Stanford Encyclopedia of Philosophy

Library of Congress Catalog Data
ISSN: 1095-5054

Stanford Encyclopedia of Philosophy
Copyright © 2016 by the publisher
The Metaphysics Research Lab
Center for the Study of Language and Information
Stanford University, Stanford, CA 94305

Epistemic Foundations of Game Theory

First published Fri Mar 13, 2015

Foundational work in game theory aims at making explicit the assumptions that underlie the basic concepts of the discipline. Non-cooperative game theory is the study of individual, rational decision making in situations of strategic interaction. This entry presents the *epistemic* foundations of non-cooperative game theory (this area of research is called *epistemic game theory*).

Epistemic game theory views rational decision making in games as something not essentially different from rational decision making under uncertainty. As in Decision Theory (Peterson 2009), to choose rationally in a game is to select the "best" action in light of one's beliefs or information. In a decision problem, the decision maker's beliefs are about a passive state of nature, the state of which determines the consequences of her actions. In a game, the consequences of one's decision depend on the choices of the *other* agents involved in the situation (and possibly the state of nature). Recognizing this—i.e., that one is interacting with other agents who try to choose the best course of action in the light of their own beliefs—brings *higher-order information* into the picture. The players' beliefs are no longer about a passive or external environment. They concern the choices *and the information* of the other players. What one expects of one's opponents depends on what one thinks the others expect from her, and what the others expect from a given player depends on what they think her expectations about them are.

This entry provides an overview of the issues that arise when one takes this broadly decision-theoretic view on rational decision making in games. After some general comments about information in games, we present the formal tools developed in epistemic game theory and epistemic logic that have been used to understand the role of higher-order information in interactive decision making. We then show how these tools can be used to characterize known "solution concepts" of games in terms of rational decision making in specific informational contexts. Along the way, we highlight a number of philosophical issues that arise in this area.

1. The Epistemic View of Games

1.1 Classical Game Theory

A *game* refers to any interactive situation involving a group of *self-interested* agents, or players. The defining feature of a game is that the players are engaged in an "*interdependent* decision problem" (Schelling 1960). Classically, the mathematical description of a *game* includes following components:

1. The *players*. In this entry, we only consider games with a finite set of players. We use N to denote the set of players in a game, and $i, j, \ldots$ to denote its elements.

2. The *feasible* options (typically called *actions* or *strategies*) for each player. Again, we only consider games with finitely many feasible options for each player.

3. The players' *preferences* over possible outcome. Here we represent them as von Neumann-Morgenstern utility functions u_i assigning real-valued utilities to each outcome of the game.

A game can have many other *structural properties*. It can be represented as a single-shot or multi-stage decision problem, or it can include simultaneous or stochastic moves. We start with games in *strategic form* without stochastic moves, and will introduce more sophisticated games as we go along in the entry. In a strategic game, each player i can choose from a (finite) set S_i of options, also called actions or strategies. The combination of all the players' choices, denoted $\mathbf{s}$, is called a **strategy profile**, or outcome of the game. We write $\mathbf{s}_i$ for i's component in $\mathbf{s}$, and $\mathbf{s}_{-i}$ for the profile of strategies for all agents other than i. Finally, we write $\Pi_{i \in N} S_i$ for the set of all strategy profiles of a given game. Putting everything together, a strategic game is a tuple $\langle N, \{S_i, u_i\}_{i \in N}\rangle$ where N is a finite set of players, for each $i \in N$, S_i is a finite set of actions and $u_i : \Pi_{i \in N} S_i \to \mathbb{R}$ is player i's utility function.

The game in Figure 1 is an example of a game in strategic form. There are two players, Ann and Bob, and each has to choose between two options: $N = \{Ann, Bob\}$, $S_{Ann} = \{u, d\}$ and $S_{Bob} = \{l, r\}$. The value of u_{Ann} and u_{Bob}, representing their respective preferences over the possible outcomes of the game, are displayed in the cell of the matrix. If Bob chooses l, for instance, Ann prefers the outcome she would get by choosing u to the one she would get by choosing d, but this preference is reversed in the case Bob chooses r. This game is called a "pure coordination game" in the literature because the players have a clear interest in coordinating their choices—i.e., on (u, l) or (d, r)—but they are indifferent about which way they coordinate their choices.

		Bob	
		l	r
Ann	u	1,1	0,0
	d	0,0	1,1

Figure 1: A coordination game

In a game, no single player has total control over which outcome will be realized at the end of the interaction. This depends on the decisions of *all players*. Such abstract models of *interdependent decisions* are capable of representing a whole array of social situations, from strictly competitive to cooperative ones. See Ross (2010) for more details about classical game theory and key references.

The central analytic tool of classical game theory are *solution concepts*. They provide a top-down perspective specifying which outcomes of a game are deemed "rational". This can be given both a *prescriptive* or a *predictive* reading. Nash equilibrium is one of the most well-known solution concepts, but we will encounter others below. In the game above, for instance, there are two Nash equilibria in so-called "pure strategies."[1] These are the two coordination profiles: (u, l) and (d, r).

From a prescriptive point of view, a solution concept is a set of practical recommendations—i.e., recommendations about what the players should do in a game. From a predictive point of view, solution concepts describe what the players will actually do in certain interactive situation. Consider again the pure strategy Nash equilibria in the above example. Under a prescriptive interpretation, it singles out what players *should* do in the game. That is, Ann and Bob should either play their component of (u,l) or (d,r). Under the predictive interpretation, these profiles are the ones that one would expect to observe in a actual play of that game.

This solution-concept-driven perspective on games faces many foundational difficulties, which we do not survey here. The interested reader can consult Ross (2010), Bruin (2010), and Kadane & Larkey (1983) for a discussion.

1.2 Epistemic Game Theory

Epistemic game theory is a broad area of research encompassing a number of different mathematical frameworks that are used to analyze games. The details of the frameworks are different, but they do share a common perspective. In this Section, we discuss two key features of this common perspective.

(1) *Epistemic game theory takes a broadly* Bayesian *perspective on decision-making in strategic situations.*

This point of view is nicely explain by Robert Stalnaker:

> There is no special concept of rationality for decision making in a situation where the outcomes depend on the actions of more than one agent. The acts of other agents are, like chance events, natural disasters and acts of God, just facts about an uncertain world that agents have beliefs and degrees of belief about. The utilities of other agents are relevant to an agent only as information that, together with beliefs about the rationality of those agents, helps to predict their actions. (Stalnaker 1996: 136)

In other words, epistemic game theory can be seen as an attempt to bring back the theory of decision making in games to its decision-theoretic roots.

In decision theory, the decision-making units are individuals with preferences over the possible consequences of their actions. Since the consequence of a given action depend on the state of the environment, the decision-maker's beliefs about the state of the environment are crucial to assess the rationality of a particular decision. So, the formal description of a decision problem includes the possible outcomes and states of the environment, the decision maker's preferences over these outcome, *and* a description of the decision maker's *beliefs* about the state of nature (i.e., the decision maker's *doxastic state*). A decision-theoretic *choice rule* can be used to make recommendations to the decision maker about what she *should* choose (or to predict what the decision-maker *will* choose). A standard example of a choice rule is maximization of (subjective) expected utility, underlying the *Bayesian* view of rationality. It presupposes that the decision maker's preferences and beliefs can be represented by numerical utilities and probabilities, respectively.[2] (We postpone the formal representation of this, and the other choice rules such as weak and strict dominance, until we have presented the formal models of beliefs in games in Section 2.)

From an epistemic point of view, the classical ingredients of a game (players, actions, outcomes, and preferences) are thus not enough to formulate recommendations or predictions about how the players should or will choose. One needs to specify the (interactive) decision problem the players are in, i.e., also the *beliefs* players have about each other's possible actions (and beliefs). In a terminology that is becoming increasingly popular in epistemic game theory, games are played in specific *contexts* (Friedenberg & Meier 2010, Other Internet Resources), in which the players have specific knowledge and/or beliefs about each other. The recommendations and/or predictions that are appropriate for one context may not transfer to another, even if the underlying situation may correspond to precisely the same strategic game.

(2) *In epistemic game theory, uncertainty about opponents' strategies takes center-stage.*

There are various types of information that a player has access to in a game situation. For instance, a player may have

- imperfect information about the play of the game (which moves have been played?);
- incomplete information about the structure of the game (what are the actions/payoffs?);
- strategic information (what will the other players do?); or
- higher-order information (what are the other players thinking?).

While all types of uncertainty may play a role in an epistemic analysis of a game, a distinguishing feature of epistemic game theory is an insistence that rational decisions are assessed in terms of the players' preferences *and* beliefs about what their opponents are going to do. Again we turn to Stalnaker to summarize this point of view:

> …There are no special rules of rationality telling one what to do in the absence of degrees of belief [about the opponents' choices], except this: decide what you believe, and then maximize expected utility. (Stalnaker 1996: 136)

The four types of uncertainty in games introduced above are conceptually important, but not necessarily exhaustive nor mutually exclusive. John Harsanyi, for instance, argued that all uncertainty about the structure of the game, that is all possible incompleteness in information, can be reduced to uncertainty about the payoffs (Harsanyi 1967–68). (This was later formalized and proved by Stuart and Hu 2002). In a similar vein, Kadane & Larkey (1982) argue that only strategic uncertainty is relevant for the assessment of decision in game situations. Contemporary epistemic game theory takes the view that, although it may ultimately be reducible to strategic uncertainty, making higher-order uncertainty explicit can clarify a great deal of what interactive or strategic rationality means.

The crucial difference from the classical "solution-concept" analysis of a game is that epistemic game theory takes a bottom-up perspective. Once the context of the game is specified, the rational outcomes are derived, given assumptions about how the players are making their choices and what they know and believe about how the others are choosing. In the remainder of this section, we briefly discuss some general issues that arise from taking an epistemic perspective on games. We postpone discussion of higher-order and strategic uncertainty until Sections 3, 4 and 5.

1.3 Stages of Decision Making

It is standard in the game theory literature to distinguish three stages of the decision making process: *ex ante*, *ex interim* and *ex post*. At one extreme is the *ex ante* stage where no decision has been made yet. The other extreme is the *ex post* stage where the choices of all players are openly disclosed. In between these two extremes is the *ex interim* stage where the players have made their decisions, but they are still uninformed about the decisions and intentions of the other players.

These distinctions are not intended to be sharp. Rather, they describe various stages of information disclosure during the decision-making process. At the *ex-ante* stage, little is known except the structure of the game, who is taking part, and possibly (but not necessarily) some aspect of the agents' character. At the *ex-post* stage the game is basically over: all player have made their decision and these are now irrevocably out in the open. This does not mean that all uncertainty is removed as an agent may remain uncertain about what exactly the others were expecting of her. In between these two extreme stages lies a whole gradation of states of information disclosure that we loosely refer to as "the" *ex-interim* stage. Common to these stages is the fact that the agents have made *a* decision, although not necessarily an irrevocable one.

In this entry, we focus on the *ex interim* stage of decision making. This is in line with much of the literature on the epistemic foundations of game theory as it allows for a straightforward assessment of the agents' rationality given their expectations about how their opponents will choose. Focusing on the *ex interim* stage does raise some interesting questions about possible *correlations* between a player's strategy choice, what Stalnaker (1999) calls "active knowledge", and her information about the choices of others, her "passive knowledge" (*idem*). The question of how a player should react, that is eventually revise her decision, upon learning that she did not choose "rationally" is an interesting and important one, but we do not discuss it in the entry. Note that this question is different from the one of how agents should revise their beliefs upon learning that *others* did not choose rationally. This second question is very relevant in games in which players choose sequentially, and will be addressed in Section 4.2.3.

1.4 Incomplete Information

A natural question to ask about *any* mathematical model of a game situation is *how does the analysis change if the players are uncertain about some of the parameters of the model?* This motivated Harsanyi's fundamental work introducing the notion of a game-theoretic **type** and defining a **Bayesian game** in Harsanyi 1967–68. Using these ideas, an extensive literature has developed that analyzes games in which players are uncertain about some aspect of the game. (Consult Leyton-Brown & Shoham (2008: ch. 7) for a concise summary of the current state-of-affairs and pointers to the relevant literature.) One can naturally wonder about the precise relationship between this literature and the literature we survey in this entry on the epistemic foundations of game theory. Indeed, the foundational literature we discuss here largely focuses on Harsanyi's approach to modeling higher-order beliefs (which we discuss in Section 2.3).

There are two crucial differences between the literature on Bayesian games and the literature we discuss in this entry (cf. the discussion in Brandenburger 2010: sec. 4 and 5).

1. In a Bayesian game, players are uncertain about the payoffs of the game, what other players believe are the correct payoffs, what other players believe that the other players believe about the payoffs, and so on, and this is the only source of uncertainty. That is, the players' (higher-order) beliefs about the payoffs in a game completely determine the (higher-order) beliefs about the other aspects of the game. In particular, if a player comes to *know* the payoffs of the other players, then that player is certain (and correct) about the possible (rational) choices of the other players.[3]

2. It is assumed that all players choose optimally given their information. That is, all players choose a strategy that maximizes their expected utility given their beliefs about the game, beliefs about what other players believe about the game, and so on. This means, in particular, that players do not entertain the possibility that their opponents may choose "irrationally."

Note that these assumptions are not inherent in the formalism that Harsanyi used to represent the players' beliefs in a game of incomplete information. Rather, they are better described as conventions followed by Harsanyi and subsequent researchers studying Bayesian games.

1.5 Imperfect Information and Perfect Recall

In a game with *imperfect information* (see Ross 2010 for a discussion), the players may not be perfectly informed about the moves of their opponents or the outcome of chance moves by nature. Games with imperfect information can be pictured as follows:

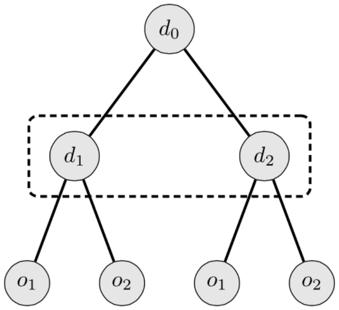

Figure 2

The interpretation is that the decision made at the first node (d_0) is forgotten, and so the decision maker is uncertain about whether she is at node d_1 or d_2. See Osborne (2003: ch. 9 & 10) for the general theory of games with imperfect information. In this section, we briefly discuss a foundational issue that arises in games with imperfect information.

Kuhn (1953) introduced the distinction between *perfect* and *imperfect* recall in games with imperfect information. Roughly, players have perfect recall provided they remember all of their own past moves (see Bonanno 2004; Kaneko & Kline 1995 for general discussions of the perfect recall assumption). It is standard in the game theory literature to assume that all players have perfect recall (i.e., they may be uncertain about previous choices of their opponents or nature, but they do remember their own moves).

As we noted in Section 1.3, there are different stages to the decision making process. Differences between these stages become even more pronounced in extensive games in which there is a temporal dimension to the game. There are two ways to think about the decision making process in an extensive game (with imperfect information). The first is to focus on the initial "planning stage". That is, initially, the players settle on a strategy specifying the (possibly random) move they will make at each of their choice nodes (this is the players' *global strategy*). Then, the players start making their respective moves (following the strategies which they have committed to without reconsidering their options at each choice node). Alternatively, we can assume that the players make "local judgements" at each of their choice nodes, always choosing the best option given the information that is currently available to them. A well-known theorem of Kuhn (1953) shows that if players have perfect recall, then a strategy is globally optimal if, and only if, it is locally optimal (see Brandenburger 2007 for a self-contained presentation of this classic result). That is, both ways of thinking about the decision making process in extensive games (with imperfect information) lead to the same recommendations/predictions.

The assumption of perfect recall is crucial for Kuhn's result. This is demonstrated by the well-known *absent-minded driver's problem* of Piccione and Rubinstein (1997a). Interestingly, their example is one where a decision maker may be tempted to change his strategy after the initial planning stage, *despite getting no new information*. They describe the example as follows:

> An individual is sitting late at night in a bar planning his midnight trip home. In order to get home he has to take the highway and get off at the second exit. Turning at the first exit leads into a disastrous area (payoff 0). Turning at the second exit yields the highest reward (payoff 4). If he continues beyond the second exit, he cannot go back and at the end of the highway he will find a motel where he can spend the night (payoff 1). The driver is absentminded and is aware of this fact. At an intersection, he cannot tell whether it is the first or the second intersection and he cannot remember how many he has passed (one can make the situation more realistic by referring to the 17th intersection). While sitting at the bar, all he can do is to decide whether or not to exit at an intersection. (Piccione & Rubinstein 1997a: 7)

The decision tree for the absent-minded driver is pictured below:

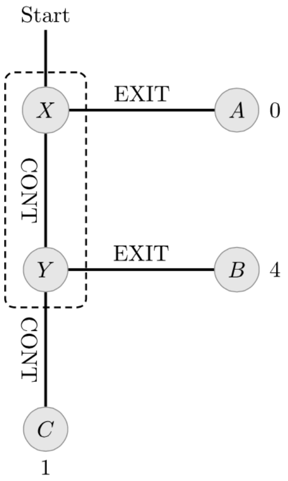

Figure 3

This problem is interesting since it demonstrates that there is a conflict between what the decision maker commits to do while planning at the bar and what he thinks is best at the first intersection:

> **Planning stage:** While planning his trip home at the bar, the decision maker is faced with a choice between “Continue; Continue” and “Exit”. Since he cannot distinguish between the two intersections, he cannot plan to “Exit” at the second intersection (he must plan the same behavior at both *X* and *Y*). Since “Exit” will lead to the worst outcome (with a payoff of 0), the optimal strategy is “Continue; Continue” with a guaranteed payoff of 1.
>
> **Action stage:** When arriving at an intersection, the decision maker is faced with a local choice of either “Exit” or “Continue” (possibly followed by another decision). Now the decision maker knows that since he committed to the plan of choosing “Continue” at each intersection, it is possible that he is at the second intersection. Indeed, the decision maker concludes that he is at the first intersection with probability 1/2. But then, his expected payoff for “Exit” is 2, which is greater than the payoff guaranteed by following the strategy he previously committed to. Thus, he chooses to “Exit”.

This problem has been discussed by a number of different researchers.[4] It is beyond the scope of this article to discuss the intricacies of the different analyses. An entire issue of *Games and Economic Behavior* (Volume 20, 1997) was devoted to the analysis of this problem. For a representative sampling of the approaches to this problem, see Kline (2002); Aumann, Hart, & Perry (1997); Board (2003); Halpern (1997); Piccione & Rubinstein (1997b).

1.6 Mixed Strategies

Mixed strategies play an important role in many game-theoretic analyses. Let $\Delta(X)$ denote the set of probability measures over the finite[5] set X. A **mixed strategy** for player i, is an element $m_i \in \Delta(S_i)$. If $m_i \in \Delta(S_i)$ assigns probability 1 to an element $s_i \in S_i$, then m_i is called a **pure strategy** (in such a case, I write s_i for m_i). Mixed strategies are incorporated into a game-theoretic analysis as follows. Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N}\rangle$ is a finite strategic game. The *mixed extension* of G is the strategic game in which the strategies for player i are the mixed strategies in G (i.e., $\Delta(S_i)$) and the utility for player i (denoted U_i) of the joint mixed strategy $m \in \Pi_{i \in N}\Delta(S_i)$ is calculated in the obvious way (let $m(s) = m_1(s_1) \cdot m_2(s_2) \cdots m_n(s_n)$ for $s \in \Pi_{i \in N} S_i$):

$$U_i(m) = \sum_{s \in \Pi_{i \in N} S_i} m(s) \cdot u_i(s)$$

Thus, the solution space of a mixed extension of the game G is the set $\Pi_{i \in N}\Delta(S_i)$.

Despite their prominence in game theory, the interpretation of mixed strategies is controversial, as Ariel Rubinstein notes:

> We are reluctant to believe that our decisions are made at random. We prefer to be able to point to a reason for each action we take. Outside of Las Vegas we do not spin roulettes. (Rubinstein 1991: 913).

In epistemic game theory, it is more natural to work with an alternative interpretation of mixed strategies: A mixed strategy for player i is a representation of the *beliefs* of i's opponent(s) about what she will do. This is nicely explained in Robert Aumann's influential paper (Aumann 1987—see, especially, Section 6 of this paper for a discussion of this interpretation of mixed strategies):

> An important feature of our approach is that it does not require explicit randomization on the part of the players. Each player always chooses a definite pure strategy, with no attempt to randomize; the probabilistic nature of the strategies reflects the uncertainties of other players about his choice. (Aumann 1987: 3)

2. Game Models

A *model* of a game is a structure that represents the informational context of a given play of the game. The states, or possible worlds, in a game model describe a possible play of the game *and* the specific information that influenced the players' choices (which may be different at each state). This includes each player's "knowledge" of her own choice and opinions about the choices and "beliefs" of her opponents. A key challenge when constructing a model of a game is how to represent the different informational attitudes of the players. Researchers interested in the foundation of decision theory, epistemic and doxastic logic and, more recently, *formal epistemology* have developed many different formal models that can describe the many varieties of informational attitudes important for assessing the choice of a *rational* agent in a decision- or game-theoretic situation.

After discussing some general issues that arise when describing the informational context of a game, we introduce the two main types of models that have been used to describe the players' beliefs (and other informational attitudes) in a game situation: *type spaces* (Harsanyi 1967–68; Siniscalchi 2008) and the so-called *Aumann*- or *Kripke-structures* (Aumann 1999a; Fagin, Halpern, Moses, & Vardi 1995). Although these two approaches have much in common, there are some important differences which we highlight below. A second, more fundamental, distinction found in the literature is between "quantitative" structures, representing "graded" attitudes (typically via probability distributions), and "qualitative" structures representing "all out" attitudes. Kripke structures are often associated with the former, and type spaces with the latter, but this is not a strict classification.

