3

The Role of Common Knowledge Assumptions in Game Theory

ADAM BRANDENBURGER AND EDDIE DEKEL

1. Introduction

The notion of common knowledge has been increasingly used in game theory, although usually in an informal, not fully appreciated or articulated, manner. By 'common knowledge' is meant the idea that something is not merely known by all the players in a game, but is also known to be known, known to be known to be known, and so on *ad infinitum*. The concept has been deployed with sufficient frequency to necessitate an assessment of its role. Thus the purpose of this chapter is to examine the common knowledge assumptions that underlie the various solution concepts in non-cooperative game theory.

Much of the work to be surveyed here is of recent origin in the context of game theory. It represents an attempt to integrate the theory of individual decision-making under uncertainty—often called Bayesian decision theory—with the theory of games. In so doing, the research marks a departure from the conventional view of the function of Bayesian decision theory. Traditionally, Bayesian decision theory has been perceived as appropriate only to 'exogenous' uncertainty, not to uncertainty about the actions of players in a game. For the latter, 'endogenous', variety, it has usually been argued that an equilibrium, or other game-theoretic, concept must be employed. Thus Harsanyi has written: 'every player i . . . will assign a *subjective* probability distribution P_i to all variables unknown to him—or at least to all unknown *independent* variables, i.e. to all variables not depending on the players' own strategy choices' (Harsanyi 1967–8: 167).

In contrast, the predominant theme of the work to be discussed in this chapter is that the players in a game assign subjective probabilities to *all* uncertainty, including the actions of other players. The distinction between the various game-theoretic solution concepts can then be seen to depend on the common knowledge assumptions respected by the players' subjective probabilities.

The basic datum for the players in a game is the structure of the game

We would like to thank Ken Binmore, Ben Polak, and Linda Pollock for helpful comments. We are indebted to Bob Aumann for providing detailed comments on an earlier version of this chapter and for sharing some of his thoughts on game theory with us.

itself, that is, the description of the payoff functions, strategy spaces, and so on. At first sight it would seem that, if the structure of a game is not taken to be common knowledge among the players, then any analysis is impossible. This seems to have been the viewpoint, at least implicitly, of early workers in game theory (see e.g. Luce and Raiffa 1957: Ch. 3). Nevertheless, in a seminal series of papers Harsanyi (1967–8) argued that, even if the structure of a game is not common knowledge, there is a well-defined larger game in which Nature first chooses the version of the game to be played, and this larger game *can* be taken to be common knowledge among the players. Harsanyi's arguments have recently been given formal mathematical expression by Böge and Eisele (1979), Mertens and Zamir (1985), and others. This work is described in Section 5 below.

The second crucial piece of common knowledge is that all the players are rational. By rationality is meant the assumption that each player conforms to the axioms of Savage (1954) (or to some related set of axioms), and hence acts to maximize expected utility calculated using some subjective probability distribution over all uncertainty that the player faces—which, as indicated above, will include the actions of the other players. Actually, as will be seen in Section 6, for the purpose of discussing refinements of Nash equilibrium, the theory of subjective expected utility is not sufficiently 'detailed'. At that point a more elaborate theory—subjective expected utility with lexicographic beliefs—will be described and rationality will be taken to mean that the players act in accordance with this modified theory. To sum up, the assumption will be made that the rationality of the players, in one or other sense, is indeed common knowledge.

What does common knowledge of rationality in a game setting amount to? The answer depends on what further *a priori* information there is about the setting. Sections 3 and 4 flesh out this observation. It is shown that, if there is no additional information, then the appropriate solution concept is rationalizability (Bernheim 1984; Pearce 1984); if the players share a common prior, then the correct concept is correlated equilibrium (Aumann 1974, 1987); if the players' beliefs are common knowledge, then common knowledge of rationality is equivalent to Nash equilibrium. Section 6 extends this last characterization to two refinements of Nash equilibrium: perfect equilibrium (Selten 1975) and proper equilibrium (Myerson 1978).

2. Formalizing the Notion of Common Knowledge

An event or proposition is common knowledge among a group of people if it is known to all, known to all that it is known to all, and so on *ad*

infinitum. The term 'common knowledge' was used in this context by Lewis (1969), who attributes the basic idea to Schelling (1960). Aumann (1976) proposed the notion independently and offered a precise mathematical formulation in terms of a model of differential information which is by now standard in economics.

There is a finite set Ω of *states of the world*. There are I individuals where each person $i \in I$ has a partition P^i of Ω, representing i's private information about the true state. That is, if $\omega \in \Omega$ is the true state, then i is informed of the element of P^i that contains ω (to be denoted $P^i(\omega)$). P denotes the partition that is the *finest common coarsening* (or meet) of $P^1, \ldots, P^n$. Write $P(\omega)$ for the element of P that contains ω. Aumann's definition of common knowledge can now be stated.

DEFINITION 1 (Aumann 1976). An event $E \subset \Omega$ is common knowledge at a state of the world $\omega \in \Omega$ if $P(\omega) \subset E$.

