2014). Pairwise epistemic conditions for Nash equilibrium. *Games and Economic Behaviour 85*, 48–59.

Bacharach, M. (1987). A theory of rational decision in games. *Erkenntnis 27*, 17–55.

Balkenborg, D. and E. Winter (1997). A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics 27*, 325–345.

Baltag, A., S. Smets, and J. Zvesper (2009). Keep hoping for rationality: a solution to the backward induction paradox. *Synthese 169*, 301–333.

Barelli, P. (2009). Consistency of beliefs and epistemic conditions for Nash and correlated equilibria. *Games and Economic Behavior 67*, 363–375.

Barelli, P. and S. Galanis (2013). Admissibility and event rationality. *Games and Economic Behavior 77*, 21–40.

Battigalli, P. (1997). On rationalizability in extensive games. *Journal of Economic Theory 74*, 40–61.

Battigalli, P. and G. Bonanno (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics 53*, 149–225.

Battigalli, P., A. Di-Tillio, and D. Samet (2013). Strategies and interactive beliefs in dynamic games. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics. Theory and Applications: Tenth World Congress*. Cambridge: Cambridge University Press.

Battigalli, P. and M. Siniscalchi (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory 106*, 356–391.

Ben-Porath, E. (1997). Nash equilibrium and backwards induction in perfect information games. *Review of Economic Studies 64*, 23–46.

van Benthem, J. (2011). *Logical Dynamics of Information and Interaction.* Cambridge: Cambridge University Press.

Bernheim, D. (1984). Rationalizable strategic behavior. *Econometrica 52*, 1002–1028.

Bertrand, J. (1883). Théorie mathématique de la richesse sociale. *Journal des Savants 67*, 499–508.

Board, O. (2004). Dynamic interactive epistemology. *Games and Economic Behavior 49*, 49–80.

Bonanno, G. (1991). The logic of rational play in games of perfect information. *Economics and Philosophy 7*, 37–65.

Bonanno, G. (1996). On the logic of common belief. *Mathematical Logic Quarterly 42*, 305–311.

Bonanno, G. (2001). Branching time logic, perfect information games and backward induction. *Games and Economic Behavior 36*, 57–73.

Bonanno, G. (2008). A syntactic approach to rationality in games with ordinal payoffs. In G. Bonanno, W. van der Hoek, and M. Wooldridge (Eds.), *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Volume 3 of *Texts in Logic and Games*, pp. 59–86. Amsterdam University Press.

Bonanno, G. (2009). Rational choice and *AGM* belief revision. *Artificial Intelligence 173*, 1194–1203.

Bonanno, G. (2011). *AGM* belief revision in dynamic games. In K. Apt (Ed.), *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK XIII, New York, pp. 37–45. ACM.

Bonanno, G. (2013). A dynamic epistemic characterization of backward induction without counterfactuals. *Games and Economics Behavior 78*, 31–43.

Bonanno, G. (2014). Reasoning about strategies and rational play in dynamic games. In J. van Benthem, S. Ghosh, and R. Verbrugge (Eds.), *Modeling strategic reasoning*, Texts in Logic and Games. Springer. forthcoming.

Bonanno, G. and K. Nehring (1998). On Stalnaker's notion of strong rationalizability and Nash equilibrium in perfect information games. *Theory and Decision 45*, 291–295.

Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In P. Dasgupta, D. Gale, O. Hart, and E. Maskin (Eds.), *Economic analysis of markets and games*, pp. 282–290. MIT Press.

Brandenburger, A. (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory 35*, 465–492.

Brandenburger, A. and E. Dekel (1987). Rationalizability and correlated equilibria. *Econometrica 55*, 1391–1402.

Brandenburger, A., A. Friedenberg, and J. Keisler (2008). Admissibility in games. *Econometrica 76*, 307–352.

Clausing, T. (2003). Doxastic conditions for backward induction. *Theory and Decision 54*, 315–336.

Clausing, T. (2004). Belief revision in games of perfect information. *Economics and Philosophy 20*, 89–115.

Cournot, A. A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses.* Paris: Hachette.

de Bruin, B. (2010). *Explaining games: the epistemic programme in game theory.* Springer.

Dekel, E. and F. Gul (1997). Rationality and knowledge in game theory. In D. Kreps and K. Wallis (Eds.), *Advances in economics and econometrics*, pp. 87–172. Cambridge University Press.

Feinberg, Y. (2005). Subjective reasoning - dynamic games. *Games and Economic Behavior 52*, 54–93.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic 17*, 157–170.

Halpern, J. (2001). Substantive rationality and backward induction. *Games and Economic Behavior 37*, 425–435.

Harsanyi, J. (1967-1968). Games with incomplete information played by "Bayesian players", Parts I-III. *Management Science 8*, 159–182, 320–334, 486–502.

Hotelling, H. (1929). Stability in competition. *Economic Journal 39*, 41–57.

Kreps, D. (1988). *Notes on the theory of choice.* Boulder: Westview Press.

Kripke, S. (1963). A semantical analysis of modal logic I: normal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 9*, 67–96.

von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen 100*, 295–320.

von Neumann, J. and O. Morgenstern (1944). *Theory of games and economic behavior.* Princeton University Press.

Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica 52*, 1029–1050.

Perea, A. (2007a). Epistemic foundations for backward induction: an overview. In J. van Benthem, D. Gabbay, and B. Löwe (Eds.), *Interactive logic. Proceedings of the 7th Augustus de Morgan Workshop*, Volume 1 of *Texts in Logic and Games*, pp. 159–193. Amsterdam University Press.

Perea, A. (2007b). A one-person doxastic characterization of Nash strategies. *Synthese 158*, 251–271.

Perea, A. (2012). *Epistemic game theory: reasoning and choice.* Cambridge: Cambridge University Press.

Polak, B. (1999). Epistemic conditions for Nash equilibrium, and common knowledge of rationality. *Econometrica 67*, 673–676.

Rubinstein, A. (1991). Comments on the interpretation of game theory. *Econometrica 59*, 909–924.

Samet, D. (1996). Hypothetical knowledge and games with perfect information. *Games and Economic Behavior 17*, 230–251.

Samuelson, L. (1992). Dominated strategies and common knowledge. *Games and Economic Behavior 4*, 284–313.

von Stackelberg, H. (1934). *Marktform und Gleichgewicht.* Vienna: Julius Springer.

Stahl, D. (1995). Lexicographic rationalizability and iterated admissibility. *Economics Letters 47*, 155–159.

Stalnaker, R. (1994). On the evaluation of solution concepts. *Theory and Decision 37*, 49–74.

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy 12*, 133–163.

Stalnaker, R. (1998). Belief revision in games: forward and backward induction. *Mathematical Social Sciences 36*, 31–56.

Tan, T. and S. Werlang (1988). The Bayesian foundation of solution concepts of games. *Journal of Economic Theory 45*, 370–391.

Zermelo, E. (1913). Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. *Proceedings Fifth International Congress of Mathematicians 2*, 501–504.