2.1 General Issues

2.1.1 Varieties of informational attitudes

Informational contexts of games can include various forms of attitudes, from the classical knowledge and belief to robust (Stalnaker 1994) and strong (Battigalli & Siniscalchi 2002) belief, each echoing in different notions of game-theoretical rationality. It is beyond the scope of this article to survey the details of this vast literature (cf. the next section for some discussion of this literature). Rather, we will introduce a general distinction between *hard* and *soft* attitudes, distinction mainly developed in dynamic epistemic logic (van Benthem 2011), which proves useful in understanding the various philosophical issues raised by epistemic game theory.

We call *hard information*, information that is *veridical*, *fully introspective* and *not revisable*. This notion is intended to capture what the agents are fully and correctly certain of in a given interactive situation. At the *ex interim* stage, for instance, the players have hard information about their *own* choice. They "know" which strategy they have chosen, they know that they know this, and no new incoming information could make them change their opinion on this. As this phrasing suggests, the term *knowledge* is often used, in absence of better terminology, to describe this very strong type of informational attitude. Epistemic logicians and game theorist are well aware of the possible discrepancies between such hard "knowledge" and our intuitive or even philosophical understanding of this notion. In the present context is it sufficient to observe that hard information shares *some* of the characteristics that have been attributed to knowledge in the epistemological literature, for instance truthfulness. Furthermore, hard information might come closer to what has been called "implicit knowledge" (see Section 5.3 below). In any case, it seems philosophically more constructive to keep an eye on where the purported counter-intuitive properties of hard information come into play in epistemic game theory, rather than reject this notion as wrong or flawed at the upshot.

Soft information is, roughly speaking, anything that is not "hard": it is not necessarily veridical, not necessarily fully introspective and/or highly revisable in the presence of new information. As such, it comes much closer to *beliefs*. Once again, philosophical carefulness is in order here. The whole range of informational attitudes that is labeled as "beliefs" indeed falls into the category of attitudes that can be described as "regarding something as true" (Schwitzgebel 2010), among which beliefs, in the philosophical sense, seem to form a proper sub-category.

2.1.2 Possible worlds models

The models introduced below describe the players' hard and soft information in interactive situations. They differ in their representation of a state of the world, but they can all be broadly described as "possible worlds models" familiar in much of the philosophical logic literature. The starting point is a non-empty (finite or infinite) set S of *states of nature* describing the *exogenous* parameters (i.e., facts about the physical world) that do not depend on the agents' uncertainties. Unless otherwise specified, S is the set of possible outcomes of the games, the set of all *strategy profiles*.[6] Each player is assumed to entertain a number of *possibilities*, called *possible worlds* or simply *(epistemic) states*. These "possibilities" are intended to represent a possible way a game situation may evolve. So each possibility will be associated with a *unique* state of nature (i.e., there is a function from possible worlds to states of nature, but this function need not be 1–1 or even onto). It is crucial for the analysis of rationality in games that there may be *different* possible worlds associated with the same state of nature. Such possible worlds are important because they open the door to representing different state of information. Such state-based modeling naturally yields a *propositional* view of the agents' informational attitudes. Agents will have beliefs/knowledge about *propositions*, which are also called *events* in the game-theory literature, and are represented as sets of possible worlds. These basic modeling choices are not uncontroversial, but such issues are not our concern in this entry.

2.2 Relational Models

We start with the models that are familiar to philosophical logicians (van Benthem 2010) and computer scientists (Fagin et al. 1995). These models were introduced to game theory by Robert Aumann (1976) in his seminal paper *Agreeing to Disagree* (see Vanderschraaf & Sillari 2009, Section 2.3, for a discussion of this result). First, some terminology: Given a set W of states, or possible worlds, let us call any subset $E \subseteq W$ an *event* or *proposition*. Given events $E \subseteq W$ and $F \subseteq W$, we use standard set-theoretic notation for intersection ($E \cap F$, read "E and F"), union ($E \cup F$, read "E or F") and (relative) complement ($-E$, read "not E"). We say that an event $E \subseteq W$ occurs at state w if $w \in E$. This terminology will be crucial for studying the following models.

Definition 2.1 (Epistemic Model) Suppose that G is a strategic game, S is the set of strategy profiles of G, and N is the set of players. An **epistemic model based on S and N** is a triple $\langle W, \{\Pi_i\}_{i\in N}, \sigma\rangle$, where W is a nonempty set, for each $i \in N$, Π_i is a partition[7] over W and $\sigma : W \to S$.

Epistemic models represent the informational context of a given game in terms of possible configurations of states of the game and the hard information that the agents have about them. The function σ assigns to each possible world a unique state of the game in which every ground fact is either true or false. If $\sigma(w) = \sigma(w')$ then the two worlds w, w' will agree on all the ground facts (i.e., what actions the players will choose) but, crucially, the agents may have different information in them. So, elements of W are *richer*, than the elements of S (more on this below).

Given a state $w \in W$, the cell $\Pi_i(w)$ is called agent i's *information set*. Following standard terminology, if $\Pi_i(w) \subseteq E$, we say the agent i *knows* the event E at state w. Given an event E, the event that agent i knows E is denoted $K_i(E)$. Formally, we define for each agent i a knowledge function assigning to every event E the event that the agent i knows E:

Definition 2.2 (Knowledge Function) Let $\mathcal{M} = \langle W, \{\Pi_i\}_{i\in N}, \sigma\rangle$ be an epistemic model. The **knowledge function** for agent i based on $\mathcal{M}$ is $K_i : \wp(W) \to \wp(W)$ with:

$$K_i(E) = \{w \mid \Pi_i(w) \subseteq E\}$$

where for any set X, $\wp(X)$ is the *powerset of* X.

Remark 2.3 It is often convenient to work with *equivalence relations* rather than partitions. In this case, an epistemic model based on S and N can also be defined as a triple $\langle W, \{\sim_i\}_{i\in N}, \sigma\rangle$ where W and σ are as above and for each $i \in N$, $\sim_i \subseteq W \times W$ is reflexive, transitive and symmetric. Given such a model $\langle W, \{\sim_i\}_{i\in N}, \sigma\rangle$, we write

$$[w]_i = \{v \in W \mid w \sim_i v\}$$

for the equivalence class of w. Since there is a 1–1 correspondence between equivalence relations and partitions,[8] we will abuse notation

> and use $\sim_i$ and Π_i interchangeably.
>
> Applying the above remark, an alternative definition of $K_i(E)$ is that E is true in all the states the agent i considers possible (according to i's hard information). That is, $K_i(E) = \{w \mid [w]_i \subseteq E\}$.

Partitions or equivalence relations are intended to represent the agents' *hard information* at each state. It is well-known that the knowledge operator satisfies the properties of the epistemic logic **S5** (see Hendricks & Symons 2009 for a discussion). We do not discuss this and related issues here and instead focus on how these models can be used to provide the informational context of a game.

> **An Example.** Consider the following coordination game between Ann (player 1) and Bob (player 2). As is well-known, there are two pure-strategy Nash equilibrium ((u,l) and (d,r)).

		Bob	
		l	r
Ann	u	3,3	0,0
	d	0,0	1,1

Figure 4: A strategic coordination game between Ann and Bob

The utilities of the players are not important for us at this stage. To construct an epistemic model for this game, we need first to specify what are the states of nature we will consider. For simplicity, take them to be the set of strategy profiles $S = \{(u,l),(d,l),(u,r),(d,l)\}$. The set of agents is of course $N = \{A,B\}$. What will be the set of states W? We start by assuming $W = S$, so there is exactly one possible world corresponding to each state of nature. This needs not be so, but here this will help to illustrate our point.

There are many different partitions for Ann and Bob that we can use to complete the description of this simple epistemic model. Not all of the partitions are appropriate for analyzing the *ex interim* stage of the decision-making process, though. For example, suppose $\Pi_A = \Pi_B = \{W\}$ and consider the event $U = \{(u,l),(u,r)\}$ representing the situation where Ann chooses u. Notice that $K_A(U) = \emptyset$ since for all $w \in W, \Pi_A(w) \not\subseteq U$, so there is no state where Ann *knows* that she chooses u. This means that this model is appropriate for reasoning about the *ex ante* stage rather than the *ex interim* stage. This is easily fixed with an additional technical assumption: Suppose S is a set of strategy profiles for some (strategic or extensive) game with players $N = \{1,\dots,n\}$.

> A model $\mathcal{M} = \langle W, \{\Pi_i\}_{i\in N}, \sigma\rangle$ is said to be an ***ex interim* epistemic model** if for all $i \in N$ and $w, v \in W$, if $v \in \Pi_i(w)$ then $\sigma_i(w) = \sigma_i(v)$

where $\sigma_i(w)$ is the i^{th} component of the strategy profile $s \in S$ assigned to w by σ. An example of an *ex interim* epistemic model with states W is:

- $\Pi_A = \{\{(u,l),(u,r)\},\{(d,l),(d,r)\}\}$ and
- $\Pi_B = \{\{(u,l),(d,l)\},\{(u,r),(d,r)\}\}$.

Note that this simply reinterprets the game matrix in Figure 1 as an epistemic model where the rows are Ann's information sets and the columns are Bob's information sets. Unless otherwise stated, we will always assume that our epistemic models are *ex interim*. The class of *ex interim* epistemic models is very rich with models describing the (hard) information the agents have about their own choices, the (possible) choices of the other players *and* higher-order (hard) information (e.g., "Ann knows that Bob knows that…") about these decisions.

We now look at the epistemic model described above in more detail. We will often use the following diagrammatic representation of the model to ease exposition. States are represented by nodes in a graph where there is a (undirected) edge between states w_i and w_j when w_i and w_j are in the same partition cell. We use a solid line labeled with A for Ann's partition and a dashed line labeled with B for Bob's partition (reflexive edges are not represented for simplicity). The event $U = \{w_1, w_3\}$ representing the proposition "Ann decided to choose option u" is the shaded gray region:

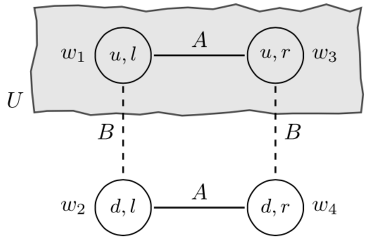

Figure 5

Notice that the following events are true at all states:

1. $-K_B(U)$: "Bob does not know that Ann decided to choose u"
2. $K_B(K_A(U) \vee K_A(-U))$: "Bob knows that Ann knows whether she has decided to choose u"
3. $K_A(-K_B(U))$: "Ann knows that Bob does not know that she has decided to choose u"

In particular, these events are true at state w_1 where Ann has decided to choose u (i.e., $w_1 \in U$). The first event makes sense given the assumptions about the available information at the *ex interim* stage: each player knows their own choice but not the other players' choices. The second event is a concrete example of another assumption about the available information: Bob has the information that Ann has, in fact, made *some* choice. But what warrants Ann to conclude that Bob does not know she has chosen u (the third event)? This is a much more significant statement about what Ann knows about what Bob expects her to do. Indeed, in certain contexts, Ann may have very good reasons to think it is possible that Bob actually *knows* she will choose u. We can find an *ex interim* epistemic model where this event $(-K_A(-K_B(U)))$ is true at w_1, but this requires adding a new possible world:

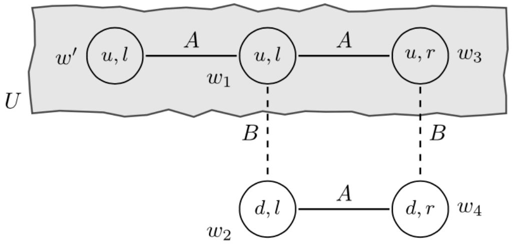

Figure 6

Notice that since $\Pi_B(w') = \{\{w'\}\} \subseteq U$ we have $w' \in K_B(U)$. That is, Bob knows that Ann chooses u at state w'. Finally, a simple calculation shows that $w_1 \in -K_A(-K_B(U))$, as desired. Of course, we can question the other *substantive assumptions* built-in to this model (e.g., at w_1, Bob knows that Ann does not know he will choose L) and continue modifying the model. This raises a number of interesting conceptual and technical issues which we discuss in Section 7.

2.2.1 Adding Beliefs

So far we have looked at relational models of hard information. A small modification of these models allows us to model a softer informational attitude. Indeed, by simply replacing the assumption of reflexivity of the relation $\sim_i$ with seriality (for each state w there is a state v such that $w \sim_i v$), but keeping the other aspects of the model the same, we can capture what epistemic logicians have called "*beliefs*". Formally, a **doxastic model** is a tuple $\langle W, \{R_i\}_{i \in N}, V \rangle$ where W is a nonempty set of states, R_i is a transitive, Euclidean and serial relation on W and V is a valuation function (cf. Definition 2.1). This notion of belief is very close to the above hard informational attitude and, in fact, shares all the properties of K_i listed above except *Veracity* (this is replaced with a weaker assumption that agents are "consistent" and so cannot believe contradictions). This points to a logical analysis of both informational attitudes with various "bridge principles" relating knowledge and belief (such as knowing something implies believing it or if an agent believes ϕ then the agent knows that he believes it). However, we do not discuss this line of research here since these models are not our preferred ways of representing the agents' soft information (see, for example, Halpern 1991 and Stalnaker 2006).

Plausibility Orderings

A key aspect of beliefs which is not yet represented in the above models is that they are *revisable* in the presence of new information. While there is an extensive literature on the theory of belief revision in the "AGM" style (Alchourrón, Gärdenfors, & Makinson 1985), we focus on how to extend an epistemic model with a representation of softer, revisable informational attitudes. The standard approach is to include a *plausibility ordering* for each agent: a preorder (reflexive and transitive) denoted $\preceq_i \subseteq W \times W$. If $w \preceq_i v$ we say "player i considers w at least as plausible as v." For an event $X \subseteq W$, let

$$Min_{\preceq_i}(X) = \{v \in W \mid v \preceq_i w \text{ for all } w \in X\}$$

denote the set of minimal elements of X according to $\preceq_i$. Thus while the $\sim_i$ partitions the set of possible worlds according to the agents' hard information, the plausibility ordering $\preceq_i$ represents which of the possible worlds the agent considers more likely (i.e., it represents the players soft information).

Definition 2.4 (Epistemic-Plausibility Models) Suppose that G is a strategic game, S is the set of strategy profiles of G, and N is the set of players. An **epistemic-plausibility model** is a tuple $\langle W, \{\Pi_i\}_{i\in N}, \{\preceq_i\}_{i\in N}, \sigma\rangle$ where $\langle W, \{\Pi_i\}_{i\in N}, \sigma\rangle$ is an epistemic model, $\sigma : W \to S$ and for each $i \in N$, $\preceq_i$ is a well-founded,[9] reflexive and transitive relation on W satisfying the following properties, for all $w, v \in W$

1. *plausibility implies possibility:* if $w \preceq_i v$ then $v \in \Pi_i(w)$.
2. *locally-connected:* if $v \in \Pi_i(w)$ then either $w \preceq_i v$ or $v \preceq_i w$.

Remark 2.5 Note that if $v \notin \Pi_i(w)$ then $w \notin \Pi_i(v)$. Hence, by property 1, $w \not\preceq_i v$ and $v \not\preceq_i w$. Thus, we have the following equivalence: $v \in \Pi_i(w)$ iff $w \preceq_i v$ or $v \preceq_i w$.

Local connectedness implies that $\preceq_i$ totally orders $\Pi_i(w)$ and well-foundedness implies that $Min_{\preceq_i}(\Pi_i(w))$ is nonempty. This richer model allows us to formally define a variety of (soft) informational attitudes. We first need some additional notation: the plausibility relation $\preceq_i$ can be lifted to subsets of W as follows[10]

$$X \preceq_i Y \text{ iff } x \preceq_i y \text{ for all } x \in X \text{ and } y \in Y$$

Suppose $\mathcal{M} = \langle W, \{\Pi_i\}_{i\in N}, \{\preceq_i\}_{i\in N}, \sigma\rangle$ is an epistemic-plausibility model, consider the following operators (formally, each is a function from $\wp(W)$ to $\wp(W)$ similar to the knowledge operator defined above):

- *Belief:* $B_i(E) = \{w \mid Min_{\preceq_i}(\Pi_i(w)) \subseteq E\}$
  This is the usual notion of belief which satisfies the standard properties discussed above (e.g., consistency, positive and negative introspection).

- *Robust Belief:* $B_i^r(E) = \{w \mid v \in E, \text{ for all } v \text{ with } w \preceq_i v\}$
  So, E is robustly believed if it is true in all worlds more plausible then the current world. This stronger notion of belief has also been called *certainty* by some authors (cf. Shoham & Leyton-Brown 2008: sec. 13.7).

- *Strong Belief:*

$$B_i^s(E) = \{w \mid E \cap \Pi_i(w) \neq \emptyset \text{ and } E \cap \Pi_i(w) \preceq_i -E \cap \Pi_i(w)\}$$

  So, E is strongly believed provided it is epistemically possible and agent i considers *any* state in E more plausible than *any* state in the complement of E.

It is not hard to see that if agent i knows that E then i (robustly, strongly) believes that E. However, much more can be said about the logical relationship between these different notions. (The logic of these notions has been extensively studied by Alexandru Baltag and Sonja Smets in a series of articles, see Baltag & Smets 2009 in Other Internet Resources for references.)

As noted above, a crucial feature of these informational attitudes is that they may be defeated by appropriate evidence. In fact, we can characterize these attitudes in terms of the type of evidence which can prompt the agent to adjust her beliefs. To make this precise, we introduce the notion of a *conditional belief:* suppose $\mathcal{M} = \langle W, \{\Pi_i\}_{i\in N}, \{\preceq_i\}_{i\in N}, \sigma\rangle$ is an epistemic-plausibility model and E and F are events, then the **conditional** belief operator is defined as follows:

$$B_i^F(E) = \{w \mid Min_{\preceq_i}(F \cap \Pi_i(w)) \subseteq E\}$$

So, 'B_i^F' encodes what agent i will believe upon receiving (possibly misleading) evidence that F is *true*.

We conclude this section with an example to illustrate the above concepts. Recall again the coordination game of Figure 4: there are two actions for player 1 (Ann), u and d, and two actions for player 2 (Bob), r and l. Again, the preferences (or utilities) of the players are not important at this stage since we are only interested in describing the players' information. The following epistemic-plausibility model is a possible description of the players' informational attitudes that can be associated with this game. The solid lines represent player 1's informational attitudes and the dashed line represents player 2's. The arrows correspond to the players plausibility orderings with an i-arrow from w to v meaning $v \preceq_i w$ (we do not draw all the arrows: each plausibility ordering can be completed by filling in arrows that result from reflexivity and transitivity). The different regions represent the players' hard information.

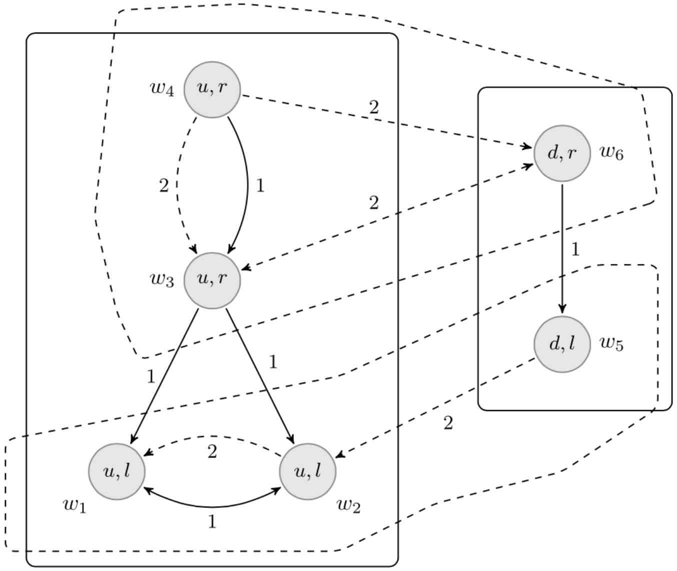

Figure 7

Suppose that the actual state of play is w_4. So, player 1 (Ann) chooses u and player 2 (Bob) chooses r. Further, suppose that $L = \{w_1, w_2, w_5\}$ is the event where where player 2 chooses l (similarly for U, D, and R)

1. $B_1(L)$: "player 1 believes that player 2 is choosing L"
2. $B_1(B_2(U))$: "player 1 believes that player 2 believes that player 1 chooses u"
3. $B_1^R(-B_2(U))$: "given that player 2 chooses r, player 1 believes that player 2 does not believe she is choosing u"

This last formula is interesting because it "pre-encodes" what player 1 would believe upon learning that player 2 is choosing R. Note that upon receiving this *true* information, player 1 drops her belief that player 2 believes she is choosing u. The situation can be even more interesting if there are statements in the language that reveal only *partial* information about the player strategy choices. Suppose that E is the event $\{w_4, w_6\}$. Now E is true at w_4 and player 2 believes that *player 1 chooses d* given that E is true (i.e., $w_4 \in B_2^E(D)$). So, player 1 can "bluff" by revealing the true (though partial) information E.