In order to relate this definition to the intuitive notion of common knowledge, it will be helpful to have the following definitions. Given an event $E \subset \Omega$, say, person i knows E at a state ω if $P^i(\omega) \subset E$. This captures the idea that i is informed that the true state lies in $P^i(\omega)$, and hence also in any set that contains $P^i(\omega)$. The event that i knows E, to be written K^iE, is then given by

$$K^iE = \{\omega \in \Omega : P^i(\omega) \subset E\}.$$

It is easy to check that K^i, considered as a function from 2^Ω to 2^Ω (where 2^Ω is the set of all subsets of Ω), has the following properties:

(P1) For any $E \subset \Omega$, $K^iE \subset E$;
(P2) For any $E \subset \Omega$, $K^iE \subset K^iK^iE$
(P3) For any $E, F \subset \Omega$, $K^i(E \cap F) = K^iE \cap K^iF$;
(P4) For any $E \subset \Omega$, $(K^iE)^c \subset K^i(K^iE)^c$ where the superscript c denotes complement.

(P1)–(P4) capture some intuitive (and some *not* so intuitive!) aspects of knowledge formalized in terms of partitions. (P1) says that i can know E only if E happens. (P2) says that if i knows E then i knows that i knows E. (P3) says that i knows E and F if and only if i knows E and i knows F. (P4) says that if i does not know E, then i knows that i does not know E.[1]

Suppose for simplicity that there are just two people, i and j. Using K^i and K^j (the latter being defined analogously to K^i using the partition P^j in place of P^i), it is now easy to write down directly the statement that E is

[1] Actually, (P1)–(P4) are characteristic of partition information in the following sense. If a function $K^i : 2^\Omega \to 2^\Omega$ satisfies (P1)–(P4) then there is a partition P^i of Ω such that $K^iE = \{\omega : P^i(\omega) \subset E\}$. To see this, define the class of sets $F^i = \{F \subset \Omega : K^iF = F\}$. F^i is a field. Let P^i be the partition which generates F^i and set $K^iE = U\{\pi \in P^i : \pi \subset E\}$. K^i defined in this way is easily seen to satisfy (P1)–(P4). (See Bacharach 1985).

common knowledge at ω. Let

$$L^iE = K^iE \cap K^iK^jE \cap K^iK^jK^iE \cap \ldots$$

$$L^jE = K^jE \cap K^jK^iE \cap K^jK^iK^jE \cap \ldots$$

Then one would say that E is common knowledge at ω if $\omega \in L^iE \cap L^jE$. It is straightforward to check that this definition is indeed equivalent to Aumann's definition.

PROPOSITION 1. $P(\omega) \subset E$ if and only if $\omega \in L^iE \cap L^jE$.

The proof of this proposition is essentially contained in the discussion of 'reachability' in Aumann (1976: 1237).

A potentially troublesome aspect of this formalization of common knowledge lies in the interpretation of the condition $P(\omega) \subset E$. To interpret this as the statement that E is common knowledge at ω, it must be assumed that the information partitions are themselves common knowledge in an informal sense. Aumann (1976, 1987) has argued that this is not really an extra assumption since, if a state of the world ω is to be *all-inclusive*, it should include lists of those other states ω' that are, for i and j respectively, indistinguishable from ω. That is, ω should describe the manner in which information is imparted to i and j.

The element of self-reference built into this line of argument raises the possibility of some version of the Liar's Paradox being applicable, and hence of a contradiction arising. In fact, problems of self-reference already arise in the single-person case. Properties in the spirit of (P1)–(P4) above, when combined with the axioms necessary for mathematics, lead to the 'Knower's Paradox' (Montague 1974). The nature of Aumann's argument has been much discussed (see e.g. Brandenburger and Dekel 1985; Gilboa 1986; Kaneko 1987; Tan and Werlang 1985).[2] On a related issue, Samet (1987) and Shin (1987) have explored an alternative model of knowledge that satisfies the properties (P1)–(P3) defined earlier, but not the rather less palatable (P4).

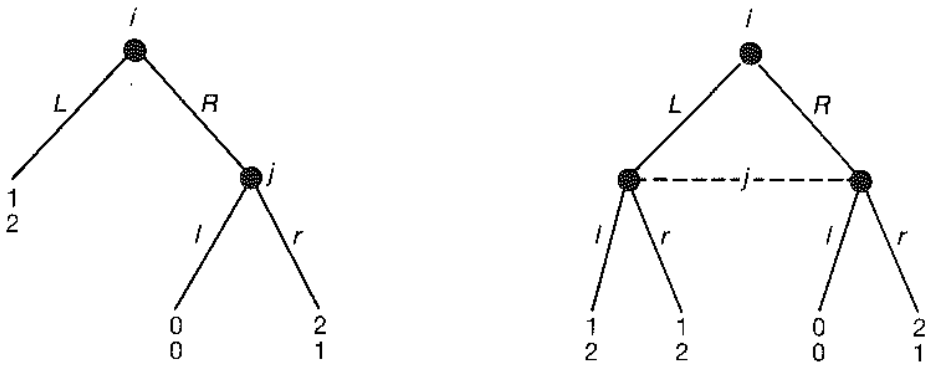
3. Common Knowledge of Rationality in Games