Probabilities

The above models use a "crisp" notion of uncertainty, i.e., for each agent and state w, any other state $v \in W$ either is or is not possible at/more plausible than w. However, there is an extensive body of literature focused on *graded*, or *quantitative*, models of uncertainty (Huber 2009; Halpern 2003). For instance, in the Game Theory literature it is standard to represent the players' *beliefs* by probabilities (Aumann 1999b; Harsanyi 1967–68). The idea is simple: replace the plausibility orderings with probability distributions:

Definition 2.6 (Epistemic-Probability Model) Suppose that G is a strategic game, S is the set of strategy profiles of G, and N is the set of players. An **epistemic-probabilistic model** is a tuple

$$\mathcal{M} = \langle W, \{\sim_i\}_{i\in N}, \{P_i\}_{i\in N}, \sigma\rangle$$

where $\langle W, \{\Pi_i\}_{i\in N}, \sigma\rangle$ is an epistemic model and

$$P_i : W \to \Delta(W)$$
$$\Delta(W) = \{p : W \to [0,1] \mid p \text{ is a probability measure}\}$$

assigns to each state a probability measure over W. Write p_i^w for the i's probability measure at state w. We make two natural assumptions (cf. Definition 2.4):

1. For all $v \in W$, if $p_i^w(v) > 0$ then $p_i^w = p_i^v$; and
2. For all $v \notin \Pi_i(w)$, $p_i^w(v) = 0$.

Property 1 says that if i assigns a non-zero probability to state v at state w then the agent uses the same probability measure at both states. This means that the players "know" their own probability measures. The second property implies that players must assign a probability of zero to all states outside the current (hard) information cell. These models provide a very precise description of the players' hard and soft informational attitudes. However, note that writing down a model requires us to specify a different probability measure for each partition cell which can be quite cumbersome. Fortunately, the properties in the above definition imply that, for each agent, we can view the agent's probability measures as arising from one probability measure through conditionalization. Formally, for each $i \in N$, agent i's **(subjective) prior probability** is any element of $p_i \in \Delta(W)$. Then, in order to define an epistemic-probability model we need only give for each agent $i \in N$, (1) a prior probability $p_i \in \Delta(W)$

and (2) a partition Π_i on W such that for each $w \in W, p_i(\Pi_i(w)) > 0$. The probability measures for each $i \in N$ are then defined by:

$$P_i(w) = p_i(\cdot \mid \Pi_i(w)) = \frac{p_i(\cdot \cap \Pi_i(w))}{p_i(\Pi_i(w))}$$

Of course, the side condition that for each $w \in W$, $p_i(\Pi_i(w)) > 0$ is important since we cannot divide by zero—this will be discussed in more detail in later sections. Indeed, (assuming W is finite[11]) given any epistemic-plausibility model we can find, for each agent, a prior (possibly different ones for different agents) that generates the model as described above. This is not only a technical observation: it means that we are assuming that the players' beliefs about the outcome of the situation are fixed *ex ante* with the *ex interim* beliefs being derived through conditionalization on the agent's *hard information*. (See Morris 1995 for an extensive discussion of the situation when there is a *common* prior.) We will return to these key assumptions throughout the text.

As above we can define belief operators, this time specifying the precise degree to which an agent believes an event:

- *Probabilistic belief:* $B_i^r(E) = \{w \mid p_i^w(E) = r\}$
  Here, r can be any real number in the unit interval; however, it is often enough to restrict attention to the rational numbers in the unit interval.

- *Full belief:* $B_i(E) = B_i^1(E) = \{w \mid p_i^w(E) = 1\}$
  So, full belief is defined to belief with probability one. This is a standard assumption in this literature despite a number of well-known conceptual difficulties (see Huber 2009 for an extensive discussion of this and related issues). It is sometimes useful to work with the following alternative characterization of full-belief (giving it a more "modal" flavor): Agent i believes E at state w provided all the states that i assigns positive probability to at w are in E. Formally,

  $$B_i(E) = \{w \mid \text{for all } v, \text{ if } p_i^w(v) > 0 \text{ then } v \in E\}$$

These models have also been subjected to sophisticated logical analyses (Fagin, Halpern, & Megiddo 1990; Heifetz & Mongin 2001) complementing the logical frameworks discussed above (cf. Baltag & Smets 2006).

We conclude this section with an example of an epistemic-probability model. Recall again the coordination game of Figure 4: there are two actions for player 1 (Ann), u and d, and two actions for player 2 (Bob), r and l. The preferences (or utilities) of the players are not important at this stage since we are only interested in describing the players' information.

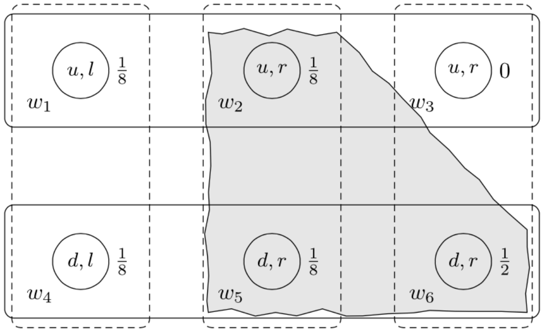

Figure 8

The solid lines are Ann's information partition and the dashed lines are Bob's information partition. We further assume there is a common prior p_0 with the probabilities assigned to each state written to the right of the state. Let $E = \{w_2, w_5, w_6\}$ be an event. Then, we have

- $B_1^{\frac{1}{2}}(E) = \{w \mid p_0(E \mid \Pi_1(w)) = \frac{p_0(E\cap\Pi_1(w))}{p_0(\Pi_1(w))} = \frac{1}{2}\} = \{w_1, w_2, w_3\}$: "Ann assigns probability 1/2 to the event E given her information cell $\Pi_1(w_1)$.

- $B_2(E) = B_2^1(E) = \{w_2, w_5, w_3, w_6\}$. In particular, note that at w_6, the agent believes (with probability 1) that E is true, but does not *know* that E is true as $\Pi_2(w_6) \nsubseteq E$. So, there is a distinction between states the agent considers possible (given their "hard information") and states to which players assign a non-zero probability.

- Let $U = \{w_1, w_2, w_3\}$ be the event that Ann plays u and $L = \{w_1, w_4\}$ the event that Bob plays l. Then, we have

    - $K_1(U) = U$ and $K_2(L) = L$: Both Ann and Bob know that strategy they have chosen;

    - $B_1^{\frac{1}{2}}(L) = U$: At all states where Ann plays u, Ann believes that Bob plays L with probability 1/2; and

    - $B_1(B_2^{\frac{1}{2}}(U)) = \{w_1, w_2, w_3\} = U$: At all states where Ann plays u, she believes that Bob believes with probability 1/2 that she is playing u.

2.3 Harsanyi Type Spaces

An alternative approach to modeling beliefs was initiated by Harsanyi in his seminal paper (Harsanyi 1967–68). Rather than "possible worlds", Harsanyi takes the notion of the players' *type* as primitive. Formally, the players are assigned a nonempty set of types. Typically, players are assumed to *know* their own type but not the types of the other players. As we will see, each type can be associated with a specific hierarchy of belief

Definition 2.7 (Qualitative Type Space) A **Qualitative type space** for a (nonempty) set of states of nature S and agents N is a tuple $\langle \{T_i\}_{i\in N}, \{\lambda_i\}_{i\in N}, S\rangle$ where for each $i \in N$, T_i is a nonempty set and

$$\lambda_i : T_i \to \wp(\times_{j\neq i} T_j \times S).$$

So, each type $t \in T_i$ is associated with a set of tuples consisting of types of the other players and a state of nature. For simplicity, suppose there are only two players, Ann and Bob. Intuitively, $(t', o') \in \lambda_{Ann}(t)$ means that Ann's type t considers it possible that the outcome is o' *and* Bob is of type t'. Since the players' uncertainty is directed at the choices and types of the *other* players, the informational attitude captured by these models will certainly not satisfy the Truth axiom. In fact, qualitative type spaces can be viewed as simply a "re-packaging" of the relational models discussed above (cf. Zvesper 2010 for a discussion).

Consider again the running example of the coordination game between Ann and Bob (pictured in Figure 1). In this case, the set of states of nature is $S = \{(u,l), (d,l), (u,r), (d,r)\}$. In this context, it is natural to modify the definition of the type functions λ_i to account for the fact that the players are only uncertain about the other players' choices: let $S_A = \{u, d\}$ and $S_B = \{l, r\}$ and suppose T_A and T_B are nonempty sets of types. Define λ_A and λ_B as follows:

$$\lambda_A : T_A \to \wp(T_B \times S_B) \qquad \lambda_B : T_B \to \wp(T_A \times S_A)$$

Suppose that there are two types for each player: $T_A = \{t_1^A, t_2^A\}$ and $T_B = \{t_1^B, t_2^B\}$. A convenient way to describe the maps λ_A and λ_B is:

$$\begin{array}{cc|c|c|} & & l & r \\ \cline{3-4} \lambda_A(T_1^A) & t_1^B & 1 & 0 \\ \cline{3-4} & t_2^B & 1 & 0 \\ \cline{3-4} \end{array} \qquad \begin{array}{cc|c|c|} & & l & r \\ \cline{3-4} \lambda_A(T_2^A) & t_1^B & 0 & 0 \\ \cline{3-4} & t_2^B & 1 & 0 \\ \cline{3-4} \end{array}$$

$$\lambda_B(T_1^B)\ \begin{array}{c|c|c|} & u & d \\ \hline t_1^A & 1 & 0 \\ \hline t_2^A & 0 & 0 \\ \hline \end{array} \qquad \lambda_B(T_2^B)\ \begin{array}{c|c|c|} & u & d \\ \hline t_1^A & 0 & 0 \\ \hline t_2^A & 0 & 1 \\ \hline \end{array}$$

Figure 9

where a 1 in the (t', s) entry of the above matrices corresponds to assuming $(t', s) \in \lambda_i(t)$ $(i = A, B)$. What does it mean for Ann (Bob) to *believe* an event E in a type structure? We start with some intuitive observations about the above type structure:

- Regardless of what type we assign to Ann, she believes that Bob will choose l since in both matrices, $\lambda_A(t_1^A)$ and $\lambda_A(t_2^A)$, the only places where a 1 appears is under the l column. So, fixing a type for Ann, in all of the situations Ann considers possible it is true that Bob chooses l.

- If Ann is assigned the type t_1^A, then she considers it possible that Bob believes she will choose u. Notice that type t_1^A has a 1 in the row labeled t_1^B, so she considers it possible that Bob is of type t_1^B, and type t_1^B believes that Ann chooses u (the only places where 1 appears is under the u column).

- If Ann is assigned the type t_2^A, then Ann believes that Bob believes that Ann believes that Bob will choose l. Note that type t_2^A "believes" that Bob will choose l and furthermore t_2^A believes that Bob is of type t_2^B who in turn believes that Ann is of type t_2^A.

We can formalize the above informal observations using the following notions: Fix a qualitative type space $\langle \{T_i\}_{i\in N}, \{\lambda_i\}_{i\in N}, S\rangle$ for a (nonempty) set of states of nature S and agents N.

- A **(global) state**, or **possible world** is a tuple $(t_1, t_2, \ldots, t_n, s)$ where $t_i \in T_i$ for each $i = 1, \ldots, n$ and $s \in S$. If $S = \times S_i$ is the set of strategy profiles for some game, then we write a possible world as: $(t_1, s_1, t_2, s_2, \ldots, t_n, s_n)$ where $s_i \in S_i$ for each $i = 1, \ldots, n$.

- Type spaces describe the players beliefs about the other players' choices, so the notion of an *event* needs to be relativized to an agent. An **event for agent** i is a subset of $\times_{j\neq i} T_j \times S$. Again if S is a set of strategy profiles (so $S = \times S_i$), then an event for agent i is a subset of $\times_{j\neq i}(T_j \times S_j)$.

- Suppose that E is an event for agent i, then we say that agent i **believes** E **at** $(t_1, t_2, \ldots, t_n, s)$ provided $\lambda(t_1, s) \subseteq E$.

In the specific example above, an event for Ann is a set $E \subseteq T_B \times S_B$ and we can define the set of pairs (t^A, s^A) that believe this event:

$$B_A(E) = \{(t^A, s^A) \mid \lambda_A(t^A, s^A) \subseteq E\}$$

similarly for Bob. Note that the event $B_A(E)$ is an event for Bob and *vice versa*. A small change to the above definition of a type space (Definition 2.7) allows us to represent *probabilistic* beliefs (we give the full definition here for future reference):

> **Definition 2.8 (Type Space)** A **type space** for a (nonempty) set of states of nature S and agents N is a tuple $\langle \{T_i\}_{i\in N}, \{\lambda_i\}_{i\in N}, S\rangle$ where for each $i \in N$, T_i is a nonempty set and
>
> $$\lambda_i : T_i \to \Delta(\times_{j\neq i} T_j \times S).$$
>
> where $\Delta(\times_{j\neq i} T_j \times S)$ is the set of probability measures on $\times_{j\neq i} T_j \times S$.

Types and their associated image under λ_i encode the players' (probabilistic) information about the others' information. Indeed, each type is associated with a hierarchy of belief. More formally, recall that an event E for a type t_i is a set of pairs (σ_{-j}, t_{-j}), i.e., a set of strategy choices and types for all the other players. Given an event E for player i, let $\lambda_i(t_i)(E)$ denote the sum of the probabilities that $\lambda_i(t_i)$ assigns to the elements of E. The type t_i of player i is said to *(all-out) believe* the event E whenever $\lambda_i(t_i)(E) = 1$. Conditional beliefs are computed in the standard way: type t_i believes that E given F whenever:

$$\frac{\lambda_i(t_i)(E \cap F)}{\lambda_i(t_i)(F)} = 1$$

A *state* in a type structure is a tuple (σ, t) where σ is a strategy profile and t is "type profile", a tuple of types, one for each player. Let $B_i(E) = \{(\sigma_{-j}, t_{-j}) : t_i \text{ believes that } E\}$ be the event (for j) that i believes that E. Then agent j believes that i believes that E when $\lambda_j(t_j)(B_i(E)) = 1$. We can continue in this manner computing any (finite) level of such higher-order information.

Example

Returning again to our running example game where player 1 (Ann) has two available actions $\{u, d\}$ and player 2 (Bob) has two available actions $\{l, r\}$. The following type space describes the players' information: there is one type for Ann (t_1) and two for Bob (t_2, t_2') with the corresponding probability measures given below:

$\lambda_1(t_1)$		l	r
	t_2	0.5	0
	t_1'	0.4	0.1

Figure 10: Ann's beliefs about Bob

u d u d

$\lambda_2(t_2)$		
t_1	1	0

$\lambda_2(t_2')$		
t_1	0.75	0.25

Figure 11: Bob's belief about Ann

In this example, since there is only one type for Ann, both of Bob's types are *certain* about Ann's beliefs. If Bob is of type t_2 then he is certain Ann is choosing u while if he is of type t_2' he thinks there is a 75% chance she plays u. Ann assigns equal probability (0.5) to Bob's types; and so, she believes it is equally likely that Bob is certain she plays u as Bob thinking there is a 75% chance she plays u. The above type space is a very compact description of the players' informational attitudes. An epistemic-probabilistic model can describe the same situation (here p_i for $i = 1, 2$ is player i's prior probability):

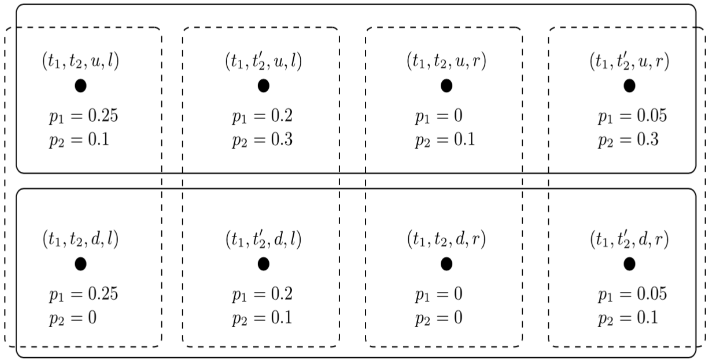

Figure 12

Some simple (but instructive!) calculations can convince us that these two models represent the same situation. The more interesting question is how do these probabilistic models relate to the epistemic-doxastic models of Definition 2.4. Here the situation is more complex. On the one hand, probabilistic models with a graded notion of belief which is much more fine-grained than the "all-out" notion of belief discussed in the context of epistemic-doxastic models. On the other hand, in an epistemic-doxastic model, conditional believes are defined for *all* events. In the above models, they are only defined for events that are assigned nonzero probabilities. In other words, epistemic-probabilistic models do not describe what a player may believe upon learning something "surprising" (i.e., something currently assigned probability zero).

A number of extensions to basic probability theory have been discussed in the literature that address precisely this problem. We do not go into details here about these approaches (a nice summary and detailed comparison between different approaches can be found in Halpern (2010) and instead sketch the main ideas. The first approach is to use so-called *Popper functions* which take *conditional probability measures* as primitive. That is, for each non-empty event E, there is a probability measure $p_E(\cdot)$ satisfying the usual Kolmogrov axioms (relativized to E, so for example $p_E(E)=1$). A second approach assigns to each agent a finite sequence of probability measures $(p_1,p_2,\dots,p_n)$ called a *lexicographic probability system*. The idea is that to condition on F, first find the first probability measure not assigning zero to F and use that measure to condition on F. Roughly, one can see each of the probability measures in a lexicographic probability system as corresponding to a level of a plausibility ordering. We will return to these notions in Section 5.2.

2.4 Common Knowledge

States in a game model not only represent the player's beliefs about what their opponents will do, but also their *higher-order* beliefs about what their opponents are thinking. This means that outcomes identified as "rational" in a particular informational context will depend, in part, on these higher-order beliefs. Both game theorists and logicians have extensively discussed different notions of knowledge and belief for a group, such as common knowledge and belief. In this section, we briefly recount the standard definition of common knowledge. For more information and

pointers to the relevant literature, see Vanderschraaf & Sillari (2009) and Fagin et al., (1995: ch. 6).

Consider the statement "everyone in group I knows that E". This is formally defined as follows:

$$K_I(E) \quad := \quad \bigcap_{i\in I} K_i(E)$$

where I is any nonempty set of players. If E is common knowledge for the group I, then not only does everyone in the group know that E is true, but this fact is completely transparent to all members of the group. We first define $K_I^n(E)$ for each $n\geq 0$ by induction:

$$K_I^0(E)=E \qquad \text{and for } n\geq 1, \quad K_I^n(E)=K_I(K_I^{n-1}(E)$$

Then, following Aumann (1976), **common knowledge** of E is *defined* as the following infinite conjunction:

$$C_I(E)=\bigcap_{n\geq 0} K_I^n(E)$$

Unpacking the definitions, we have

$$C_I(E)=E\cap K_I\phi(E)\cap K_I(K_I(E))\cap K_I(K_I(K_I(E)))\cap\cdots$$

The approach to defining common knowledge outlined above can be viewed as a recipe for defining common (robust/strong) belief (simply replace the knowledge operators K_i with the appropriate belief operator). See Bonanno (1996) and Lismont & Mongin (1994, 2003) for more information about the logic of common belief. Although we do not discuss it in this entry, a probabilistic variant of common belief was introduced by Monderer & Samet (1989).

3. Choice Rules, or Choosing Optimally

There are many philosophical issues that arise in decision theory, but that is not our concern here. See Joyce 2004 and reference therein for discussions of the main philosophical issues. This section provides enough background on decision theory to understand the key results of epistemic game theory presented in the remainder of this entry.

Decision rules or *choice rules* determine what each individual player will, or should do, given her preferences and her information in a given context. In the epistemic game theory literature the most commonly used choice rules are: (strict) *dominance*, *maximization of expected utility* and *admissibility* (also known as weak dominance). One can do epistemic analysis of games using alternative choice rules, e.g., minmax regret (Halpern & Pass 2011). In this entry, we focus only on the most common ones.

Decision theorists distinguish between choice under *uncertainty* and choice under *risk*. In the latter case, the decision maker has probabilistic information about the possible states of the world. In the former case, there is no such information. There is an extensive literature concerning decision making in both types of situations (see Peterson 2009 for a discussion and pointers to the relevant literature). In the setting of epistemic game theory, the appropriate notion of a "rational choice" depends on the type of game model used to describe the informational context of the game. So, in general, "rationality" should be read as following a given choice rule. The general approach is to start with a definition of an *ir*rational choice (for instance, one that is *strictly dominated* given one's beliefs), and then define rationality as not being irrational. Some authors have recently looked at the consequences of lifting this simplifying assumption (cf., the tripartite notion of a *categorization* in Cubitt & Sugden (2011) and Pacuit & Roy (2011)), but the presentation of this goes beyond the scope of this entry.

Finally, when the underlying notion of rationality goes *beyond* maximization of expected utility, some authors have reserved the word "optimal" to qualify decisions that meet the latter requirement, but not necessarily the full requirements of rationality. See the remarks in Section 5.2 for more on this.

3.1 Maximization of Expected Utility

Maximization of expected utility is the most well-known choice rule in decision theory. Given an agent's preferences (represented as utility functions) and beliefs (represented as subjective probability measures), the expected utility of an action, or option, is the sum of the utilities of the outcomes of the action weighted by the probability that they will occur (according to the agent's beliefs). The recommendation is to choose the action that maximizes this weighted average. This idea underlies the *Bayesian* view on practical rationality, and can be straightforwardly defined in type spaces.[12] We start by defining expected utility for a player in a game.

Expected utility

Suppose that $G = \langle N, \{S_i, u_i\}_{i\in N}\rangle$ is a strategic game. A **conjecture** for player i is a probability on the set S_{-i} of strategy profiles of i's opponents. That is, a conjecture for player i is an element of $\Delta(S_{-i})$, the set of probability measures over S_{-i}. The **expected utility** of $s_i \in S_i$ with respect to a conjecture $p \in \Delta(S_{-i})$ is defined as follows:

$$EU(s_i,p) \ := \sum_{s_{-i}\in S_{-i}} p(s_{-i})u(s_i,s_{-i})$$

A strategy $s_i \in S_i$ **maximizes expected utility** for player i with respect to

$p \in \Delta(S_{-i})$ provided for all $s_i' \in S_i$, $EU(s_i,p) \geq EU(s_i',p)$. In such a case, we also say s_i is a **best response** to p in game G.

We now can define an event in a type space or epistemic-probability model where all players "choose rationally", in the sense that their choices maximize expected utility with respect to their beliefs.

Expected utility in type spaces

Let $G = \langle N, \{S_i, u_i\}_{i\in N}\rangle$ be a strategic game and $\mathcal{T} = \langle \{T_i\}_{i\in N}, \{\lambda_i\}_{i\in N}, S\rangle$ a type space for G. Recall that each t_i is associated with a probability measure $\lambda(t_i) \in \Delta(S_{-i} \times T_{-i})$. Then, for each $t_i \in T_i$, we can define a probability measure $p_{t_i} \in \Delta(S_{-i})$ as follows:

$$p_{t_i}(s_{-i}) = \sum_{t_{-i}\in T_{-i}} \lambda_i(t_i)(s_{-i}, t_{-i})$$

The set of states (pairs of strategy profiles and type profiles) where player i chooses rationally is then defined as:

$$\mathsf{Rat_i} \ := \ \{(s_i, t_i) \mid s_i \text{ is a best response to } p_{t_i}\}$$

The event that all players are *rational* is

$$\mathsf{Rat} = \{(s,t) \mid \text{ for all } i, (s_i, t_i) \in \mathsf{Rat}_i\}.$$

Notice that here *types*, as opposed to players, maximize expected utility. This is because in type structure, beliefs are associated to types (see Section 2.3 above). The reader acquainted with decision theory will recognize that this is just the standard notion of maximization of expected utility, where the space of uncertainty of each player, i.e., the possible "states of the world" on which the consequences of her action depend, is the possible combinations of types and strategy choices of the other players.

To illustrate the above definitions, consider the game in Figure 4 and the type space in Figure 11. The following calculations show that $(u, t_1) \in \mathsf{Rat}_1$ (u is the best response for player 1 given her beliefs defined by t_1):

$$\begin{aligned} EU(u, p_{t_1}) &= p_{t_1}(l)u_1(u,l) + p_{t_1}(r)u_1(u,r) \\ &= [\lambda_1(t_1)(l,t_2) + \lambda_1(t_1)(l,t_2')]\cdot u_1(u,l) \\ &\qquad + [\lambda_1(t_1)(r,t_2) + \lambda_1(t_1)(r,t_2')]\cdot u_1(u,r) \\ &= (0.5+0.4)\cdot 3 + (0+0.1)\cdot 0 \\ &= 2.7 \end{aligned}$$

$$\begin{aligned} EU(d, p_{t_1}) &= p_{t_1}(l)u_1(d,l) + p_{t_1}(r)u_1(d,r) \\ &= [\lambda_1(t_1)(l,t_2) + \lambda_1(t_1)(l,t_2')]\cdot u_1(d,l) \\ &\qquad + [\lambda_1(t_1)(r,t_2) + \lambda_1(t_1)(r,t_2')]\cdot u_1(d,r) \\ &= (0.5+0.4)\cdot 0 + (0+0.1)\cdot 1 \\ &= 0.1 \end{aligned}$$

A similar calculation shows that $(l, t_2) \in \mathsf{Rat}_2$.

Expected utility in epistemic-probability models

The definition of a rationality event is similar in an epistemic-probability model. For completeness, we give the formal details. Suppose that

$$G = \langle N, \{S_i, u_i\}_{i\in N}\rangle$$

is a strategic game and

$$\mathcal{M} = \langle W, \{\sim_i\}_{i\in N}, \{p_i\}_{i\in N}, \sigma\rangle$$

is an epistemic probability models with each p_i a prior probability measure over W. Each state $w \in W$, let

$$E_{s\_i} = \{w \in W \mid (\sigma(w))_{-i} = s_{-i}\}.$$

Then, for each state $w \in W$, we define a measure $p_w \in \Delta(S_{-i})$ as follows:

$$p_w(s_{-i}) = p(E_{s_{-i}} \mid \Pi_i(w))$$

As above,

$$\mathsf{Rat}_i := \{w \mid \sigma_i(w) \text{ is a best response to } p_w\}$$

and

$$\mathsf{Rat} := \bigcap_{i \in N} \mathsf{Rat}_i.$$

3.2 Dominance Reasoning

When a game model does not describe the players' probabilistic beliefs, we are in a situation of choice under *uncertainty*. The standard notion of "rational choice" in this setting is based on *dominance reasoning* (Finetti 1974). The two standard notions of dominance are:

> **Definition 3.1 (Strict Dominance)** Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N}\rangle$ is a strategic game and $X \subseteq S_{-i}$. Let $m_i, m_i' \in \Delta(S_i)$ be two mixed strategies for player i. The strategy m_i **strictly dominates** m_i' **with respect to** X provided
>
> $$\text{for all } s_{-i} \in X, U_i(m_i, s_{-i}) > U_i(m_i', s_{-i}).$$
>
> We say m_i is **strictly dominated** provided there is some $m_i' \in \Delta(S_i)$ that strictly dominates m_i.

A strategy $m_i \in \Delta(S_i)$ strictly dominates $m_i' \in \Delta(S_i)$ provided m_i is better than m_i' (i.e., gives higher payoff to player i) *no matter what* the other players do. There is also a weaker notion:

> **Definition 3.2 (Weak Dominance)** Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N}\rangle$ is a strategic game and $X \subseteq S_{-i}$. Let $m_i, m_i' \in \Delta(S_i)$ be two mixed strategies for player i. The strategy m_i **weakly dominates** m_i' **with respect to** X provided
>
> $$\text{for all } s_{-i} \in X, U_i(m_i, s_{-i}) \geq U_i(m_i', s_{-i})$$
>
> and
>
> $$\text{there is some } s_{-i} \in X \text{ such that } U_i(m_i, s_{-i}) > U_i(m_i', s_{-i}).$$
>
> We say m_i is **weakly dominated** provided there is some $m_i' \in \Delta(S_i)$ that weakly dominates m_i.

So, a mixed strategy m_i weakly dominates another strategy m_i' provided m_i is at least as good as m_i' no matter what the other players do *and* there is at least one situation in which m_i is strictly better than m_i'.

Before we make use of these choice rules, we need to address two potentially confusing issues about these definitions.

1. The definitions of strict and weak dominance are given in terms of mixed strategies even though we are assuming that players *only select pure strategies*. That is, we are not considering situations in which players explicitly randomize. In particular, recall that only pure strategies are associated with states in a game model. Nonetheless, it is important to define strict/weak dominance in terms of mixed strategies because there are games in which a pure strategy is strictly (weakly) dominated by a mixed strategy, but not by any of the other pure strategies.

2. Even though it is important to consider situations in which a player's pure strategy is strictly/weakly dominated by a mixed strategy, we do

not extend the above definitions to probabilities over the opponents' strategies. That is, we do not replace the above definition with

> m_i is strictly p-dominates m_i' with respect to $X \subseteq \Delta(S_{-i})$, provided for all $q \in X$, $U_i(m_i, q) > U_i(m_i', q)$.

This is because both definitions are equivalent. Obviously, p-strict dominance implies strict dominance. To see the converse, suppose that m_i' is dominated by m_i with respect to $X \subseteq S_{-i}$. We show that for all $q \in \Delta(X)$, $U_i(m_i, q) > U_i(m_i', q)$ (and so m_i' is p-strictly dominated by m_i with respect to X). Suppose that $q \in \Delta(X)$. Then,

$$U_i(m_i, q) = \sum_{s_{-i} \in S_{-i}} q(s_{-i}) U_i(m_i, s_{-i}) > \sum_{s_{-i} \in S_{-i}} q(s_{-i}) U_i(m_i', s_{-i}) = U_i(m_i', q).$$

The parameter X in the above definitions is intended to represent the set of strategy profiles that the player i take to be "live possibilities". Each state in an epistemic (-plausibility) model is associated with a such a set of strategy profiles. Given a possible world w in a game model, let $S_{-i}(w)$ denote the set of states that player i "thinks" are possible. The precise definition depends on the type of game model:

Epistemic models Suppose that

$$G = \langle N, \{S_i, u_i\}_{i \in N} \rangle$$

is a strategic game and

$$\mathcal{M} = \langle W, \{\Pi_i\}_{i \in N}, \sigma \rangle$$

is an epistemic model of G. For each player i and $w \in W$, define the set $S_{-i}(w)$ as follows:

$$S_{-i}(w) = \{\sigma_{-i}(v) \mid v \in \Pi_i(w)\}$$

Epistemic-Plausibility Models Suppose that

$$G = \langle N, \{S_i, u_i\}_{i \in N} \rangle$$

is a strategic game and

$$\mathcal{M} = \langle W, \{\sim_i\}_{i \in N}, \{\preceq_i\}_{i \in N}, \sigma \rangle$$

is an epistemic-plausibility model of G. For each player i and $w \in W$, define the set $S_{-i}(w)$ as follows:

$$S_{-i}(w) = \{\sigma_{-i}(v) \mid v \in Min_{\preceq_i}([w]_i)\}$$

In either case, we say that a choice at state w is sd-rational for player i at state w provided it is not strictly dominated with respect to $S_{-i}(w)$. The event in which i chooses rationality is then defined as

$$\mathsf{Rat}_i^{sd} := \{w \mid \sigma_i(w) \text{ is not strictly dominated with respect to } S_{-i}(w)\}.$$

In addition, we have $\mathsf{Rat}^{sd} := \bigcap_{i \in N} \mathsf{Rat}_i^{sd}$. Similarly, we can define the set of states in which player i is playing a strategy that is not weakly dominated, denoted Rat_i^{wd} and Rat^{wd} using weak dominance.

Knowledge of one's own action, the trademark of *ex-interim* situations, plays an important role in the above definitions. It enforces that $\sigma_i(w') = \sigma_i(w)$ whenever $w' \in \Pi_i(w)$. This means that player i's rationality is assessed on the basis of the result of her *current* choice according to different combinations of actions of *the other players*.

An important special case is when the players consider *all* of their opponents' strategies possible. It should be clear that a rational player will *never* choose a strategy that is strictly dominated with respect to S_{-i}. That is, if s_i is strictly dominated with respect that S_{-i}, then there is no informational context in which it is rational for player i to choose s_i. This can be made more precise using the following well-known Lemma.

> **Lemma 3.1** Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N} \rangle$ is a strategic game. A strategy $s_i \in S_i$ is strictly dominated (possibly by a mixed strategy) with respect to $X \subseteq S_{-i}$ iff there is no probability measure $p \in \Delta(X)$ such that s_i is a best response with respect to p.

The proof of this Lemma is given in the supplement, Section 1.

The general conclusion is that no dominated strategy can maximize expected utility at a given state; and, conversely, if there is a strategy that is not a best in a specific context, then it is not strictly dominated.

Similar facts hold about *weak dominance*, though the situation is more subtle. The crucial observation is that there is a characterization of weak dominance in terms of best response to certain types of probability measures. A probability measure $p \in \Delta(X)$ is said to have **full support** (with respect to X) if p assigns positive probability to every element of X (formally, $supp(p) = \{x \in X \mid p(x) > 0\} = X$). Let $\Delta^{>0}(X)$ be the set of full support probability measures on X. A full support probability on S_{-i} means that player i does not completely rule out (in the sense, that she assigns zero probability to) any strategy profiles of her opponents. The following analogue of Lemma 3.1 is also well-known:

> **Lemma 3.2** Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N} \rangle$ is a strategic game. A strategy $s_i \in S_i$ is weakly dominated (possibly by a mixed strategy) with respect to $X \subseteq S_{-i}$ iff there is no full support probability measure $p \in \Delta^{>0}(X)$ such that s_i is a best response with respect to p.

The proof of this Lemma is more involved. See Bernheim (1984: Appendix A) for a proof. In order for a strategy s_i to not be strictly dominated, it is sufficient for s_i to be a best response to a belief, whatever that belief is, about the opponents' choices. Admissibility requires something more: the strategy must be a best response to a belief that does not explicitly rule-out any of the opponents' choices. Comparing these two Lemmas, we see that strict dominance implies weak dominance, but not necessarily *vice versa*. A strategy might not be a best response to any full-support probability measure while being a best response to some particular beliefs, those assigning probability one to a state where the player is indifferent between the outcome of its present action and the potentially inadmissible one.

There is another, crucial, difference between weak and strict dominance. The following observation is immediate from the definition of strict dominance:

> **Observation 3.3** If s_i is strictly dominated with respect to X and $X' \subseteq X$, then s_i is strictly dominated with respect to X'.

If a strategy is strictly dominated, it remains so if the player gets more information about what her opponents (might) do. Thus, if a strategy s_i is strictly dominated in a game G with respect to the *entire* set of her opponents' strategies S_{-i}, then it will never be rational (according to the above definitions) in any epistemic (-plausibility) model for G. I.e., there are no beliefs player i can have that makes s_i rational. The same observation does not hold for weak dominance. The existential part of the definition of weak dominance means that the analogue of Observation 3.3 does not hold for weak dominance: if s_i is weakly dominated with respect to X then it need not be the case that s_i is weakly dominated with respect to some $X' \subseteq X$.

4. Fundamentals

The epistemic approach to game theory focuses on the choices of *individual* decision makers in specific informational contexts, assessed on the basis of decision-theoretic choice rules. This is a bottom-up, as opposed to the classical top-down, approach. Early work in this paradigm include Bernheim (1984) and Pearce's (1984) notion of *rationalizability* and Aumann's *derivation* of correlated equilibrium from the minimal assumption that the players are "Bayesian rational" (Aumann 1987).

An important line of research in epistemic game theory asks under what *epistemic* conditions will players follow the recommendations of particular solution concept? Providing such conditions is known as an *epistemic characterization* of a solution concept.

In this section, we present two fundamental epistemic characterization results. The first is a characterization of iterated removal of strictly dominated strategies (henceforth ISDS), and the second is a characterization of backward induction. These epistemic characterization results are historically important. They mark the beginning of epistemic game theory as we know it today. Furthermore, they are also conceptually important. The developments in later sections build on the ideas presented in this section.

4.1 Iterated Removal of Strictly Dominated Strategies

The central result of epistemic game theory is that "rationality and common belief in rationality *implies* iterated elimination of strictly dominated strategies." This result is already covered in Vanderschraaf & Sillari (2009). For that reason, instead of focusing on the formal details, the emphasis here will be on its significance for the epistemic foundations of game theory. One important message is that the result highlights the importance of *higher-order information*.

4.1.1 The Result

Iterated elimination of strictly dominated strategies (ISDS) is a solution concept that runs as follows. First, remove from the original game any strategy that is strictly dominated for player i (with respect to all of the opponents' strategy profiles). After having removed the strictly dominated strategies in the original game, look at the resulting sub-game, remove the strategies which have become strictly dominated there, and repeat this process until the elimination does not remove any strategies. The profiles that survive this process are said to be *iteratively non-dominated*.

For example, consider the following strategic game:

		Bob		
		l	*c*	*r*
Ann	*t*	3,3	1,1	0,0
	m	1,1	3,3	1,0
	b	0,4	0,0	4,0

Figure 13

Note that r is strictly dominated for player 2 with respect to $\{t,m,b\}$. Once r is removed from the game, we have b is strictly dominated for player 1 with respect to $\{l,c\}$. Thus, $\{(t,l),(t,c),(m,l),(m,c)\}$ are iteratively undominated. That is, iteratively removing strictly dominated strategies generates the following sequence of games:

$$\begin{array}{c|c|c|c|} & l & c & r \\ \hline t & 3,3 & 1,1 & 0,0 \\ \hline m & 1,1 & 3,3 & 1,0 \\ \hline b & 0,4 & 0,0 & 4,0 \end{array} \rightarrowtail \begin{array}{c|c|c|} & l & c \\ \hline t & 3,3 & 1,1 \\ \hline m & 1,1 & 3,3 \\ \hline b & 0,4 & 0,0 \end{array} \rightarrowtail \begin{array}{c|c|c|} & l & c \\ \hline t & 3,3 & 1,1 \\ \hline m & 1,1 & 3,3 \\ \hline \end{array}$$

Figure 14

For arbitrary large (finite) strategic games, if all players are *rational* and there is **common belief that all players are rational**, then they will choose a strategy that is iteratively non-dominated. The result is credited to Bernheim (1984) and Pearce (1984). See Spohn (1982) for an early version, and Brandenburger & Dekel (1987) for the relation with correlated equilibrium.

Before stating the formal result, we illustrate the result with an example. We start by describing an "informational context" of the above game. To that end, define a type space $\mathcal{T} = \langle \{T_1, T_2\}, \{\lambda_1, \lambda_2\}, S\rangle$, where S is the strategy profiles in the above game, there are two types for player 1 ($T_1 = \{t_1, t_2\}$) and three types for player 2 ($T_2 = \{s_1, s_2, s_3\}$). The type functions λ_i are defined as follows:

$\lambda_1(t_1)$	l	c	r
s_1	0.5	0.5	0
s_2	0	0	0
s_3	0	0	0

$\lambda_1(t_2)$	l	c	r
s_1	0	0.5	0
s_2	0	0	0.5
s_3	0	0	0

$\lambda_2(s_1)$	t	m	b
t_1	0.5	0.5	0
t_2	0	0	0

$\lambda_2(s_2)$	t	m	b
t_1	0.25	0.25	0
t_2	0.25	0.25	0

$\lambda_2(s_3)$	t	m	b
t_1	0.5	0	0
t_2	0	0	0.5

Figure 15

We then consider the pairs (s, t) where $s \in S_i$ and $t \in T_i$ and identify all the rational pairs (i.e., where s is a best response to $\lambda_i(t)$, see the previous section for a discussion):

- $\mathsf{Rat}_1 = \{(t, t_1), (m, t_1), (b, t_2)\}$
- $\mathsf{Rat}_2 = \{(l, s_1), (c, s_1), (l, s_2), (c, s_2), (l, s_3)\}$

The next step is to identify the types that *believe* that the other players are rational. In this context, belief means *probability 1*. For the type t_1, we have $\lambda_1(t_1)(\mathsf{Rat}_2) = 1$; however,

$$\lambda_1(t_2)(s_2, r) = 0.5 > 0,$$

but $(r, s_2) \notin \mathsf{Rat}_2$, so t_2 does not believe that player 2 is rational. This can be turned into an iterative process as follows: Let $R_i^1 = \mathsf{Rat}_\mathsf{i}$. We first need some notation. Suppose that for each i, R_i^n has been defined. Then, define R_{-i}^n as follows:

$$R_{-i}^n = \{(s, t) \mid s \in S_{-i}, t \in T_{-j}, \text{and for each } j \neq i, (s_j, t_j) \in R_j^n\}.$$

For each $n > 1$, define R_i^n inductively as follows:

$$R_i^{n+1} = \{(s, t) \mid (s, t) \in R_i^n \text{ and } \lambda_i(t) \text{ assigns probability 1 to } R_{-i}^n\}$$

Thus, we have $R_1^2 = \{(t, t_1), (m, t_1)\}$. Note that s_2 assigns non-zero probability to the pair (m, t_2) which is not in R_1^1, so s_2 does not believe that 1 is rational. Thus, we have $R_2^2 = \{(l, s_1), (c, s_1), (l, s_3)\}$. Continuing with this process, we have $R_1^2 = R_1^3$. However, s_3 assigns non-zero probability to (b, t_2) which is not in R_1^2, so $R_2^3 = \{(l, s_1), (c, s_1)\}$. Putting everything together, we have

$$\bigcap_{n \geq 1} R_1^n \times \bigcap_{n \geq 1} R_2^n = \{(t, t_1), (m, t_1)\} \times \{(l, s_1), (c, s_1)\}.$$

Thus, all the profiles that survive iteratively removing strictly dominated strategies ($\{(t,l),(m,l),(t,c),(m,c)\}$) are consistent with states where the players are rational and commonly believe they are rational.

Note that, the above process need not generate *all* strategies that survive iteratively removing strictly dominated strategies. For example, consider a type space with a single type for player 1 assigning probability 1 to the single type of player 2 and l, and the single type for player 2 assigning probability 1 to the single type for player 1 and u. Then, (u,l) is the only strategy profile in this model and obviously rationality and common belief of rationality is satisfied. However, for any type space, if a strategy profile is consistent with rationality and common belief of rationality, then it must be a strategy that is in the set of strategies that survive iteratively removing strictly dominated strategies.

> **Theorem 4.1** Suppose that G is a strategic game and $\mathcal{T}$ is any type space for G. If (s,t) is a state in $\mathcal{T}$ in which all the players are rational and there is common belief of rationality—formally, for each i,
>
> $$(s_i,t_i) \in \bigcap_{n\geq 1} R_i^n$$
>
> —then s is a strategy profile that survives iteratively removal of strictly dominated strategies.