This section discusses how to formalize the idea of common knowledge of rationality in the context of games described in normal form. An n-person game in normal form is a $2n$-tuple $\Gamma = \langle A^1, \ldots, A^n; u^1, \ldots u^n \rangle$ where, for each $i = 1, \ldots, n$, A^i is a finite set of pure strategies (henceforth actions) of player i and $u^i : \times_{j=1}^n A^j \to R$ is i's payoff function (assigning a von Neumann–Morgenstern utility to each combination of strategies chosen by the players). In games in economics it is typically

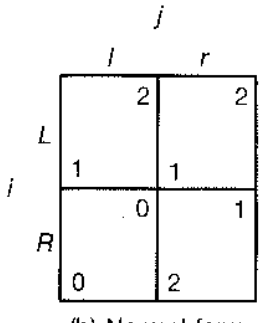
[2] There are also relevant literatures in computer science, artificial intelligence, linguistics, and philosophy—see Halpern (1986) and the references therein.

assumed that the players' strategy spaces are infinite; for example, player i's strategy space might be taken to be R^+ if i's strategy is to choose a price. But for present purposes it seems preferable to restrict attention to finite games, on the premise that extending many of the results to the infinite case raises technical rather than conceptual questions.

The discussion will also be confined to games in normal form rather than to the more familiar context of games in extensive form, or trees. Although the map from extensive- to normal-form games is many-to-one (see Figure 3.1), and thus an apparent loss of information is entailed in using the normal form, the adequacy of the normal form was the traditional position among game theorists (von Neumann and Morgenstern 1944). Such a stance must be founded, at least implicitly, on an argument of strategic equivalence, for example between the two game trees in the figure. This type of argument can be constructed from the work of Thompson (1952), Dalkey (1953), and Elmes and Reny (1987). These authors demonstrated that, given two game trees with the same normal form (up to duplicated pure strategies), one can be transformed



(a) Extensive forms



(b) Normal form

Fig. 3.1

into the other by a sequence of elementary and 'inessential' transformations of the game tree. (Kohlberg and Mertens 1986, from whom this discussion is derived, supply a description of the transformations.) If one is convinced that the transformations are inessential, then two trees with the same normal form must be equivalent. From this it follows that a 'good' solution concept should be defined on the normal rather than on the extensive form, in order to avoid any dependence on the presentation of the decision problem facing the players. A further point is worth emphasizing. Despite being defined on the normal form, a 'good' solution concept may well be required to conform to some notion of extensive-form rationality (e.g. backwards induction or sequential rationality—see Kreps and Wilson 1982: 272 and Krep's Chapter 2 above) in any tree with that normal form. Proper equilibrium (Myerson 1978) is such a solution concept—see Section 6.

Ultimately, the choice of normal versus extensive form is best viewed as a choice of axiom. It is of course possible to argue that the Dalkey/Thompson transformations are *not* inessential, in which case two trees with the same normal form may well have to be analysed separately. Chapter 2 above by Kreps surveys the literature on solution concepts defined on the extensive form. There is a final caveat to this discussion: the arguments in favour of the normal form certainly rely on the standard assumption that the players have unlimited computational power. As soon as one takes explicit account of computational complexity, it may be that a single n-way choice for a player is *not* equivalent to a sequence of $n-1$ binary choices (as is implied by a normal-form approach). Related arguments against a normal-form approach can be found in Binmore (1987b).

Returning to the objective of this section, the question to be addressed is, Which normal-form actions $a^i \in A^i$ can a player i choose under the assumption that the rationality of the players is common knowledge? As described in the introduction, by 'rationality of player i' is meant that i chooses an action to maximize expected utility calculated using some subjective probability distribution over the uncertainty that i faces. This uncertainty is the choice of actions by the other players. Given sets $X^1, \ldots, X^n$, let X^{-i} denote the set $X^1 \times \ldots \times X^{i-1} \times X^{i+1} \times \ldots \times X^n$. For any finite set X, let $\Delta(X)$ denote the set of probability distributions on X. Rationality requires that i choose an action a^i which solves

$$\max_{\bar{a}^i \in A^i} \sum_{a^{-i} \in A^{-i}} \sigma(a^{-i}) u^i(\bar{a}^i, a^{-i})$$

where $\sigma \in \Delta(A^{-i})$ is a subjective probability distribution over the actions of the other players. Following standard game-theoretic terminology, a^i will be called a best reply to σ. So the knowledge that i is a rational player immediately rules out certain actions of i, in that they are not best

replies to any $\sigma \in \Delta(A^{-i})$. Formally, let $A_0^i = A^i$ and

$$A_1^i = \{a^i \in A_0^i : a^i \text{ is a best reply to some } \sigma \in \Delta(A^{-i})\}.$$

A_1^i is the set of actions that a rational player i can choose. If i knows that each other player j is also rational, then i knows that j can only choose actions a^j in A_1^j (defined in analogous fashion to A_1^i). A probability distribution of i will be inconsistent with this knowledge if it assigns positive probability to an action $a^j \notin A_1^j$. So if i not only is rational, but also knows that every other player j is rational, i should only choose an action $a^i \in A_2^i$ where

$$A_2^i = \{a^i \in A_1^i : a^i \text{ is a best reply to some } \sigma \in \Delta(A_1^{-i})\}.$$

If the rationality of the players is assumed to be common knowledge, then this reasoning can be extended indefinitely. Define inductively

$$A_k^i = \{a^i \in A_{k-1}^i : a^i \text{ is a best reply to some } \sigma \in (A_{k-1}^{-i})\}.$$

By finiteness, there must be a K such that $A_k^i = A_K^i \neq \emptyset$ for all $k \geq K$. The sets $A_K^1, \ldots, A_K^n$ are the sets of actions that the players can choose under the assumption of common knowledge of rationality.

It should be emphasized that the structure of the game $\Gamma = \langle A^1, \ldots, A^n; u^1, \ldots, u^n \rangle$ itself, as well as the rationality of the players, is assumed to be common knowledge in deriving the sets A_K^i. If player i does not know each other player j's utility function u^j, then i cannot calculate the sets A_1^j, and hence cannot restrict him/herself to a probability distribution on A_1^{-i}. Extending this argument shows that the structure of Γ must be taken to be common knowledge.

At first sight, it appears somewhat perverse to be assuming that the utility functions of the players are common knowledge but that their beliefs are not. In Savage (1954), utility functions and subjective probability distributions are derived jointly from preferences over actions. If it is common knowledge of the players' preferences that gives rise to common knowledge of the utility functions, then why are not beliefs also common knowledge? Aumann (1987), Bernheim (1985), and Brandenburger and Dekel (1987) present expanded models with private information in which this objection does not bite. The players' *prior* probability distributions, as well as their utility functions, can be assumed to be common knowledge. But the players' beliefs about other players, which are their *posteriors* calculated using their private information, need not be common knowledge.

The sets $A_K^1, \ldots, A_K^n$ are the sets of actions of the players that remain after iterated deletion of strongly dominated actions. Recall that an action a^i of player i is strongly dominated if there is a $\sigma^i \in \Delta(A^i)$ such that

$$\sum_{\bar{a}^i \in A^i} \sigma^i(\bar{a}^i) u^i(\bar{a}^i, a^{-i}) > u^i(a^i, a^{-i}) \; \forall a^{-i} \in A^{-i}.$$

It is well known (e.g. Ferguson 1967; Pearce 1984: Appendix B) that a^i is strongly dominated if and only if there is no $\sigma \in \Delta(A^{-i})$ to which a^i is a best reply. Hence the sets $A_1^1, \ldots, A_1^n$ are the sets of actions remaining after deletion of strongly dominated actions. Proceeding inductively, $A_k^1, \ldots, A_k^n$ are the sets of actions remaining after k rounds of deletion.

Arguments based on dominance and iterated dominance are nothing new in game theory (see Luce and Raiffa 1957: 108–9), but, rather surprisingly, it is only recently that the foundations have been spelled out in detail by Bernheim (1984, 1985), Pearce (1984), and Tan and Werlang (1988). In fact, the 'rationalizable' actions of Bernheim and Pearce are not quite the same as the sets $A_K^1, \ldots, A_K^n$, the difference being that Bernheim and Pearce require it to be common knowledge that each player's probability distribution on the actions of the other players is stochastically independent. While this makes no difference in the context of two-person games (since a player faces only one opponent), there are three-person games in which the rationalizable actions are proper subsets of $A_K^1, \ldots, A_K^n$. Stochastic independence may be an appropriate assumption in a scenario where each player supposes that his/her opponents are selected independently and play independently. However, in many situations it seems more natural to allow for correlated beliefs.

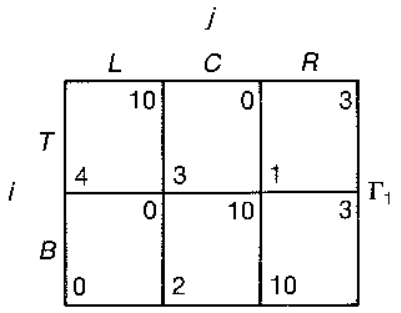


Fig. 3.2

In some games, iterated dominance arguments are successful in the sense of reducing the set of possible actions for each player to a singleton. Consider for example the game Γ_1 of Figure 3.2. R is strongly dominated by the mixed strategy of j which assigns probability 1/2 to L, 1/2 to C. In the reduced game, after deletion of R, T strongly dominates B. Finally, after deletion of B, L strongly dominates C. So iterated dominance leads to i playing T and j playing L. Nevertheless, iterated dominance is often a 'weak' solution concept, placing few or no restrictions on which actions the players can choose. In the game Γ_2 of Figure 3.3 (based on a game in Bernheim 1984), iterated dominance does not eliminate any actions of i or j from consideration.