This result establishes *sufficient* conditions for ISDS. It has also a converse direction: given any strategy profile that survives iterated elimination of strictly dominated strategies, there is a model in which this profile is played where all players are rational and this is common knowledge. In other words, one can always *view* or *interpret* the choice of a strategy profile that would survive the iterative elimination procedure as one that results from common knowledge of rationality. Of course, this form of the converse is not particularly interesting as we can always define a type space where all the players assign probability 1 to the given strategy profile (and everyone playing their requisite strategy). Much more interesting is the question whether the *entire* set of strategy profiles that survive iteratively removal of strictly dominated strategies is consistent with rationality and common belief in rationality. This is covered by the following theorem of Brandenburger & Dekel (1987) (cf. also Tan & Werlang 1988):

> **Theorem 4.2** For any game G, there is a type structure for that game in which the strategy profiles consistent with rationality and common belief in rationality is the set of strategies that survive iterative removal of strictly dominated strategies.

See Friedenberg & Keisler (2010) for the strongest versions of the above results. Analogues of the above results have been proven using different game models (e.g., epistemic models, epistemic-plausibility models, etc.). For example, see Apt & Zvesper (2010) proofs of corresponding theorems using Kripke models.

4.1.2 Philosophical Issues

Many authors have pointed out the strength of the common belief assumption in the results of the previous section (see, e.g., Gintis 2009; Bruin 2010). It requires that the players not only believe that the others are not choosing an irrational strategy, but also to believe that everybody believes that nobody is choosing an irrational strategy, and everyone believes that everyone believes that everyone believes that nobody is choosing an irrational strategy, and so on. It should be noted, however, that this unbounded character is there only to ensure that the result holds for *arbitrary* finite games. For a particular game and a model for it, a finite iteration of "everybody believes that" suffices to ensure a play that survives the iterative elimination procedure.

A possible reply to the criticism of the infinitary nature of the common belief assumption is that the result should be seen as the analysis of a *benchmark* case, rather than a description of genuine game playing situations or a prescription for what rational players *should* do (Aumann 2010). Indeed, common knowledge/belief of rationality has long been used as an informal explanation of the idealizations underlying classical game-theoretical analyses (Myerson 1991). The results above show that, once formalized, this assumption does indeed lead to a classical solution concept, although, interestingly, *not* the well-known Nash equilibrium, as is often informally claimed in early game-theoretic literature. Epistemic conditions for Nash equilibrium are presented in Section 5.1.

The main message to take away from the results in the previous section is: Strategic reasoning in games involves higher-order information. This means that, in particular,

> "Bayesian rationality" alone—i.e., maximization of expected utility—is not sufficient to ensure a strategy profile is played that is iteratively undominated, in the general case.

In general, first-order belief of rationality will not do either. Exactly how many levels of beliefs is needed to guarantee "rational play" in game situations is still the subject of much debate (Kets 2014; Colman 2003; de Weerd, Verbrugge, & Verheij 2013; Rubinstein 1989). There are two further issues we need to address.

First of all, how can agents arrive at a context where rationality is commonly believed? The above results do not answer that question. This has been the subject of recent work in Dynamic Epistemic Logic (van Benthem 2003). In this literature, this question is answered by showing that the agents can eliminate all higher-order uncertainty regarding each others' rationality, and thus ensure that no strategy is played that would not

survive the iterated elimination procedure, by repeatedly and publicly "*announcing*" that they are not irrational. In other words, iterated public announcement of rationality makes the players' expectations converge towards sufficient epistemic conditions to play iteratively non-dominated strategies. For more on this dynamic view on solution epistemic characterization see van Benthem (2003); Pacuit & Roy (2011); van Benthem & Gheerbrant (2010); and van Benthem, Pacuit, & Roy (2011).

Second of all, when there are more than two players, the above results only hold if players can believe that the choices of their opponents are *correlated* (Bradenburger & Dekel 1987; Brandenburger & Friedenberg 2008). The following example from Brandenburger & Friedenberg (2008) illustrates this point. Consider the following three person game where Ann's strategies are $S_A = \{u, d\}$, Bob's strategies are $S_B = \{l, r\}$ and Charles' strategies are $S_C = \{x, y, z\}$ and their respective preferences for each outcome are given in the corresponding cell:

	l	r
u	1,1,3	1,0,3
d	0,1,0	0,0,0

x

	l	r
u	1,1,2	1,0,0
d	0,1,0	1,1,2

y

	l	r
u	1,1,0	1,0,0
d	0,1,3	0,0,3

z

Figure 16

Note that y is not strictly dominated for Charles. It is easy to find a probability measure $p \in \Delta(S_A \times S_B)$ such that y is a best response to p. Suppose that $p(u,l) = p(d,r) = \frac{1}{2}$. Then, $EU(x,p) = EU(z,p) = 1.5$ while $EU(y,p) = 2$. However, there is no probability measure $p \in \Delta(S_A \times S_B)$ such that y is a best response to p and $p(u,l) = p(u) \cdot p(l)$ (i.e., Charles believes that Ann and Bob's choices are independent). To see this, suppose that a is the probability assigned to u and b is the probability assigned to l. Then, we have:

- The expected utility of y is

$$2ab + 2(1-a)(1-b);$$

- The expected utility of x is

$$3ab + 3a(1-b) = 3a(b + (1-b)) = 3a;$$

  and

- The expected utility of z is

$$\begin{aligned} 3(1-a)b + 3(1-a)(1-b) &= 3(1-a)(b + (1-b)) \\ &= 3(1-a). \end{aligned}$$

There are three cases:

1. Suppose that $a = 1 - a$ (i.e., $a = 1/2$). Then,

$$\begin{aligned} 2ab + 2(1-a)(1-b) &= 2ab + 2a(1-b) \\ &= 2a(b + (1-b)) \\ &= 2a < 3a. \end{aligned}$$

   Hence, y is not a best response.

2. Suppose that $a > 1 - a$. Then,

$$2ab + 2(1-a)(1-b) < 2ab + 2a(1-b) = 2a < 3a.$$

   Hence, y is not a best response.

3. Suppose that $1 - a > a$. Then,

$$\begin{aligned} 2ab + 2(1-a)(1-b) &< 2(1-a)b + 2(1-a)(1-b) \\ &= 2(1-a) \\ &< 3(1-a). \end{aligned}$$

   Hence, y is not a best response.

In all of the cases, y is not a best response.

4.2 Backward induction

The second fundamental result analyzes the consequences of rationality and common belief/knowledge of rationality in *extensive games* (i.e., trees instead of matrices). Here, the most well-known solution concept is the so-called *subgame perfect equilibrium*, also known as *backward induction* in games of perfect information. The epistemic characterization of this solution concept is in terms of “substantive rationality” and common belief that all players are substantively rational (cf. also Vanderschraaf & Sillari 2009: sec. 2.8). The main point that we highlight in this section, which is by now widely acknowledged in the literature, is:

> Belief revision policies play a key role in the epistemic analysis of extensive games

The most well-known illustration of this is through the comparison of two apparently contradictory results regarding the consequences of assuming rationality and common knowledge of rationality in extensive games. Aumann (1995) showed that this epistemic condition implies that the players will play according to the backward induction solution while Stalnaker (1998) argued that this is not necessarily true. The crucial difference between these two results is the way in which they model the players’ belief change upon (hypothetically) learning that an opponent has deviated from the backward induction path.

4.2.1 Extensive games: basic definitions

Extensive games make explicit the sequential structure of choices in a game situation. In this section, we focus on games of *perfect information* in which there is no uncertainty about earlier choices in the game. These games are represented by tree-like structures:

Definition 4.3 (Perfect Information Extensive Game) An extensive game is a tuple $\langle N, T, Act, \tau, \{u_i\}_{i\in N}\rangle$, where

- N is a finite set of players;
- T is a tree describing the temporal structure of the game situation: Formally, T consists of a set of nodes and an immediate successor relation $\rightarrowtail$. Let Z denote the set of terminal nodes (i.e., nodes without any successors) and V the remaining nodes (called decision nodes). Let v_0 denote the initial node (i.e., the **root** of the tree). The edges at a decision node $v \in V$ are each labeled with **actions** from a set Act. Let $Act(v)$ denote the set of actions available at v. Let $\rightsquigarrow$ be the transitive closure of $\rightarrowtail$.
- τ is a turn function assigning a player to each node $v \in V$ (for a player $i \in N$, let $V_i = \{v \in V \mid \tau(v) = i\}$).
- $u_i : Z \to \mathbb{R}$ is the utility function for player i assigning real numbers to outcome nodes.

A **strategy** is a term of art in extensive games. It denotes a plan for every eventuality, which tells an agent what to do at all histories she is to play, even those which are excluded by the strategy itself.

Definition 4.4 (Strategies) A **strategy** for player i is a function $s_i : V_i \to Act$ where for all $v \in V_i$, $s_i(v) \in Act(v)$. A strategy profile, denoted $\mathbf{s}$, is an element of $\Pi_{i\in N} S_i$. Given a strategy profile $\mathbf{s}$, let $\mathbf{s}_i$ be player i's component of $\mathbf{s}$ and $\mathbf{s}_{-i}$ the sequence of strategies form $\mathbf{s}$ for all players except i.

Each strategy profile $\mathbf{s}$ generates a path through an extensive game, where a path is a maximal sequence of nodes from the extensive game ordered by the immediate successor relations $\rightarrowtail$. We say that v is **reached** by a strategy profile $\mathbf{s}$ is v is on the path generated by $\mathbf{s}$. Suppose that v is any node in an extensive game. Let $out(v, \mathbf{s})$ be the terminal node that is reached if, starting at node v, all the players move according to their respective strategies in the profile $\mathbf{s}$. Given a decision node $v \in V_i$ for player i, a strategy s_i for player i, and a set $X \subseteq S_{-i}$ of strategy profiles of the opponents of i, let $Out_i(v, s_i, X) = \{out(v, (s_i, s_{-i})) \mid s_{-i} \in X\}$. That is, $Out_i(v, s_i, X)$ is the set of terminal nodes that may be reached if, starting at node v, player i uses strategy s_i and i's opponents use strategy profiles from X.

The following example of a perfect information extensive game will be used to illustrate these concepts. The game is an instance of the well-known *centipede game*, which has played an important role in the epistemic game theory literature on extensive games.

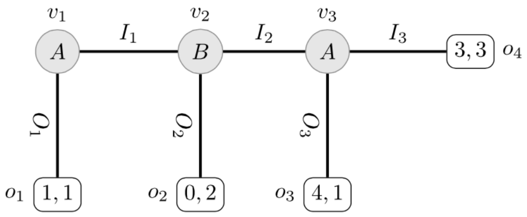

Figure 17: An extensive game

The decision nodes for A and B respectively are $V_A = \{v_1, v_3\}$ and $V_B = \{v_2\}$; and the outcome nodes are $O = \{o_1, o_2, o_3, o_4\}$. The labels of the edges in the above tree are the actions available to each player. For instance, $Act(v_1) = \{O_1, I_1\}$. There are four strategies for A and two strategies for B. To simplify notation, we denote the players' strategies by the sequence of choices at each of their decision nodes. For example, A's strategy s_A^1 defined as $s_A^1(v_1) = O_1$ and $s_A^1(v_3) = O_3$ is denoted by the sequence O_1O_3. Thus, A's strategies are: $s_A^1 = O_1O_3$, $s_A^2 = O_1I_3$, $s_A^3 = I_1O_3$ and $s_A^4 = I_1I_3$. Note that A's strategy s_A^2 specifies a move at v_3, even though the earlier move at v_1, O_1, means that A will not be given a chance to move at v_3. Similarly, Bob's strategies will be denoted by $s_B^1 = O_2$ and $s_B^2 = I_2$, giving the actions chosen by B at his decision node. Then, for example, $out(v_2, (s_A^2, s_B^2)) = o_4$. Finally, if $X = \{s_A^1, s_A^4\}$, then $Out_B(v_2, s_B^2, X) = \{o_3, o_4\}$.

4.2.2 Epistemic Characterization of Backward Induction

There are a variety of ways to describe the players' knowledge and beliefs in an extensive game. The game models vary according to which epistemic attitudes are represented (e.g., knowledge and/or various notions of beliefs) and precisely how the players' disposition to revise their beliefs during a play of the game is represented. Consult Battigalli, Di Tillio, & Samet (2013); Baltag, Smets, & Zvesper (2009); and Battigalli & Siniscalchi (2002) for a sampling of the different types of models found in the literature.

One of the simplest approaches is to use the epistemic models introduced in Section 2.2 (cf. Aumann 1995; Halpern 2001b). An epistemic model of an extensive game $G = \langle N, T, Act, \tau, \{u_i\}_{i\in N}\rangle$ is a tuple $\langle W, \{\Pi_i\}_{i\in N}, \sigma\rangle$ where W is a nonempty set of states; for each $i \in N$, Π_i is a partition on W; and $\sigma : W \to \Pi_{i\in N} S_i$ is a function assigning to each state w, a strategy profile from G. If $\sigma(w) = \mathbf{s}$, then we write $\sigma_i(w)$ for $\mathbf{s}_i$ and $\sigma_{-i}(w)$ for $\mathbf{s}_{-i}$. As usual, we assume that players know their own strategies: for all $w \in W$, if $w' \in \Pi_i(w)$, then $\sigma_i(w) = \sigma_i(w')$.

The rationality of a strategy at a decision node depends both on what actions the strategy prescribes at all future decision nodes *and* what the players know about the strategies that their opponents are following. Let $S_{-i}(w) = \{\sigma_{-i}(w') \mid w' \in \Pi_i(w)\}$ be the set of strategy profiles of player i's opponents that i thinks are possible at state w. Then, $Out_i(v, s_i, S_{-i}(w))$ is the set of outcomes that player i thinks are possible starting at node v if she follows strategy s_i.

> **Definition 4.5 (Rationality at a decision node)** Player i is **rational at node** $v \in V_i$ **in state** w provided, for all strategies s_i such that $s_i \neq \sigma_i(w)$, there is an $o' \in Out_i(v, s_i, S_{-i}(w))$ and $o \in Out_i(v, \sigma_i(w), S_{-i}(w))$ such that $u_i(o) \geq u_i(o')$.

So, a player i is rational at a decision node $v \in V_i$ in state w provided that i does not know that there is an alternative strategy that would give her a higher payoff.

> **Definition 4.6 (Substantive rationality)** Player i is **substantively rational at state** w provided for all decision nodes $v \in V_i$, i is rational at v in state w.

We can define the event that player i is substantively rational is the standard way: $\mathsf{Rat}_i = \{w \mid \text{player } i \text{ is substantively rational at state } w\}$; and so, the event that all players are substantively rational is $\mathsf{Rat} = \bigcap_{i\in N} \mathsf{Rat}_i$.

This notion of rationality at a decision node v is forward-looking in the sense that it only takes account of the possibilities that can arise *from that point on* in the game. It does not take account of the previous moves leading to v—i.e., which choices have or could have lead to v. We shall return to this in the discussion below.

An important consequence of this is that the rationality of choices at nodes that are only followed by terminal nodes are independent of the relevant player's knowledge. Call a node v **pre-terminal** if all of v's immediate successors are terminal nodes. At such nodes, it does not matter what strategies the player thinks are possible: If v is a pre-terminal node and player i is moving at v, then for all states in w in an epistemic model of the game and for all strategies $s_i \in S_i$, $Out_i(v, s_i, S_{-i}(w)) = \{s_i(v)\}$. This means, for example, that for any state w in an epistemic model for the extensive game in Figure 17, the only strategies that are rational at node v_3 in w are those that prescribe that A chooses O_3 at node v_3. Therefore, if $w \in \mathsf{Rat}_A$, then $\sigma_A(w)(v_3) = O_3$. Whatever A knows, or rather knew about what B would do, if the game reaches the node v_3, then the only rational choice for A is O_3.

Information about the rationality of players at pre-terminal nodes is very important for players choosing earlier in the game. Returning to the game in Figure 17, if B knows that A is substantively rational at a state w in an epistemic model of the game, then $\Pi_B(w) \subseteq \mathsf{Rat}_A$. Given the above argument, this means that if $w' \in \Pi_B(w)$, then $\sigma_A(w')(v_3) = O_3$. Thus, we have for any state w in an epistemic model of the game,

$$Out_B(v_2, I_2, S_{-i}(w)) = \{o_3\};$$

and, of course,

$$Out_B(v_2, O_2, S_{-i}(w)) = \{o_2\}.$$

But then, (O_2) is the only strategy that is rational for B at v_2 in any state w (this follows since $u_B(o_2) = 2 \geq 1 = u_B(o_3)$). This means that if $w \in \mathsf{Rat}_B$ and $\Pi_B(w) \subseteq \mathsf{Rat}_A$, then $\sigma_B(w)(v_2) = O_2$. Finally, if A knows that B knows that A is substantively rational, then

$$\Pi_A(w) \subseteq K_B\mathsf{Rat}_A = \{w' \mid \Pi_B(w') \subseteq \mathsf{Rat}_A\}.$$

A similar argument shows that if $w \in \mathsf{Rat}_A$ and $w \in K_A(K_B(\mathsf{Rat}_A))$, then $\sigma_A(w)(v_1) = O_1$.

The strategy profile (O_1O_3, O_2) is the unique pure-strategy *sub-game perfect equilibrium* (Selten 1975) of the game in Figure 17. Furthermore, the reasoning that we went through in the previous paragraphs is very close to *backward induction* algorithm. This algorithm can be used to calculate the sub-game perfect equilibrium in any perfect information game in which all players receive unique payoffs at each outcome.[13] The algorithm runs as follows:

> **BI Algorithm** At terminal nodes, players already have the nodes marked with their utilities. At a non-terminal node v, once all immediate successors are marked, the node is marked as follows: find the immediate successor d that has the highest utility for player $\tau(v)$ (the players whose turn it is to move at v). Copy the utilities from d onto v.

Given a marked game tree, the unique path that leads from the root v_0 of the game tree to the outcome with the utilities that match the utilities assigned to v_0 is called the **backward induction path**. In fact, the markings on each and every node (even nodes not on the backward induction path) defines a unique path through the game tree. These paths can be used to define strategies for each player: At each decision node v, choose the action that is consistent with the path from v. Let BI denote the resulting **backward induction profile** (where each player is following the strategy given by the backward induction algorithm).

Aumann (1995) showed that the above reasoning can be carried out for any extensive game of perfect information.

> **Theorem 4.7 (Aumann 1995)** Suppose that G is an extensive game of perfect information and $\mathbf{s}$ is a strategy profile for G. The following are equivalent:

1. There is a state w in an epistemic model of G such that $\sigma(w) = \mathbf{s}$ and $w \in C_N(\mathsf{Rat})$ (there is common knowledge that all players are substantively rational).

2. $\mathbf{s}$ is a sub-game perfect equilibrium of G.

This result has been extensively discussed. The standard ground of contention is that common knowledge of rationality used in this argument seems *self-defeating*, at least intuitively. Recall that we asked what would B do at node v_2 under common knowledge of rationality, and we concluded that he would choose O_2. But, if the game ever reaches that state, then, by the theorem above, B has to conclude that either A is not rational, or that she does not know that he is. Both violate common knowledge of rationality. Is there a contradiction here? This entry will not survey the extensive literature on this question. The reader can consult the references in Bruin 2010. Our point here is rather that how one looks at this potential paradox hinges on the way the players will revise their beliefs in "future" rationality in the light of observing a move that would be "irrational" under common knowledge of rationality.

4.2.3 Common Knowledge of Rationality without Backward Induction

Stalnaker (1996, 1998) offers a different perspective on backward induction. The difference with Aumann's analysis is best illustrated with the following example:

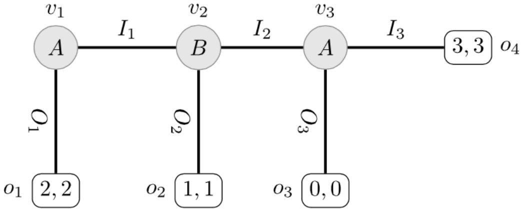

Figure 18: An extensive game

In the above game the backward induction profile is (I_1I_3, I_2) leading to the outcome o_4 with both players receiving a payoff of 3. Consider an epistemic model with a single state w where $\sigma(w) = (O_1I_3, O_2)$. This is not the backward induction profile, and so, by Aumann's Theorem (Theorem 4.7) it cannot be common knowledge among A and B at state w that both A and B are substantively rational.

Recall that a strategy for a player i specifies choices at *all* decision nodes for i, even those nodes that are impossible to reach given earlier moves prescribed by the strategy. Thus, strategies include "counterfactual" information about what players would do if they were given a chance to move at each of their decision nodes. In the single state epistemic model, B knows that A is following the strategy O_1I_3. This means that B knows two things about A's choice behavior in the game. The first is that A is choosing O_1 initially. The second is that if A where given the opportunity to choose at v_3, then she would choose I_3. Now, given B's knowledge about what A is doing, there is a sense in which *whatever B would choose at v_2*, his choice is rational. This follows trivially since A's initial choice prescribed by her strategy at w makes it impossible for B to move. Say that a player i is **materially rational** at a state w in an epistemic model of a game if i is rational at all decision nodes $v \in V_i$ in state w that are reachable according to the strategy profile $\sigma(w)$. We have seen that B is trivially materially rational. Furthermore, A is materially rational since she knows that B is choosing O_2 (i.e., $S_{-A}(w) = \{O_2\}$). Thus, $Out_A(v_1, O_1, S_{-i}(w)) = \{o_1\}$ and $Out_A(v_1, O_I, S_{-i}(w)) = \{o_2\}$; and so, A's choice of O_1 at v_1 makes her materially rational at w. The main point of contention between Aumann and Stalnaker boils down to whether the single state epistemic model includes enough information about what exactly B thinks about A's choice at v_3 when assessing the rationality of B's *hypothetical* choice of O_2 at v_2.