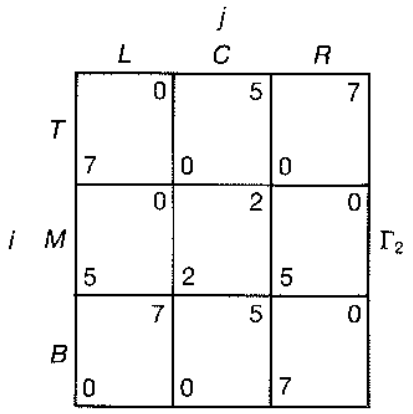


Fig. 3.3

4. Nash Equilibrium

Consider again the game Γ_2. Player i can justify choosing T by the hierarchy of beliefs: i believes j chooses L, i believes j believes i chooses B, i believes j believes i believes j chooses R, and so on in the cycle T–L–B–R–T–. . . . Notice that i's choice of T can be justified only if i believes that j's belief about i's action is wrong: if j believed that i's choice was T then j would choose R, to which T is not a best reply for i. Although this 'inconsistency' seems to be an accurate reflection of the players' ignorance about each other's beliefs, such a situation can be eliminated by supposing that the players' beliefs are common knowledge. This assumption leads to a characterization of Nash equilibrium (and in the game Γ_2 singles out the unique Nash equilibrium, in which i plays M and j plays C). In order to state the characterization, a preliminary result (Proposition 2 below) will be needed. Write player i's belief over j's choice of action as $\sigma^i \in \Delta(A^j)$, and j's belief over i's choice of action as $\sigma^j \in \Delta(A^i)$.

PROPOSITION 2. Assume that the beliefs $(\sigma^i, \sigma^j) \in \Delta(A^j) \times \Delta(A^i)$ are common knowledge. Then common knowledge of rationality is satisfied if and only if

$$\sigma^i(a^i) > 0 \Rightarrow a^i \text{ is a best reply to } \sigma^i \tag{1}$$

$$\sigma^j(a^j) > 0 \Rightarrow a^j \text{ is a best reply to } \sigma^j. \tag{2}$$

Proof. To prove this proposition, assume first that σ^i, σ^j are common knowledge and that common knowledge of rationality is satisfied. If i assigns positive probability to an action a^j of j, that is, if $\sigma^j(a^j) > 0$, then by common knowledge of rationality a^j must be optimal for j given some belief over A^i. But since beliefs are common knowledge, this belief is just σ^i, so condition (2) is satisfied. A similar argument establishes condition (1). To prove the converse direction, assume that σ^i, σ^j are common

knowledge and that conditions (1), (2) hold; i knows that j is rational since, by condition (2), i's belief σ^j assigns positive probability only to actions a^j of j which are optimal for j given the belief σ^j which i knows j to have. Continuing in this fashion establishes that all sentences of the form 'i (or j) knows that j (or i) . . . is rational' are true. So common knowledge of rationality is satisfied. □

Recall the conventional definition of a Nash equilibrium: it is a pair of *mixed strategies* $(\sigma^i, \sigma^j) \in \Delta(A^i) \times \Delta(A^j)$ such that σ^i is optimal against σ^j and σ^j is optimal against σ^i. It is easy to check that σ^i is optimal against σ^j (resp. σ^j is optimal against σ^i) if and only if condition (1) (resp. (2)) holds. Hence the following result follows immediately from Proposition 2.

COROLLARY 1. Assume that the beliefs $(\sigma^i, \sigma^j) \in \Delta(A^i) \times \Delta(A^j)$ are common knowledge. Then common knowledge of rationality is satisfied if and only if (σ^i, σ^j) is a Nash equilibrium.

The characterization of Nash equilibrium contained in Corollary 1 can be extended to n-person games $(n > 2)$ by supposing that there are commonly known beliefs $(\sigma^1, \ldots, \sigma^n) \in \times_{i=1}^n \Delta(A^i)$. In writing the beliefs this way, two assumptions are being made: first, that, for any player i, all players other than i share the same belief σ^i about i's choice of action; second, that (as for rationalizability) each player's probability distribution on the actions of the other players is stochastically independent.

Corollary 1 offers a characterization of Nash equilibrium which seems preferable to the conventional interpretation of Nash equilibrium. According to the orthodox view, player i actually performs a randomization over the set of actions A^i in accordance with the probabilities prescribed by σ^i. Player i is prepared to carry out this randomization provided player j performs the randomization σ^j. But i is indifferent between the precise randomization and any other randomization over the same set of actions, or indeed between σ^i and choosing for sure any action assigned positive probability by σ^i. Why in fact should i perform the randomization σ^i? Yet if i does 'deviate' from σ^i, then j may no longer be prepared to play σ^j. The lack of a clear rationale for a player to randomize is a serious drawback of the conventional view of Nash equilibrium. The view espoused here is to think in terms of an equilibrium of *beliefs* rather than of strategies: σ^i is no longer a randomization of player i but rather reflects the uncertainty of player j over i's choice of action.

An early attempt to advance this viewpoint of equilibrium was made in Harsanyi (1973a). In Harsanyi's formulation a mixed-strategy equilibrium is interpreted as a pure-strategy equilibrium in an augmented game where there is some exogenous uncertainty. The notion of an equilibrium of beliefs is developed in Aumann (1987). Aumann shows that, under the

assumption that the players share a common prior (the Common Prior Assumption), his 1974 concept of objective correlated equilibrium can be viewed as a consequence of common knowledge of rationality.

5. Games with Incomplete Information

It was assumed in Sections 3 and 4 that the structure of the game is common knowledge among the players. However, this may not always be the case: players may begin a game with different private information about their possibilities, preferences, and the like. The term 'incomplete information' was introduced by von Neumann and Morgenstern (1944) to describe games of this type, and the key work on these games is that by Harsanyi (1967–8).

Incomplete information about the game may arise in several ways: it may concern how many players there are in the game, the spaces of actions of the players, how the outcome of the game depends on the actions chosen, the players' preferences. Harsanyi (1967–8: 167–8) argued that all these cases can be reduced to uncertainty about the payoff functions by appropriate expansion of the number of players and spaces of actions. So, following Harsanyi, let the (expanded) number of players be n and, for each $i = 1, \ldots, n$, let A^i be i's (expanded) finite set of actions and $v^i : \times_{j=1}^{n} A^j \times S \to R$ be i's payoff function where S is a space of unknown parameters.

Let us try to analyse this game. Player i's optimal choice of action depends on what i thinks each player $j \neq i$ will do—and this depends on what i thinks each other player j's payoff function is. But what player j will do depends in turn on what j thinks are the payoff functions of the players $k \neq j$. And so on. This argument leads to each player having an infinite hierarchy of beliefs—over S, over the other players' beliefs over S, and so on.

The basic result for games with incomplete information is that a model of infinite hierarchies of beliefs can be closed in the following sense. For each player i there is a well-defined space T^i of all possible infinite hierarchies of beliefs of i such that T^i is homeomorphic to $\Delta(S \times T^{-i})$. That is, it is possible to summarize an infinite hierarchy of beliefs of i in a single object, namely, i's type $t^i \in T^i$, which is associated with a joint probability distribution over the parameter space S and the types of the other players. The proof of this result when S is compact is contained in Mertens and Zamir (1985); for an alternative proof which covers the case when S is a complete separable metric space, see Brandenburger and Dekel (1985). (See also Myerson 1985.) It is worth pointing out that, even if the parameter space S is finite, the type spaces T^i are uncountably

[3] The notation $\Delta(X)$ has not been formally defined when X is infinite. $\Delta(S \times T^{-i})$ denotes the set of all probability measures on the Borel field of $S \times T^{-i}$; see Billingsley (1968) for definitions.

infinite.[3] In fact, it is clear that the type spaces T^i could not be finite: $\Delta(S \times T^{-i})$ is uncountably infinite and so could not be homeomorphic to a finite T^i. It is only once the type spaces are uncountable that they can be used to 'encode' the set of probability distributions on themselves.

The tools with which to analyse a game with incomplete information are now assembled. Each type $t^i \in T^i$ of player i induces a belief $q^i(t^i)$ over $S \times T^{-i}$. The parameter s can be integrated out by defining new payoff functions $u^i(a, t)$ and beliefs $p^i(t^i)$ by

$$u^i(a, t) = \int_S v^i(a, s) q^i(t^i)(ds \mid t^{-i})$$

$$p^i(t^i) = \text{marg}_{T^{-i}} q^i(t^i).$$

(Here $\text{marg}_{T^{-i}} q^i(t^i)$ denotes the marginal of $q^i(t^i)$ on T^{-i}.) A game with incomplete information, often called a *Bayesian game*, is then a $4n$-tuple $\Gamma_B = \langle A^1, \ldots, A^n;\ T^1, \ldots, T^n;\ p^1, \ldots, p^n;\ u^1, \ldots, u^n \rangle$ where, for each i and $t^i \in T^i$, $p^i(t^i)$ is a probability distribution over T^{-i} and $u^i : \times_{j=1}^n A^j \times \times_{j=1}^n T^j \to R$. Unlike the game Γ we began with, the structure of Γ_B *can* be assumed to be common knowledge among the players.

The Bayesian game Γ_B in its full generality is too unwieldy a tool to be useful in practice. In applications it is typically assumed that the players' types lie in a finite 'belief-closed' subset of $T^1 \times \ldots \times T^n$. A finite subset $V^1 \times \ldots \times V^n$ of $T^1 \times \ldots \times T^n$ is said to be belief-closed if, for every i and each $t^i \in V^i$, $p^i(t^i)(V^{-i}) = 1$. That is, each player knows that all the players' types lie in $V^1 \times \ldots \times V^n$, each player knows that all the players know this, and so on. In other words, it is assumed to be common knowledge that the players' types lie in $V^1 \times \ldots \times V^n$.

6. Refinements of Nash Equilibrium

In many games it is agreed that some Nash equilibria are more 'reasonable' than others. There is by now a considerable literature on refinements of Nash equilibrium, which is concerned with formulating criteria for choosing among the set of Nash equilibria. Starting with Selten (1965), one strand of this literature focuses on refinements of the extensive form; the other on refinements on the normal form. This section will be predominantly, though not exclusively, concerned with the second approach, since Chapter 2 above provides a detailed discussion of the first. Some simple examples will be provided to motivate refining the set of Nash equilibria, and some of the proposed solution concepts will be reviewed. Finally, in keeping with aims of this chapter, some recent work will be described revealing how refinements of Nash equilibrium can be understood as a consequence of common knowledge of rationality—albeit of a different type of rationality from that described in Section 3.