According to Aumann, B is not substantively rational: Since $S_{-B}(w) = \{O_1I_3\}$, we have

$$Out_B(v_2, O_2, S_{-B}(w)) = \{o_2\}$$

and

$$Out_B(v_2, I_2, S_{-B}(w)) = \{o_4\};$$

and so, B is not rational at v_2 in w (note that $u_B(o_4) = 3 > 1 = u_B(o_2)$). Stalnaker suggests that the players should be endowed with a *belief revision policy* that describes which informational state they would revert to in case they were to observe moves that are inconsistent with what they know about their opponents' strategies. If B does learn that he can in fact move, then he has learned something about A's strategy. In particular, he now knows that she cannot be following any strategy that prescribes that she chooses O_1 at v_1 (so, in particular, she cannot be following the strategy O_1I_3.) Suppose that B is disposed to react to surprising information about A's choice of strategy as follows: Upon learning that A is not following a strategy in which she chooses O_1 at v_1, he concludes that she is following strategy I_1O_3. That is, B's "belief revision policy" can be summarized as follows: If A makes one "irrational move", then she will make another one. Stalnaker explains the apparent tension between this belief revision policy and his knowledge that if A where given the opportunity to choose at v_3, then she would choose I_3 as follows:

> To think there is something incoherent about this combination of beliefs and belief revision policy is to confuse epistemic with causal counterfactuals—it would be like thinking that because I believe that if Shakespeare hadn't written Hamlet, it would have never been written by anyone, I must therefore be disposed to conclude that Hamlet was never written, were I to learn that Shakespeare was in fact not its author. (Stalnaker 1996: 152)

Then, with respect to B's appropriately updated knowledge about A's choice at v_3 (according to his specified belief revision policy), his strategy O_2 is in fact rational. According to Stalnaker, the rationality of a choice at a node v should be evaluated in the (counterfactual) epistemic state the player *would be in if that node was reached*. Assuming A knows that B is using the belief revision policy described above, then A knows that B is substantively rational in Stalnaker's sense. If the model includes explicit information about the players' belief revision policy, then there can be common knowledge of substantive rationality (in Stalnaker's sense) yet the players' choices do not conform to the backward induction profile.

4.3 Common strong belief and forward induction

In the previous section, we assumed that the players interpret an opponent's deviation from expected play in an extensive game (e.g., deviation from the backward induction path) as an indication that that player will choose "irrationally" at future decision nodes. However, this is just one example of a belief revision policy. It is not suggested that this is the belief revision policy that players *should* adopt. Stalnaker's central claim is that models of extensive games should include a component that describes the players' disposition to change their beliefs during a play of the game, which may vary from model to model or even among the players in a single model:

> Faced with surprising behavior in the course of a game, the players must decide what then to believe. Their strategies will be based on how their beliefs would be revised, which will in turn be based on their epistemic priorities—whether an unexpected action should be regarded as an isolated mistake that is thereby epistemically independent of beliefs about subsequent actions, or whether it reveals, intentionally or inadvertently, something about the player's expectations, and so about the way she is likely to behave in the

> future. The players must decide, but the theorists should not—at least they should not try to generalize about epistemic priorities that are meant to apply to any rational agent in all situations. (Stalnaker 1998: 54)

One belief revision policy that has been extensively discussed in the epistemic game theory literature is the **rationalizability principle**. Battigalli (1997) describes this belief revision policy as follows:

> **Rationalizability Principle** A player should always try to interpret her information about the behavior of her opponents assuming that they are not implementing 'irrational' strategies.

This belief revision policy is closely related to so-called *forward induction reasoning*. To illustrate, consider the following imperfect information game:

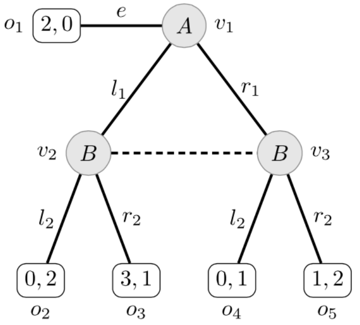

Figure 19

In the above game, A can either exit the game initially (by choosing e) for a guaranteed payoff of 2 or decide to play a game of imperfect information with B. Notice that r_1 is strictly dominated by e: No matter what B chooses at v_3, A is better off choosing e. This means that if A is following a rational strategy, then she will not choose r_1 at v_1. According to the rationalizability principle, B is disposed to believe that A did not choose r_1 if he is given a chance to move. Thus, assuming that B knows the structure of the game and revises his beliefs according to the rationalizability principle, his only rational strategy is to choose l_2 at his informational cell (consisting of $\{v_2, v_3\}$). If A can anticipate this reasoning, then her only rational strategy is to choose e at v_1. This is the forward induction outcome of the above game.

Battigalli & Siniscalchi (2002) develop an epistemic analysis of forward induction reasoning in extensive games (cf. also, Stalnaker 1998: sec. 6). They build on an idea of Stalnaker (1998, 1996) to characterize forward induction solution concepts in terms of common *strong belief* in rationality. We discussed the definition of "strong belief" in Section 2.4. The mathematical representation of beliefs in Battigalli & Siniscalchi (2002) is different, although the underlying idea is the same. A player strongly believes an event E provided she believes E is true at the beginning of the game (in the sense that she assigns probability 1 to E) and continues to believe E as long as it is not falsified by the evidence. The *evidence* available to a player in an extensive game consists of the observations of the previous moves that are consistent with the structure of the game tree—i.e., the paths through a game tree. A complete discussion of this approach is beyond the scope of the entry. Consult Battigalli & Siniscalchi (2002); Baltag et al. (2009); Battigalli & Friedenberg (2012); Bonanno (2013); Perea (2012, 2014); and van Benthem & Gheerbrant (2010) for a discussion of this approach and alternative epistemic analyses of backward and forward induction.

5. Developments

In this section, we present a number of results that build on the methodology presented in the previous section. We discuss the characterization of the Nash equilibrium, incorporate considerations of weak dominance into the players' reasoning and allow the players to be *unaware*, as opposed to *uncertain*, about some aspects of the game.

5.1 Nash Equilibrium

5.1.1 The Result

Iterated elimination of strictly dominated strategies is a very intuitive concept, but for many games it does not tell anything about what the players will or should choose. In coordination games (Figure 1 above) for instance, all profiles, can be played under rationality and common belief of rationality.

Looking again at Figure 1, one can ask what would happen if Bob *knew* (that is had correct beliefs about) Ann's strategy choice? Intuitively, it is quite clear that his rational choice is to coordinate with her. If he *knows* that she plays t, for instance, then playing l is clearly the only rational choice for him, and similarly, if he knows that she plays b, then r is the only rational choice. The situation is symmetric for Ann. For instance, if she knows that Bob plays l, then her only rational choice is to choose t. More formally, the only states where Ann is rational and her type *knows* (i.e., is correct and assigns probability 1 to) Bob's strategy choice and where Bob is also rational and his type *knows* Ann's strategy choices are states where they play either (t,l) or (b,r), the pure-strategy Nash equilibria of the game.

A *Nash equilibrium* is a profile where no player has an incentive to unilaterally deviate from his strategy choice. In other words, a Nash equilibrium is a combination of (possibly mixed) strategies such that they all play their best response given the strategy choices of the others. Again, (t,l) and (b,r) are the only pure-strategy equilibria of the above coordination game. Nash equilibrium, and its numerous refinements, is arguably the game theoretical solution concept that has been most used in game theory (Aumann & Hart 1994) and philosophy (e.g., famously in Lewis 1969).

The seminal result of Aumann & Brandenburger 1995 provides an epistemic characterization of the Nash equilibrium in terms of *mutual knowledge* of strategy choices (and the structure of the game). See, also, Spohn (1982) for an early statement. Before stating the theorem, we discuss an example from Aumann & Brandenburger (1995) that illustrates the key ideas. Consider the following coordination game:

		B	
		l	r
A	u	2,2	0,0
	d	0,0	1,1

Figure 20

The two pure-strategy Nash equilibria are (u,l) and (d,r) (there is also a mixed-strategy equilibrium). As usual, we fix an informational context for this game. Let $\mathcal{T}$ be a type space for the game with three types for each player $T_A = \{a_1, a_2, a_3\}$ and $T_B = \{b_1, b_2, b_3\}$ with the following type functions:

	l	r
b_1	0.5	0.5
b_2	0	0
b_3	0	0

$\lambda_A(a_1)$

	l	r
b_1	0.5	0
b_2	0	0
b_3	0	0.5

$\lambda_A(a_2)$

	l	r
b_1	0	0
b_2	0	0.5
b_3	0	0.5

$\lambda_A(a_3)$

	l	r
a_1	0.5	0
a_2	0	0.5
a_3	0	0

$\lambda_B(b_1)$

	l	r
a_1	0.5	0
a_2	0	0
a_3	0	0.5

$\lambda_B(b_2)$

	l	r
a_1	0	0
a_2	0	0.5
a_3	0	0.5

$\lambda_B(b_3)$

Figure 21

Consider the state (d,r,a_3,b_3). Both a_3 and b_3 correctly believe (i.e., assign probability 1 to) that the outcome is (d,r) (we have $\lambda_A(a_3)(r) = \lambda_B(b_3)(d) = 1$). This fact is not common knowledge: a_3 assigns a 0.5 probability to Bob being of type b_2, and type b_2 assigns a 0.5 probability to Ann playing l. Thus, Ann does not know that Bob knows that she is playing r (here, "knowledge" is identified with "probability 1" as it is in Aumann & Brandenburger 1995). Furthermore, while it is true that both Ann and Bob are rational, it is not common knowledge that they are rational. Indeed, the type a_3 assigns a 0.5 probability to Bob being of type b_2 and choosing r; however, this is irrational since b_2 believes that both of Ann's options are equally probable.

The example above is a situation where there is mutual knowledge of the choices of the players. Indeed, it is not hard to see that in any type space for a 2-player game G, if (s,t) is a state where there is mutual knowledge that player i is choosing s_i and the players are rational, then, s constitutes a (pure-strategy) Nash Equilibrium. There is a more general theorem concerning mixed strategy equilibrium. Recall that a conjecture for player i is a probability measure over the strategy choices of her opponents.

> **Theorem 5.1 (Aumann & Brandenburger 1995: Theorem A)** Suppose that G is a 2-person strategic game, (p_1,p_2) are conjectures for players 1 and 2, and $\mathcal{T}$ is a type space for G. If (s,t) is a state in $\mathcal{T}$ where for $i = 1,2$, t_i assigns probability 1 to the events (a) both players are rational (i.e., maximize expected utility), (b) the game is G and (c) for $i = 1,2$, player i's conjecture is p_i, then (p_1,p_2) constitutes a Nash equilibrium.

The general version of this result, for arbitrary finite number of agents and allowing for mixed strategies, requires *common knowledge* of *conjectures*, i.e., of each player's probabilistic beliefs in the other's choices. See Aumann & Brandenburger (1995: Theorem B) for precise formulation of the result, and, again, Spohn (1982) for an early version. See, also, Perea (2007) and Tan & Werlang (1988) for similar results about the Nash equilibrium.

5.1.2 Philosophical Issues

This epistemic characterization of Nash equilibrium requires mutual *knowledge* and rather than beliefs. The result fails when agents can be mistaken about the strategy choice of the others. This has lead some authors to criticize this epistemic characterization: See Gintis (2009) and Bruin (2010), for instance. How could the players ever *know* what the others are choosing? Is it not contrary to the very idea of a game, where the players are free to choose whatever they want (Baltag et al. 2009)?

One popular response to this criticism (Brandenburger 2010; Perea 2012) is that the above result tells us something about Nash equilibrium *as a solution concept*, namely that *it alleviates strategic uncertainty*. Indeed, returning to the terminology introduced in Section 1.3, the epistemic conditions for Nash equilibrium are those that correspond to the *ex post* state of information disclosure, "when all is said and done", to put it figuratively. When players have reached full knowledge of what the others are going to do, there is nothing left to think about regarding the other players as rational, deliberating agents. The consequences of each of the players' actions are now certain. The only task that remains is to compute which action is recommended by the adopted choice rule, and this does not involve any specific information about the other players' beliefs. Their choices are fixed, after all.

The idea here is not to reject the epistemic characterization of Nash Equilibrium on the grounds that it rests on unrealistic assumptions, but, rather, to view it as a lesson learned about Nash Equilibrium itself. From an epistemic point of view, where one is focused on *strategic reasoning* about what others are going to do and are thinking, this solution concepts might be of less interest.

There is another important lesson to draw from this epistemic characterization result. The widespread idea that game theory "assumes common knowledge of rationality", perhaps in conjunction with the extensive use of equilibrium concepts in game-theoretic analysis, has lead to misconception that the Nash Equilibrium either *requires* common knowledge of rationality, or that common knowledge of rationality is sufficient for the players to play according to a Nash equilibrium. To be sure, game theoretic models do assume that the structure of the game is common knowledge (though, see Section 5.3). Nonetheless, the above result shows that both of these ideas are incorrect:

> Common knowledge of rationality is neither necessary nor sufficient for Nash Equilibrium.

In fact, as we just stressed, Nash equilibrium can be played under full uncertainty, and *a fortiori* under higher-order uncertainty, about the rationality of others.

5.1.3 Remarks on "Modal" Characterizations of Nash Equilibrium

In recent years, a number of so-called "modal" characterizations of Nash Equilibrium have been proposed, mostly using techniques from modal logic (see Hoek & Pauly 2007 for details). These results typically devise a modal logical language to describe games in strategic form, typically including modalities for the players' actions and preference, and show that the notion of profile being a Nash Equilibrium language is *definable* in such a language.

Most of these characterizations are not epistemic, and thus fall outside the scope of this entry. In context of this entry, it is important to note that most of these results aim at something different than the epistemic characterization which we are discussing in this section. Mostly developed in Computer Sciences, these logical languages have been used to verify properties of multi-agents systems, not to provide epistemic foundations to this solution concept. However, note that in recent years, a number of logical characterizations of Nash equilibrium do explicitly use epistemic concepts (see, for example, van Benthem et al. 2009; Lorini & Schwarzentruber 2010).

5.2 Incorporating Admissibility and "Cautious" Beliefs

It is not hard to find a game and an informational context where there is at least one player without a *unique* "rational choice". How should a rational player incorporate the information that more than one action is classified as "choice-worthy" or "rationally permissible" (according to some choice rule) for her opponent(s)? In such a situation, it is natural to require that the player does not *rule out* the possibility that her opponent will pick a "choice-worthy" option. More generally, the players should be "cautious" about which of their opponents' options they *rule out*.

Assuming that the players' beliefs are "cautious" is naturally related to weak dominance (recall the characterization of weak dominance, Section

3.2 in which a strategy is weakly dominated iff it does not maximize expected utility with respect to any *full support* probability measure). A key issue in epistemic game theory is the epistemic analysis of iterated removal of weakly dominated strategies. Many authors have pointed out puzzles surrounding such an analysis (Asheim & Dufwenberg 2003; Brandenburger, Friedenberg & Keisler 2008; Cubitt & Sugden 1994; Samuelson 1992). For example, Samuelson (1992) showed (among other things) that the analogue of Theorem 4.1 is not true for iterated removal of weakly dominated strategies. The main problem is illustrated by the following game:

		Bob	
		l	r
Ann	u	1,1	1,0
	d	1,0	0,1

Figure 22

In the above game, d is weakly dominated by u for Ann. If Bob knows that Ann is rational (in the sense that she will not choose a weakly dominated strategy), then he can rule out option d. In the smaller game, action r is now strictly dominated by l for Bob. If Ann knows that Bob is rational and that Bob knows that she is rational (and so, rules out option d), then she can rule out option r. Assuming that the above reasoning is transparent to both Ann and Bob, it is common knowledge that Ann will play u and Bob will play l. But now, what is the reason for Bob to rule out the possibility that Ann will play d? He knows that Ann knows that he is going to play l and both u and d are best responses to l. The problem is that assuming that the players' beliefs are cautious conflicts with the logic of iterated removal of weakly dominated strategies. This issue is nicely described in a well-known microeconomics textbook:

> [T]he argument for deletion of a weakly dominated strategy for player i is that he contemplates the possibility that every strategy combination of his rivals occurs with positive probability. However, this hypothesis clashes with the logic of iterated deletion, which assumes, precisely that eliminated strategies are not expected to occur. (Mas-Colell, Winston, & Green 1995: 240)

The extent of this conflict is nicely illustrated in Samuelson (1992). In particular, Samuelson (1992) shows that there is no epistemic-probability model[14] of the above game with a state satisfying common knowledge of *rationality* (where "rationality" means that players do not choose weakly dominated strategies). *Prima facie*, this is puzzling: What about the epistemic-probability model consisting of a single state w assigned the profile (u, l)? Isn't this a model of the above game where there is a state satisfying common knowledge that the players do not choose weakly dominated strategies? The problem is that the players do not have "cautious" beliefs in this model (in particular, Bob's beliefs are not cautious in the sense described below). Recall that having a cautious belief means that a player cannot *know* which options her opponent(s) will *pick*[15] from a set of choice-worthy options (in the above game, if Ann *knows* that Bob is choosing l, then both u and d are "choice-worthy", so Bob cannot *know* that Ann is choosing u). This suggests an additional requirement on a game model: Let $\mathcal{M} = \langle W, \{\Pi_i\}_{i\in N}, \{p_i\}_{i\in N}, \sigma\rangle$ be an epistemic-probability model. For each action $a \in \cup_{i\in N} S_i$, let $[\![a]\!] = \{w \mid (\sigma(w))_i = a\}$.

> If $a \in S_i$ is **rational** for player i at state w, then for all players $j \neq i$, $[\![a]\!] \cap \Pi_j(w) \neq \emptyset$.

This means that a player cannot *know* that her opponent will not choose an action at a state w which is deemed rational (according to some choice rule). This property is called "privacy of tie-breaking" by Cubitt and

Sugden (2011: 8) and "no extraneous beliefs" by Asheim and Dufwenberg (2003).[16] For an extended discussion of the above assumption see Cubitt & Sugden (2011).

Given the above considerations, the epistemic analysis of iterated weak dominance is not a straightforward adaptation of the analysis of iterated strict dominance discussed in the previous section. In particular, any such analysis must resolve the conflict between strategic reasoning where players *rule out* certain strategy choices of their opponent(s) and admissibility considerations where players must consider all of their opponents' options *possible*. A number of authors have developed frameworks that do resolve this conflict (Brandenburger et al. 2008; Asheim & Dufwenberg 2003; Halpern & Pass 2009). We sketch one of these solutions below:

The key idea is to represent the players' beliefs as a *lexicographic probability system* (LPS). An LPS is a finite sequence of probability measures $(p_1, p_2, \dots, p_n)$ with supports (The **support** of a probability measure p defined on a set of states W is the set of all states that have nonzero probability; formally, $Supp(p) = \{w \mid p(w) > 0\}$) that do not overlap. This is interpreted as follows: if $(p_1, \dots, p_n)$ represents Ann's beliefs, then p_1 is Ann's "initial hypothesis" about what Bob is going to do, p_2 is Ann's secondary hypothesis, and so on. In the above game, we can describe Bob's beliefs as follows: his initial hypothesis is that Ann will choose U with probability 1 and his secondary hypothesis is that she will choose D with probability 1. The interpretation is that, although Bob does not rule out the possibility that Ann will choose D (i.e., choose irrationally), he does consider it *infinitely less likely* than her choosing U (i.e., choosing rationally).

So, representing beliefs as lexicographic probability measures resolves the conflict between strategic reasoning and the assumption that players do not

play weakly dominated strategies. However, there is another, more fundamental, issue that arises in the epistemic analysis of iterated weak dominance:

> Under admissibility, Ann considers everything possible. But this is only a decision-theoretic statement. Ann is in a game, so we imagine she asks herself: "What about Bob? What does he consider possible?" If Ann truly considers everything possible, then it seems she should, in particular, allow for the possibility that Bob does not! Alternatively put, it seems that a full analysis of the admissibility requirement should include the idea that other players do not conform to the requirement. (Brandenburger et al. 2008: 313)

There are two main ingredients to the epistemic characterization of iterated weak dominance. The first is to represent the players' beliefs as lexicographic probability systems. The second is to use a stronger notion of belief: A player **assumes** an event E provided E is infinitely more likely than $\overline{E}$ (on finite spaces, this means each state in E is infinitely more likely than states not in E). The key question is: What is the precise relationship between the event "rationality and common assumption of rationality" and the strategies that survive iterated removal of weakly dominated strategies? The precise answer turns out to be surprisingly subtle—the details are beyond the scope of this article (see Brandenburger et al. 2008).

5.3 Incorporating Unawareness

The game models introduced in Section 2 have been used to describe the uncertainty that the players have about what their opponents are going to do and are thinking in a game situation. In the analyses provided thus far, the *structure* of the game (i.e., who is playing, what are the preferences of the different players, and which actions are available) is assumed to be common knowledge among the players. However, there are many situations where the players do not have such *complete* information about the game. There is no inherent difficulty in using the models from Section 2 to describe situations where players are not *perfectly informed* about the structure of the game (for example, where there is some uncertainty about available actions).