The basic idea behind normal-form refinements is to use 'trembles' to

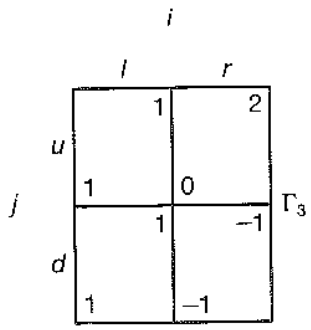


Fig. 3.4

rule out Nash equilibria that are not robust to small perturbations, in particular, equilibria that involve weakly dominated strategies. Consider for example the game Γ_3 of Figure 3.4. (l, d) is a Nash equilibrium but does not seem robust. If j thinks there is any chance of i choosing r, then j would strictly prefer u to d. Expressed differently, d is weakly dominated by u. The only sensible Nash equilibrium of Γ_3 is (r, u). The idea of a (trembling-hand) perfect equilibrium was introduced by Selten (1975) in order to rule out equilibria such as (l, d).[4]

In order to state Selten's definition, some extra notation will be useful. Given a game $\Gamma = \langle A^1, \ldots, A^n; u^1, \ldots, u^n \rangle$, let $v^i(a^i, \sigma^{-i})$ denote i's expected payoff from choosing the action a^i when the other players choose the mixed strategies $\sigma^{-i} \in \times_{j \neq i} \Delta(A^j)$. For any finite set X, let $\Delta^0(X)$ denote the set of all strictly positive probability distributions on X. A mixed strategy of i is called completely mixed if it lies in $\Delta^0(A^i)$.

DEFINITION 2 (Selten 1975). An n-tuple of completely mixed strategies $(\sigma^1, \ldots, \sigma^n) \in \times_{i=1}^n \Delta^0(A^i)$ is an ε-perfect equilibrium (for $\varepsilon > 0$) if, for each i and every $a^i, \tilde{a}^i \in A^i$, $v^i(a^i, \sigma^{-i}) < v^i(\tilde{a}^i, \sigma^{-i})$ implies $\sigma^i(a^i) < \varepsilon$.

So an ε-perfect equilibrium is an n-tuple of mixed strategies such that every pure strategy receives positive probability, but only best replies get more than ε weight.

DEFINITION 3 (Selten 1975). A perfect equilibrium is a limit (as $\varepsilon \to 0$) of ε-perfect equilibria.

Selten showed that a perfect equilibrium is a Nash equilibrium, and every finite game possesses a perfect equilibrium.

Perfect equilibrium behaves as desired on the game Γ_3: in any ε-perfect equilibrium d must receive weight less than ε so the unique perfect equilibrium is (r, u).

Γ_3 is also the normal form of the game in Figure 2.3 in Kreps's chapter.

[4] In fact, most of Selten's paper is concerned with developing a notion of perfection for extensive-form games. The reference is to Section 13 of Selten (1975), which deals with normal-form games.

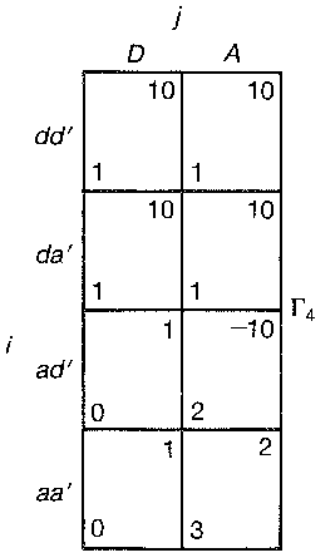


Fig. 3.5

There the Nash equilibrium (l, d) was ruled out by a backwards induction argument. This suggests that there is a close relationship between robustness/weak dominance arguments on the normal form and backwards induction/sequentiality arguments on the extensive form. Indeed, one may ask whether a perfect equilibrium in the normal form always gives rise to backwards induction, or perhaps even sequentially rational, behaviour in the extensive form. The answer is no, as the game Γ_4 of Figure 3.5 shows. (dd', D) is a perfect equilibrium (it is the limit of $((1-3\varepsilon, \varepsilon, \varepsilon, \varepsilon), (1-\varepsilon, \varepsilon))$ strategies). Γ_4 is the normal form of the game in Figure 2.9 in Kreps's chapter, which has the unique backwards induction solution (aa', A). As Kreps discusses, the equilibrium (dd', D) can be ruled out by arguing that a costlier 'mistake' by a player is infinitely less likely than a less costly one. In Γ_4, given j's strategy $(1-\varepsilon, \varepsilon)$, ad' is a costlier mistake for i than aa'. But, once i puts much less weight on ad' than aa', j will prefer A to D and the equilibrium (dd', D) will collapse. A normal-form refinement that captures this line of reasoning is the notion of proper equilibrium due to Myerson (1978).

DEFINITION 4 (Myerson 1978). $(\sigma^1, \ldots, \sigma^n) \in \times_{i=1}^n \Delta^0(A^i)$ is an ε-proper equilibrium (for $\varepsilon > 0$) if for each i and every $a^i, \bar{a}^i \in A^i$, $v^i(a^i, \sigma^{-i}) < v^i(\bar{a}^i, \sigma^{-i})$ implies $\sigma^i(a^i) < \varepsilon\ \sigma^i(\bar{a}^i)$. A proper equilibrium is a limit (as $\varepsilon \to 0$) of ε-proper equilibria.

It is not hard to show that the unique proper equilibrium of Γ_4 is (aa', A). In fact, one can demonstrate: a proper equilibrium of a normal-form game gives rise to a sequential equilibrium in any tree with that normal form. For a precise statement of this relationship and proofs, see van Damme (1982) and Kohlberg and Mertens (1986). The latter

paper should also be consulted for a comprehensive discussion of the issue of refinements and for the definitions of further refinements (hyperstable, fully stable, and stable sets of equilibria).

The ideas in this section might appear somewhat removed from the strictly 'decision-theoretic' approach taken in the preceding sections. The emphasis there was on understanding solution concepts in game theory from the perspective of single-person decision theory: players were assumed to be rational in the sense of Savage (1954), and different solution concepts were characterized in terms of varying common knowledge assumptions. But how are the 'trembles' and 'mistakes' of this section to be interpreted within the context of decision theory? Two recent papers (Blume 1986 and Brandenburger and Dekel 1986) have shown that perfect and proper equilibrium can in fact be understood as a consequence of common knowledge of a modified form of rationality *à la* Savage.

In Section 3 rationality of a player i was taken to mean that i has a subjective probability distribution over the actions of the other players, and chooses an action to maximize expected utility calculated using this distribution. This type of behaviour is justified by supposing that player i conforms to the axioms of subjective expected utility in Savage (1954). The papers by Blume and by Brandenburger and Dekel propose a different set of axioms which leads to an alternative theory of decision-making under uncertainty—subjective expected utility with lexicographic beliefs. According to this theory, player i has some finite hierarchy of subjective probability distributions over the actions of the other players, say $(\sigma_1, \ldots, \sigma_K)$, where $\sigma_k \in \Delta(A^{-i})$ for each $k = 1, \ldots, K$, with the property that, for each $a^{-i} \in A^{-i}$, $\sigma_k(a^{-i}) > 0$ for some k. i chooses an action $a^i \in A^i$ over another action $\tilde{a}^i \in A^i$ if the first yields a higher subjective expected utility in a lexicographic sense, that is if

$$\left[\sum_{a^{-i} \in A^{-i}} \sigma_k(a^{-i}) u^i(a^i, a^{-i})\right]_{k=1}^{k} >^L \left[\sum_{a^{-i} \in A^{-i}} \sigma_k(a^{-i}) u^i(\tilde{a}^i, a^{-i})\right]_{k=1}^{K}.$$

Here $>^L$ means that the first vector is lexicographically greater than the second. In other words, player i has a 'primary' belief σ_1 over the actions of the other players and computes expected utilities using σ_1. Only if two actions are deemed indifferent under σ_1 does i then consult his/her 'secondary' belief σ_2 and compute expected utilities using σ_2, and so on.

The connection of these ideas with refinements of Nash equilibrium is apparent from the condition on the hierarchy $(\sigma_1, \ldots, \sigma_K)$ that each $a^{-i} \in A^{-i}$ must be assigned positive probability by some σ_k. This condition says that player i considers all choices of actions by the other players to be possible, although actions assigned positive probability by σ_1 are infinitely more likely than those assigned zero probability by σ_1 but positive probability by σ_2, and so on.

Subjective expected utility with lexicographic beliefs can be used to provide axiomatic characterizations of perfect and proper equilibrium which are analogous to the characterization of Nash equilibrium discussed in Section 4. Both refinements arise from the assumption that the players' lexicographic hierarchies of beliefs are common knowledge—the distinction between perfect and proper equilibrium lies in the precise way in which common knowledge of rationality is formulated. Blume (1986) and Brandenburger and Dekel (1986) should be consulted for further details of the characterizations.

7. Concluding Remarks

The work surveyed in this chapter only analyses certain 'coherent' situations of common knowledge in games and does not explain how the common knowledge itself arises. As in any analysis founded on Bayesian decision theory, the issue of formation of beliefs is side stepped. This deficiency could be remedied by supplementing the formal analysis with either an informal model or a model of learning behaviour by the players. In the first case, the analyst must evaluate the economic or other situation being modelled and judge which common knowledge assumptions seem appropriate. In this respect this chapter complements the one by Kreps, which emphasizes the importance of making a choice, depending on the context, among alternative theories of out-of-equilibrium behaviour in extensive-form games.

The second approach would aim to develop an understanding of how a player learns to predict another player's choice of action. This appears to lead to a new version of the old infinite regress problem: I learn that you learn that I learn . . . . Introducing elements of bounded rationality and computational complexity into the analysis may help to avoid this pitfall.[5] The paper in this volume by Hahn (Chapter 5) contains examples of learning behaviour which highlight the problems in this area. Although this route will probably prove to be the more fruitful, at present there are few results along these lines.

[5] The recent work on applying automata theory to repeated games (Neyman 1985, Rubinstein 1986 and others) marks a first step in bringing complexity theory to bear on game theory.