There is, however, a foundational issue that arises here. Suppose that Ann considers it *impossible* that her opponent will choose action a. Now, there are many reasons why Ann would hold such an opinion. On the one hand, Ann may know something about what her opponent is going to do or is thinking which allows her to rule out action a as a live possibility—i.e., given all the evidence Ann has about her opponent, she concludes that action a is just not something her opponent will do. On the other hand, Ann may not even conceive of the possibility that her opponent will choose action a. She may have a completely different model of the game in mind than her opponents. The foundational question is: Can the game models introduced in Section 2 faithfully represent this latter type of uncertainty?

The question is not whether one can formally describe what Ann knows and believes under the assumption that she considers it impossible that her opponent will choose action a. Indeed, an epistemic-probability model where Ann assigns probability zero to the event that her opponent chooses action a is a perfectly good description of Ann's epistemic state. The problem is that this model blurs an important distinction between Ann being *unaware* that action a is a live possibility and Ann *ruling out* that action a is a viable option for her opponent. This distinction is illustrated by the following snippet from the well-known Sherlock Holmes' short story *Silver Blaze* (Doyle 1894):

> …I saw by the inspector's face that his attention had been keenly aroused.
> "You consider that to be important?" he [Inspector Gregory] asked.
> "Exceedingly so."
> "Is there any point to which you would wish to draw my attention?"
> "To the curious incident of the dog in the night-time."
> "The dog did nothing in the night-time."
> "That was the curious incident," remarked Sherlock Holmes.

The point is that Holmes is aware of a particular event ("the dog not barking") and uses this to come to a conclusion. The inspector is not aware of this event, and so cannot (without Holmes' help) come to the same conclusion. This is true of many detective stories: clever detectives not only have the ability to "connect the dots", but they are also *aware* of which dots need to be connected. Can we describe the inspector's unawareness in an epistemic model?[17]

Suppose that $U_i(E)$ is the event that the player i is unaware of the event E. Of course, if i is unaware of E then i does not know that E is true ($U_i(E) \subseteq \overline{K_i(E)}$, where $\overline{X}$ denotes the complement of the event X). Recall that in epistemic models (where the players' information is described by partitions), we have the negative introspection property:

$$\overline{K_i(E)} \subseteq K_i(\overline{K_i(E)}).$$

This means that if i is unaware of E, then i knows that she does not know that E. Thus, to capture a more natural definition of $U_i(E)$ where

$$U_i(E) \subseteq \overline{K_i(E)} \cap \overline{K_i(\overline{K_i(E)})},$$

we need to represent the players' knowledge in a *possibility structure* where the K_i operators do not necessarily satisfy negative introspection. A possibility structure is a tuple $\langle W, \{P_i\}_{i\in N}, \sigma\rangle$ where $P_i : W \rightarrow \wp(W)$. The only difference with an epistemic model is that the $P_i(w)$ do not necessarily form a partition of W. We do not go into details here—see Halpern (1999) for a complete discussion of possibility structures and how they relate to epistemic models. The knowledge operator is defined as it is for epistemic models: for each event E, $K_i(E) = \{w \mid P_i(w) \subseteq E\}$. However, S. Modica and A. Rustichini (1994, 1999) argue that even the more general possibility structures cannot be used to describe a player's unawareness.

A natural definition of unawareness on possibility structures is:

$$U(E) = \overline{K(E)} \cap \overline{K(\overline{K(E)})} \cap \overline{K(\overline{K(\overline{K(E)})})} \cap \cdots$$

That is, an agent is unaware of E provided the agent does not know that E obtains, does not know that she does not know that E obtains, and so on. Modica and Rustichini use a variant of the above Sherlock Holmes story to show that there is a problem with this definition of unawareness.

Suppose there are two signals: A dog barking (d) and a cat howling (c). Furthermore, suppose there are three states w_1, w_2 in which the dog barks and w_3 in which the cat howls. The event that there is no intruder is $E = \{w_1\}$ (the lack of the two signals indicates that there was no intruder[18]). The following possibility structure (where there is an arrow from state w to state v provided $v \in P(w)$) describes the inspector's epistemic state:

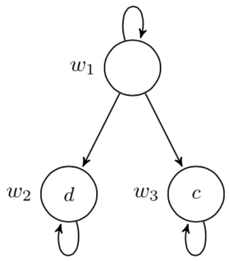

Figure 23

Consider the following calculations:

- $K(E) = \{w_2\}$ (at w_2, Watson knows there is a human intruder) and $-K(E) = \{w_1, w_3\}$
- $K(-K(E)) = \{w_3\}$ (at w_3, Watson knows that she does not know E), and $-K(-K(E)) = \{w_1, w_2\}$.
- $-K(E) \cap -K(-K(E)) = \{w_1\}$ and, in fact, $\bigcap_{i=1}^{\infty}(-K)^i(E) = \{w_1\}$
- Let $U(F) = \bigcap_{i=1}^{\infty}(-K)^i(F)$. Then,
    - $U(\emptyset) = U(W) = U(\{w_1\}) = U(\{w_2, w_3\}) = \emptyset$
    - $U(E) = U(\{w_3\}) = U(\{w_1, w_3\}) = U(\{w_1, w_2\} = \{w_1\}$

So, $U(E) = \{w_1\}$ and $U(U(E)) = U(\{w_1\}) = \emptyset$. This means that at state w_1, the Inspector is unaware of E, but is not unaware that he is unaware of E. More generally, Dekel et al. (1998) show that there is no nontrivial unawareness operator U satisfying the following properties:

- $U(E) \subseteq \overline{K(E)} \cap \overline{K(E)}$
- $K(U(E)) = \emptyset$

- $U(E) \subseteq U(U(E))$

There is an extensive literature devoted to developing models that can represent the players' unawareness. See Board, Chung, & Schipper (2011); Chen, Ely, & Luo (2012); E. Dekel et al. (1998); Halpern (2001a); Halpern & Rego (2008); and Heifetz, Meier, & Schipper (2006) for a discussion of issues related to this entry. The Unawareness Bibliography (see Other Internet Resources) has an up-to-date list of papers in this area.

6. A Paradox of Self-Reference in Game Models

The first step in any epistemic analysis of a game is to describe the players' knowledge and beliefs using (a possible variant of) one of the models introduced in Section 2. As we noted already in Section 2.2, there will be statements about what the players know and believe about the game situation and about each other that are commonly known in some models but not in others.

> In any particular structure, certain beliefs, beliefs about belief, …, will be present and others won't be. So, there is an important implicit assumption behind the choice of a structure. This is that it is "transparent" to the players that the beliefs in the type structure —and only those beliefs—are possible ….The idea is that there is a "context" to the strategic situation (e.g., history, conventions, etc.) and this "context" causes the players to rule out certain beliefs. (Brandenburger & Friedenberg 2010: 801)

Ruling out certain configurations of beliefs constitute *substantive assumptions* about the players' reasoning during the decision making process. In other words, substantive assumptions are about how, and how much, information is imparted to the agents, over and above those that are intrinsic to the mathematical formulation of the structures used to describe

the players' information. It is not hard to see that one always finds substantive assumptions in finite structures: Given a countably infinite set of atomic propositions, for instance, in finite structures it will always be common knowledge that some logically consistent combination of these basic facts are not realized, and *a fortiori* for logically consistent configurations of information and higher-order information about these basic facts. On the other hand, monotonicity of the belief/knowledge operator is a typical example of an assumption that is *not* substantive. More generally, there are no models of games, as we defined in Section 2, where it is not common knowledge that the players believe all the logical consequences of their beliefs.[19]

Can we compare models in terms of the number of substantive assumptions that are made? Are there models that make no, or at least as few as possible, substantive assumptions? These questions have been extensively discussed in the epistemic foundations of game theory—see the discussion in Samuelson (1992) and the references in Moscati (2009). Intuitively, a structure without any substantive assumptions must represent all possible states of (higher-order) information. Whether such a structure exists will depend, in part, on how the players' informational attitudes are represented—e.g., as (conditional/lexicographic) probability measures or set-valued knowledge/belief functions. These questions have triggered interest in the existence of "rich" models containing most, if not all, possible configurations of (higher-order) knowledge and beliefs.

There are different ways to understand what it means for a structure to minimize the substantive assumptions about the players' higher-order information. We do not attempt a complete overview of this interesting literature here (see Brandenburger & Keisler (2006: sec. 11) and Siniscalchi (2008: sec. 3) for discussion and pointers to the relevant results). One approach considers the space of all (Harsanyi type-/Kripke-/epistemic-plausibility-) structures and tries to find a single structure that, in some suitable sense, “contains” all other structures. Such a structure, often called called a *universal structure* (or a *terminal object* in the language of category theory), if it exists, incorporates any substantive assumption that an analyst can imagine. Such structure have been shown to exists for Harsanyi type spaces (Mertens & Zamir 1985; Brandenburger & Dekel 1993). For Kripke structures, the question has been answered in the negative (Heifetz & Samet 1998; Fagin, Geanakoplos, Halpern, & Vardi 1999; Meier 2005), with some qualifications regarding the language that is used to describe them (Heifetz 1999; Roy & Pacuit 2013).

A second approach takes an internal perspective by asking whether, *for a fixed set of states or types*, the agents are making any substantive assumptions about what their opponents know or believe. The idea is to identify (in a given model) a set of possible *conjectures* about the players. For example, in a knowledge structure based on a set of states W this might be the set of all subsets of W or the set definable subsets of W in some suitable logical language. A space is said to be *complete* if each agent correctly takes into account each possible conjecture about her opponents. A simple counting argument shows that there cannot exist a complete structure when the set of conjectures is *all* subsets of the set of states (Brandenburger 2003). However, there is a deeper result here which we discuss below.

The Brandenburger-Keisler Paradox

Adam Brandenburger and H. Jerome Keisler (2006) introduce the following two person, Russel-style paradox. The statement of the paradox involves two concepts: beliefs and assumptions. An *assumption* is a player’s strongest belief: it is a set of states that implies all other beliefs at a given state. We will say more about the interpretation of an assumption below. Suppose there are two players, Ann and Bob, and consider the following description of beliefs.

(S) Ann believes that Bob assumes that Ann believes that Bob’s assumption is wrong.

A paradox arises by asking the question

(Q) Does Ann believe that Bob’s assumption is wrong?

To ease the discussion, let C be Bob’s assumption in (S): that is, C is the statement “Ann believes that Bob’s assumption is wrong.” So, (Q) asks whether C is true or false. We will argue that C is true if, and only if, C is false.

Suppose that C is true. Then, Ann does believe that Bob’s assumption is wrong, and, by introspection, she believes that she believes this. That is to say, Ann believes that C is correct. Furthermore, according to (S), Ann believes that Bob’s assumption is C. So, Ann, in fact, believes that Bob’s assumption is correct (she believes Bob’s assumption is C and that C is correct). So, C is false.

Suppose that C is false. This means that Ann believes that Bob’s assumption is correct. That is, Ann believes that C is correct (By (S), Ann believes that Bob’s assumption is C). Furthermore, by (S), we have that *Ann believes that Bob assumes that Ann believes that C is wrong*. So, Ann believes that she believes that C is correct and she believes that Bob assumption is that she believes that C is wrong. So, it is true that she believes Bob’s assumptions is wrong (Ann believes that Bob’s assumption is *she believes that C is wrong*, but she believes that is wrong: *she believes that C is correct*). So, C is true.

Brandenburger and Keisler formalize the above argument in order to prove a very strong impossibility result about the existence of so-called *assumption-complete* structures. We need some notation to state this result. It will be most convenient to work in qualitative type spaces for two players (Definition 2.7). A qualitative type space for two players (cf. Definition 2.7. The set of states is not important in what follows, so we leave it out) is a structure $\langle \{T_A, T_B\}, \{\lambda_A, \lambda_B\} \rangle$ where

$$\lambda_A : T_A \to \wp(T_B) \qquad \lambda_B : T_B \to \wp(T_A)$$

A set of **conjectures about Ann** is a subset $\mathcal{C}_A \subseteq \wp(T_A)$ (similarly, the set of conjectures about Bob is a subset $\mathcal{C}_B \subseteq \wp(T_B)$). A structure $\langle \{T_A, T_B\}, \{\lambda_A, \lambda_B\} \rangle$ is said to be **assumption-complete** for the conjectures $\mathcal{C}_A$ and $\mathcal{C}_B$ provided for each conjecture in $\mathcal{C}_A$ there is a type that assumes that conjecture (similarly for Bob). Formally, for each $Y \in \mathcal{C}_B$ there is a $t_0 \in T_A$ such that $\lambda_A(t_0) = Y$, and similarly for Bob. As we remarked above, a simple counting argument shows that when $\mathcal{C}_A = \wp(T_A)$ and $\mathcal{C}_B = \wp(T_B)$, then assumption-complete models only exist in trivial cases. A much deeper result is:

> **Theorem 6.1 (Brandenburger & Keisler 2006: Theorem 5.4)** There is no assumption-complete type structure for the set of conjectures that contains the first-order definable subsets.

See the supplement for a discussion of the proof of this theorem (see Section 2).

Consult Pacuit (2007) and Abramsky & Zvesper (2010) for an extensive analysis and generalization of this result. But, it is not all bad news: Mariotti, Meier, & Piccione (2005) construct a complete structure where the set of conjectures are compact subsets of some well-behaved topological space.

7. Concluding Remarks

The epistemic view on games is that players should be seen as individual decision makers, choosing what to do on the basis of their own preferences and the information they have in specific informational contexts. What decision they will make—the descriptive question—or what decision they should make—the normative question, depends on the decision-theoretic choice rule that the player use, or should use, in a given context. We conclude with two general methodological issues about epistemic game theory and some pointers to further reading.

7.1 What is an epistemic game theory trying to accomplish?

Common knowledge of rationality is an informal assumption that game theorists, philosophers and other social scientists often appeal to when analyzing social interactive situations. The epistemic program in game theory demonstrates that there are many ways to understand what exactly it means to assume that there is “common knowledge/belief of rationality” in a game situation.

Broadly speaking, much of the epistemic game theory literature is focused on two types of projects. The goal of the first project is to map out the relationship between different mathematical representations of what the players know and believe about each other in a game situation. Research along these lines not only raises interesting technical questions about how to compare and contrast different mathematical models of the players' epistemic states, but it also highlights the benefits and limits of an epistemic analysis of games. The second project addresses the nature of rational choice in game situations. The importance of this project is nicely explained by Wolfgang Spohn:

> …game theory…is, to put it strongly, confused about the rationality concept appropriate to it, its assumptions about its subjects (the players) are very unclear, and, as a consequence, it is unclear about the decision rules to be applied.…The basic difficulty in defining rational behavior in game situations is the fact

> that in general each player's strategy will depend on his expectations about the other players' strategies. Could we assume that his expectations were given, then his problem of strategy choice would become an ordinary maximization problem: he could simply choose a strategy maximizing his own payoff on the assumption that the other players would act in accordance with his given expectations. But the point is that game theory cannot regard the players' expectations about each other's behavior as given; rather, one of the most important problems for game theory is precisely to decide what expectations intelligent players can rationally entertain about other intelligent players' behavior. (Spohn 1982: 267)

Much of the work in epistemic game theory can be viewed as an attempt to use precise representations of the players' knowledge and beliefs to help resolve some of the confusion alluded to in the above quote.

7.2 Alternatives to maximizing expected utility

In an epistemic analysis of a game, the specific recommendations or predictions for the players' choices are derived from decision-theoretic choice rules. Maximization of expected utility, for instance, underlies most of the results in the contemporary literature on the epistemic foundations of game theory. From a methodological perspective, however, the choice rule that the modeler assumes the players are following is simply a parameter that can be varied. In recent years, there have been some initial attempts to develop epistemic analyses with alternative choice rules, for instance *minregret* Halpern & Pass (2009).

7.3 Further reading

The reader interested in more extensive coverage of all or some of the topics discussed in this entry should consult the following articles and books.

Logic in Games by Johan van Benthem: This book uses the tools of modal logic broadly conceived to discuss many of the issues raised in this entry (2014, MIT Press).

The Language of Game Theory by Adam Brandenburger: A collection of Brandenburger's key papers on epistemic game theory (2014, World Scientific Series in Economic Theory).

Epistemic Game Theory by Eddie Dekel and Marciano Siniscalchi: A survey paper aimed at economists covering the main technical results of epistemic game theory (2014, Available online).

Epistemic Game Theory: Reasoning and Choice by Andrés Perea: A non-technical introduction to epistemic game theory (2012, Cambridge University Press).

The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences by Herbert Gintis: This book offers a broad overview of the social and behavioral science using the ideas of epistemic game theory (2009, Princeton University Press).

Bibliography

Abramsky, S. & J.A. Zvesper, 2012, "From Lawvere to Brandenburger-Keisler: interactive forms of diagonalization and self-reference", in *Coalgebraic Methods in Computer Science* (LNCS, Vol. 7399, pp. 1–19), *CoRR*, *abs/1006.0992*.

Alchourrón, C.E., P. Gärdenfors, & D. Makinson, 1985, "On the logic of theory change: Partial meet contraction and revision functions", *Journal of Symbolic Logic*, 50(2): 510–530.

Apt, K. & J. Zvesper, 2010, "The role of monotonicity in the epistemic analysis of strategic games", *Games*, 1(4): 381–394, doi:10.3390/g1040381

Asheim, G. & M. Dufwenberg, 2003, "Admissibility and common belief", *Game and Economic Behavior*, 42: 208–234.

Aumann, R., 1976, "Agreeing to disagree", *The Annals of Statistics*, 4(6): 1236–1239.

–––, 1987, "Correlated equilibrium as an expression of Bayesian rationality", *Econometrica*, 55(1): 1–18.

–––, 1995, "Backward induction and common knowledge of rationality", *Games and Economic Behavior*, 8(1): 6–19.

–––, 1999a, "Interactive epistemology I: Knowledge", *International Journal of Game Theory*, 28(3): 263–300.

–––, 1999b, "Interactive epistemology II: Probability", *International Journal of Game Theory*, 28(3): 301–314.

–––, 2010, "Interview on epistemic logic", in V. F. Hendricks & O. Roy (Eds.), *Epistemic logic: Five questions* (pp. 21–35). Automatic Press.

Aumann, R. J., S. Hart, & M. Perry, 1997, "The absent-minded driver", *Games and Economic Behavior*, 20(1): 102–116.

Aumann, R. & A. Brandenburger, 1995, "Epistemic conditions for Nash equilibrium", *Econometrica*, 63(5): 1161–1180.

Aumann, R. & S. Hart, 1994, *Handbook of game theory with economic applications* (Vol. 2), Amsterdam: North Holland.

Baltag, A. & S. Smets, 2006, "Conditional doxastic models: A qualitative approach to dynamic belief revision", in *Electronic notes in theoretical computer science* (Vol. 165, pp. 5–21), Springer.

Baltag, A., S. Smets, & J. Zvesper, 2009, "Keep 'hoping' for rationality: a solution to the backwards induction paradox", *Synthese*, 169: 301–333.

Battigalli, P., 1997, "On rationalizability in extensive games", *Journal of Economic Theory*, 74(1): 40–61.

Battigalli, P. & A. Friedenberg, 2012, "Forward induction reasoning revisited", *Theoretical Economics*, 7(1): 57–98.

Battigalli, P. & M. Siniscalchi, 2002, "Strong belief and forward induction reasoning", *Journal of Economic Theory*, 106(2): 356–391.

Battigalli, P., A. Di Tillio, & D. Samet, 2013, "Strategies and interactive beliefs in dynamic games", in *Advances in economics and econometrics: Theory and applications, Tenth World Congress, volume 1: economic theory*, Cambridge: Cambridge University Press.

van Benthem, J., 2003, "Rational dynamic and epistemic logic in games", in S. Vannucci (Ed.), *Logic, game theory and social choice III*, University of Siena, Department of Political Economy.

–––, 2010, *Modal logic for open minds*, Stanford, CA: CSLI Publications.

–––, 2011, *Logical dynamics of information and interaction*, Cambridge: Cambridge University Press.

van Benthem, J. & A. Gheerbrant, 2010, "Game solution, epistemic dynamics and fixed-point logics", *Fundamenta Informaticae*, 100: 1–23.

van Benthem, J., P. Girard, & O. Roy, 2009, "Everything else being equal: A modal logic for *Ceteris Paribus* preferences", *Journal of Philosophical Logic*, 38: 83–125.

van Benthem, J., E. Pacuit, & O. Roy, 2011, "Toward a theory of play: A logical perspective on games and interaction", *Games*, 2(1): 52–86.

Bernheim, D., 1984, "Rationalizable strategic behavior", *Econometrica*, 52: 1007–1028.

Board, O., 2003, "The not-so-absent-minded driver", *Research in Economics*, 57(3): 189–200.

Board, O., K.S. Chung, & B. Schipper, 2011, "Two models of unawareness: Comparing object-based and subjective-state-space approaches", *Synthese*, 179: 13–34.

Bonanno, G., 1996, "On the logic of common belief", *Mathematical*

Logical Quarterly, 42: 305–311.

–––, 2004, "Memory and perfect recall in extensive games", *Games and Economic Behavior*, 47(2): 237–256.

–––, 2013, "A dynamic epistemic characterization of backward induction without counterfactuals", *Games and Economic Behavior*, 78: 31–43.

Brandenburger, A., 2003, "On the existence of a "complete" possibility structure", in M. Basili, N. Dimitri, & I. Gilboa (Eds.), *in Cognitive processes and economic behavior* (pp. 30–34). Routledge.

–––, 2007, "A note on Kuhn's theorem", in J. van Benthem, D. Gabbay, & B. Loewe (Eds.), *Interactive logic, proceedings of the 7th Augustus de Morgan workshop, London* (pp. 71–88). Texts in Logic; Games, Amsterdam University Press.

–––, 2010, "Origins of epistemic game theory", in V. F. Hendricks & O. Roy (Eds.), *Epistemic logic: Five questions* (pp. 59–69). Automatic Press.

Brandenburger, A. & E. Dekel, 1987, "Rationalizability and correlated equilibria", *Econometrica*, 55(6): 1391–1402.

–––, 1993, "Hierarchies of beliefs and common knowledge", *Journal of Economic Theory*, 59.

Brandenburger, A. & A. Friedenberg, 2008, "Intrinsic correlation in games", *Journal of Economic Theory*, 141(1): 28–67.

–––, 2010, "Self-admissible sets", *Journal of Economic Theory*, 145: 785–811.

Brandenburger, A. & H. Keisler, 2006, "An impossibility theorem on beliefs in games", *Studia Logica*, 84(2): 211–240.

Brandenburger, A., A. Friedenberg, & H.J. Keisler, 2008, "Admissibility in games", *Econometrica*, 76(2): 307–352.

de Bruin, B., 2010, *Explaining games : The epistemic programme in game theory*, New York City: Springer.

Chen, Y.C., J. Ely, & X. Luo, 2012, "Note on unawareness: Negative introspection versus AU introspection (and KU introspection)",

International Journal of Game Theory, 41(2): 325 - 329.

Colman, A., 2003, "Cooperation, psychological game theory, and limitations of rationality in social interactions", *Behavioral and Brain Sciences*, 26: 139–198.

Cubitt, R.P. & R. Sugden, 1994, "Rationally justifiable play and the theory of non-cooperative games", *The Economic Journal*, 104(425): 798–893.

–––, 2014, " Common reasoning in games: A Lewisian analysis of common knowledge of rationality", *Economics and Philosophy*, 30(03): 285–329.

Dekel, E., B. Lipman, & A. Rustichini, 1998, "Standard state-space models preclude unawareness", *Econometrica*, 66: 159–173.

Doyle, A.C., 1894, *The Memoirs of Sherlock Holmes*, Mineola, NY: Dover Thrift Edition, 2010.

Fagin, R., J. Geanakoplos, J. Halpern, & M. Vardi, 1999, "The hierarchical approach to modeling knowledge and common knowledge", *International Journal of Game Theory*, 28(3): 331–365.

Fagin, R., J. Halpern, & N. Megiddo, 1990, "A logic for reasoning about probabilities", *Information and Computation*, 87(1–2): 78–128.

Fagin, R., J. Halpern, Y. Moses, & M. Vardi, 1995, *Reasoning about knowledge*, Cambridge: The MIT Press.

Finetti, B., 1974, *Theory of probability, vols. 1 and 2*, New York: Wiley.

Friedenberg, A. & H.J. Keisler, 2011, "Iterated dominance revisited", in *Proceedings of the behavioral and quantitative game theory: Conference on future directions*, ACM, New York, NY. [available online].

Friedenberg, A. & M. Meier, 2009, "The context of a game", in *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 134–135 [available online].

Gintis, H., 2009, *The bounds of reason: game theory and the unification of the behavioral sciences*, Princeton: Princeton University Press.

Halpern, J.Y., 1991, "The relationship between knowledge, belief, and certainty", *Annals of Mathematics and Artificial Intelligence*, 4(3): 301–322. [available online].

—–, 1997, "On ambiguities in the interpretation of game trees", *Games and Economic Behavior*, 20(1): 66–96.

—–, 1999, "Set-theoretic completeness for epistemic and conditional logic", *Annals of Mathematics and Artificial Intelligence*, 26: 1–27.

—–, 2001a, "Alternative semantics for unawareness", *Game and Economic Behavior*, 37: 321–339.

—–, 2001b, "Substantive rationality and backward induction", *Games and Economic Behavior*, 37(2): 425–435.

—–, 2003, *Reasoning about uncertainty*, Cambridge: The MIT Press.

—–, 2010, "Lexiographic probability, conditional probability and nonstandard probability", *Games and Economic Behavior*, 68(1): 155–179.

Halpern, J.Y. & R. Pass, 2009, "A logical characterization of iterated admissibility", in A. Heifetz (Ed.), *Proceedings of the twelfth conference on theoretical aspects of rationality and knowledge* (pp. 146–155).

—–, 2011, "Iterated regret minimization: A new solution concept", *Games and Economic Behavior*, 74(1): 184–207 [available online].

Halpern, J.Y. & L.C. Rego, 2008, "Interactive unawareness revisited", *Games and Economic Behavior*, 62(1): 232–262.

Harsanyi, J.C., 1967–68, "Games with incomplete information played by 'Bayesian' players, parts I–III", *Management Science*, 14: 159–182; 14: 320–334; 14: 486–502.

Heifetz, A., 1999, "How canonical is the canonical model? A comment on Aumann's interactive epistemology", *International Journal of Game Theory*, 28(3): 435–442.

Heifetz, A. & P. Mongin, 2001, "Probability Logic for Type Spaces", *Games and Economic Behavior*, 35(1–2): 31–53.

Heifetz, A. & D. Samet, 1998, "Knowledge spaces with arbitrarily high rank", *Games and Economic Behavior*, 22(2): 260–273.

Heifetz, A., M. Meier, & B. Schipper, 2006, "Interactive unawareness", *Journal of Economic Theory*, 130: 78–94.

Hendricks, V. & J. Symons, 2009, "Epistemic logic", in E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2009 Edition), URL = <https://plato.stanford.edu/archives/spr2009/entries/logic-epistemic/>.

Hoek, W. van der & M. Pauly, 2007, "Modal logic for games and information", in P. Blackburn, J. van Benthem, & F. Wolter (Eds.), *Handbook of modal logic* (Vol. 3), Amsterdam: Elsevier.

Huber, F., 2009, "Formal representations of belief", in E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2009 Edition), URL = <https://plato.stanford.edu/archives/sum2009/entries/formal-belief/>.

Joyce, J., 2004, "Bayesianism", in A. Mele & P. Rawling (Eds.), *The Oxford handbook of rationality*, Oxford: Oxford University Press.

Kadane, J.B. & P.D. Larkey, 1982, "Subjective probability and the theory of games", *Management Science*, 28(2): 113–120. [available online]

—–, 1983, "The confusion of is and ought in game theoretic contexts", *Management Science*, 29(12): 1365–1379. [available online]

Kaneko, M. & J. Kline, 1995, "Behavior strategies, mixed strategies and perfect recall", *International Journal of Game Theory*, 24: 127–145.

Kline, J., 2002, "Minimum memory for equivalence between *Ex Ante* optimality and time-consistency", *Games and Economic Behavior*, 38: 278–305.

Kuhn, H., 1953, "Extensive games and the problem of information", in H. Kuhn & A. Tucker (Eds.), *Contributions to the theory of games, vol. II*, Princeton: Princeton University Press.

Lewis, D., 1969, *Convention*, Cambridge: Harvard University Press.

Leyton-Brown, K. & Y. Shoham, 2008, *Essentials of game theory: A* *concise, multidisciplinary introduction*, New York: Morgan & Claypool.

Lismont, L. & P. Mongin, 1994, "On the logic of common belief and common knowledge", *Theory and Decision*, 37(1): 75–106.

–––, 2003, "Strong Completeness Theorems for Weak Logics of Common Belief", *Journal of Philosophical Logic*, 32(2): 115–137.

Liu, F., 2011, "A two-level perspective on preference", *Journal of Philosophical Logic*, 40(3): 421–439.

Lorini, E. & F. Schwarzentruber, 2010, "A modal logic of epistemic games", *Games*, 1(4): 478–526.

Mariotti, T., M. Meier, & M. Piccione, 2005, "Hierarchies of beliefs for compact possibility models", *Journal of Mathematical Economics*, 41: 303–324.

Mas-Colell, A., M. Winston, & J. Green, 1995, *Microeconomic theory*, Oxford: Oxford University Press.

Meier, M., 2005, "On the nonexistence of universal information structures", *Journal of Economic Theory*, 122(1): 132–139.

Mertens, J. & S. Zamir, 1985, "Formulation of Bayesian analysis for games with incomplete information", *International Journal of Game Theory*, 14(1): 1–29.

Modica, S. & A. Rustichini, 1994, "Awareness and partitional information structures", *Theory and Decision*, 37: 107–124.

–––, 1999, "Unawareness and partitional information structures", *Game and Economic Behavior*, 27: 265–298.

Monderer, D. & D. Samet, 1989, "Approximating common knowledge with common beliefs", *Games and Economic Behavior*, 1(2): 170–190.

Morris, S., 1995, "The common prior assumption in economic theory", *Economics and Philosophy*, 11(2): 227–253.

Moscati, I., 2009, *Interactive and common knowledge in the state-space model* (CESMEP Working Papers). University of Turin. [available online].

Myerson, R., 1997 [1991], *Game theory: Analysis of conflict*, Cambridge: Harvard University Press.

Osborne, M., 2003, *An introduction to game theory*, Oxford: Oxford University Press.

Pacuit, E., 2007, "Understanding the Brandenburger-Keisler paradox", *Studia Logica*, 86(3): 435–454.

Pacuit, E. & O. Roy, 2011, "A dynamic analysis of interactive rationality", in H. van Ditmarsch, J. Lang, & S. Ju (Eds.), *Proceedings of the third international workshop on logic, rationality and interaction* (Vol. 6953, pp. 244–258).

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

Perea, A., 2007, "A one-person doxastic characterization of Nash strategies", *Synthese*, 158: 251–271.

–––, 2012, *Epistemic game theory: Reasoning and choice*, Cambridge: Cambridge University Press.

–––, 2014, "Belief in the opponents' future rationality", *Games and Economic Behavior*, 83: 231–254.

Peterson, M., 2009, *An introduction to decision theory*, Cambridge: Cambridge University Press.

Piccione, M., & A. Rubinstein, 1997a, "On the interpretation of decision problems with imperfect recall", *Games and Economic Behavior*, 20(1): 3–24.

–––, 1997b, "The absent-minded driver's paradox: Synthesis and responses", *Games and Economic Behavior*, 20(1): 121–130.

Rabinowicz, W., 1992, "Tortuous labyrinth: Noncooperative normal-form games between hyperrational players", in C. Bicchieri & M. L. D. Chiara (Eds.), *Knowledge, belief and strategic interaction* (pp. 107–125).

Ross, D., 2010, "Game theory", in E. N. Zalta (Ed.), *The Stanford*

Encyclopedia of Philosophy (Fall 2010 Edition), URL = <https://plato.stanford.edu/archives/fall2010/entries/game-theory/>.

Roy, O. & E. Pacuit, 2013, “Substantive assumptions in interaction: A logical perspective”, *Synthese*, 190(5): 891–908.

Rubinstein, A., 1989, “The electronic mail game: Strategic behavior under ‘Almost common knowledge’”, *American Economic Review*, 79(3): 385–391.

–––, 1991, “Comments on the interpretation of game theory”, *Econometrica*, 59(4): 909–924.

Samuelson, L., 1992, “Dominated strategies and common knowledge”, *Game and Economic Behavior*, 4(2): 284–313.

Schelling, T., 1960, *The Strategy of Conflict*, Cambridge: Harvard University Press.

Schwitzgebel, E., 2010, “Belief”, in E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2010 Edition), URL = <https://plato.stanford.edu/archives/win2010/entries/belief/>.

Selten, R., 1975, “Reexamination of the perfectness concept for equilibrium points in extensive games”, *International Journal of Game Theory*, 4(1): 25–55. [available online].

Shoham, Y. & K. Leyton-Brown, 2008, *Multiagent systems*, Cambridge: Cambridge University Press.

Siniscalchi, M., 2008, “Epistemic game theory: Beliefs and types”, in S. Durlauf & L. Blume (Eds.), *The new Palgrave dictionary of economics*, Basingstoke: Palgrave Macmillan.

Spohn, W., 1982, “How to make sense of game theory”, *Philosophy of economics: Proceedings, Munich, July 1981*, W. Stegmüller, W. Balzer, & W. Spohn (eds), 239–270, *Studies in Contemporary Economics*, Volume 2, Berlin: Springer-Verlag.

Stalnaker, R., 1994, “On the evaluation of solution concepts”, *Theory and Decision*, 37(1): 49–73.

–––, 1996, “Knowledge, belief and counterfactual reasoning in games”, *Economics and Philosophy*, 12(02): 133–163.

–––, 1998, “Belief revision in games: forward and backward induction”, *Mathematical Social Sciences*, 36(1): 31–56.

–––, 1999, “Extensive and strategic forms: Games and models for games”, *Research in Economics*, 53(3): 293–319.

–––, 2006, “On logics of knowledge and belief”, *Philosophical Studies*, 128(1): 169–199.

Stuart Jr., H.W. & H. Hu, 2002, “An epistemic analysis of the Harsanyi transformation”, *International Journal of Game Theory*, 30(4): 517–525.

Tan, T.C.-C. & S.R. da Costa Werlang, 1988, “The Bayesian foundations of solution concepts of games”, *Journal of Economic Theory*, 45(2): 370–391, doi:10.1016/0022-0531(88)90276-1

Titelbaum, M., 2013, “Ten reasons to care about the sleeping beauty problem”, *Philosophy Compass*, 8: 1003–1017.

Ullmann-Margalit, E. & S. Morgenbesser, 1977, “Picking and choosing”, *Social Research*, 44: 757–785.

Vanderschraaf, P. & G. Sillari, 2009, “Common knowledge”, in E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2009 Edition), URL = <https://plato.stanford.edu/archives/spr2009/entries/common-knowledge/>.

de Weerd, H., R. Verbrugge, & B. Verheij, 2013, “How much does it help to know what she knows you know? An agent-based simulation study”, *Artificial Intelligence*, 199–200: 67–92.

Zvesper, J., 2010, *Playing with information* (PhD thesis), ILLC, University of Amsterdam.

Academic Tools

How to cite this entry.

- Preview the PDF version of this entry at the Friends of the SEP Society.
- Look up this entry topic at the Indiana Philosophy Ontology Project (InPhO).
- Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

- Baltag, A., and S. Smets, 2009, "Dynamic logics for interactive belief revision," slides for ESSLLI 2009 Course.
- Kets, W., 2014, "Bounded reasoning and higher-order uncertainty", *Northwestern Discussion Papers* [available online]

Related Entries

belief, formal representations of | common knowledge | epistemology: Bayesian | game theory | game theory: and ethics | prisoner's dilemma

Acknowledgments

The editors would like to thank Philippe van Basshuysen for reading this entry carefully and taking the time to inform us of a significant number of typographical errors.

Supplement to Epistemic Foundations of Game Theory

1. Proof of Lemma 3.1

Lemma 3.1 Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N} \rangle$ is a strategic game. A strategy $s_i \in S_i$ is strictly dominated (possibly by a mixed strategy) with

respect to $X \subseteq S_{-i}$ iff there is no probability measure $p \in \Delta(X)$ such that s_i is a best response with respect to p.

Proof: We start with some preliminary observations. Let $G = \langle S_1, S_2, u_1, u_2 \rangle$ be a two-player strategic game. Recall that $\Delta(S_1)$ and $\Delta(S_2)$ denote the mixed strategies for players 1 and 2, respectively. For $p_1 \in \Delta(S_1), p_2 \in \Delta(S_2)$, we write $U_1(p_1, p_2)$ (respectively $U_2(p_1, p_2)$) for the expected utility that player 1 (respectively 2) receives when 1 uses the mixed strategy p_1 and 2 uses the mixed strategy p_2. We assume that the players' choices are independent, so we have the following calculation for $i = 1, 2$:

$$U_i(p_1, p_2) = \sum_{x \in S_1} \sum_{y \in S_2} p_1(x) p_2(y) u_i(x, y)$$

A two-player game $G = \langle S_1, S_2, u_1, u_2 \rangle$ is **zero-sum** provided for each $x \in S_1$ and $y \in S_2$, $u_1(x, y) + u_2(x, y) = 0$. We make use of the following fundamental theorem of von Neumann:

> **Theorem S2 (von Neumann's minimax theorem)** For every two-player zero- sum game with finite strategy sets S_1 and S_2, there is a number v, called the **value** of the game such that:

1. $v = \max_{p \in \Delta(S_1)} \min_{q \in \Delta(S_2)} U_1(p, q)$
   $= \min_{q \in \Delta(S_2)} \max_{p \in \Delta(S_1)} U_1(p, q)$

2. The set of mixed Nash equilibria is nonempty. A mixed strategy profile (p, q) is a Nash equilibrium if and only if

$$p \in \text{argmax}_{p \in \Delta(S_1)} \min_{q \in \Delta(S_2)} U_1(p, q)$$
$$q \in \text{argmax}_{q \in \Delta(S_2)} \min_{p \in \Delta(S_1)} U_1(p, q)$$

3. For all mixed Nash equilibria (p, q), $U_1(p, q) = v$

Now, we can proceed with the proof of the Lemma. Suppose that $G = \langle N, \{S_i, u_i\}_{i \in N}\rangle$ is a strategic game where each S_i is finite.

Suppose that $s_i \in S_i$ is strictly dominated with respect to X. Then there is a $s_i \in S_i$ such that for all $s_{-i} \in X$, $u_i(s_i', s_{-i}) > u_i(s_i, s_{-i})$. Let $p \in \Delta(X)$ be any probability measure. Then for all $s_{-i} \in X$, $0 \leq p(s_{-i}) \leq 1$. This means that for all $s_{-i} \in X$, we have $p(s_{-i}) \cdot u_i(s_i', s_{-i}) \geq p(s_{-i}) \cdot u_i(s_i, s_{-i})$, and there is at least one $s_{-i} \in S_{-i}$ such that

$$p(s_{-i}) \cdot u_i(s_i', s_{-i}) > p(s_{-i}) \cdot u_i(s_i, s_{-i})$$

(this follows since p is a probability measure on X, so cannot assign probability 0 to all elements of X). Hence,

$$\sum_{s_{-i} \in S_{-i}} p(s_{-i}) \cdot u_i(s_i', s_{-i}) > \sum_{s_{-i} \in S_{-i}} p(s_{-i}) \cdot u_i(s_i, s_{-i})$$

So, $EU(s_i', p) > EU(s_i, p)$, which means s_i is not a best response to p.

For the converse direction, we sketch the proof for two player games. The proof of the more general statement uses the *supporting hyperplane theorem* from convex analysis. We do not discuss this extension here (note that it is not completely trivial to extend this result to the many agent case as we must allow players to have beliefs about *correlated* choices of their opponents). Let $G = \langle S_1, S_2, u_1, u_2\rangle$ be a two-player game. Suppose that $\alpha \in \Delta(S_1)$ is not a best response to any $p \in \Delta(S_2)$. This means that for each $p \in \Delta(S_2)$ there is a $q \in \Delta(S_1)$ such that $U_1(q, p) > U_1(\alpha, p)$. We can define a function $b : \Delta(S_2) \to \Delta(S_1)$ where, for each $p \in \Delta(S_2)$, $U_1(b(p), p) > U_1(\alpha, p)$.

Consider the game $(S_1, S_2, \overline{u}_1, \overline{u}_2)$ where

$$\overline{u}_1(s_1, s_2) = u_1(s_1, s_2) - U_1(\alpha, s_2)$$

and $\overline{u}_2(s_1, s_2) = -\overline{u}_1(s_1, s_2)$. This is a 2-person, zero-sum game, and so by the von Neumann minimax theorem, there is a mixed strategy Nash equilibrium (p_1^*, p_2^*). Then, by the minimax theorem, for all $m \in \Delta(S_2)$, we have

$$\overline{U}(p_1^*, m) \geq \overline{U}_1(p_1^*, p_2^*) \geq \overline{U}_1(b(p_2^*), p_2^*)$$

We now prove that $\overline{U}_1(b(p_2^*), p_2^*) > \overline{U}_1(\alpha, p_2^*)$:

$$\begin{aligned}
\overline{U}_1(b(p_2^*), p_2^*) &= \sum_{x \in S_1} \sum_{y \in S_2} b(p_2^*)(x) p_2^*(y) \overline{u}_1(x, y) \\
&= \sum_{x \in S_1} \sum_{y \in S_2} b(p_2^*)(x) p_2^*(y)[u_1(x, y) - U_1(\alpha, y)] \\
&= \sum_{x \in S_1} \sum_{y \in S_2} b(p_2^*)(x) p_2^*(y) u_1(x, y) \\
&\qquad - \sum_{x \in S_1} \sum_{y \in S_2} b(p_2^*)(x) p_2^*(y) U_1(\alpha, y) \\
&= U_1(b(p_2^*), p_2^*) - \sum_{x \in S_1} \sum_{y \in S_2} b(p_2^*)(x) p_2^*(y) U_1(\alpha, y) \\
&> U_1(\alpha, p_2^*) - \sum_{x \in S_1} \sum_{y \in S_2} b(p_2^*)(x) p_2^*(y) U_1(\alpha, y) \\
&> U_1(\alpha, p_2^*)
\end{aligned}$$

Since $p_2^*(y) U_1(\alpha, y)$ does not depend on x, we have

$$\begin{aligned}\sum_{x\in S_1}\sum_{y\in S_2} b(p_2^*)(x)p_2^*(y)U_1(\alpha,y) &= \sum_{x\in S_1} b(p_2^*)(x)\sum_{y\in S_2} p_2^*(y)U_1(\alpha,y)\\ &= \sum_{x\in S_1} b(p_2^*)(x)U_1(\alpha,p_2^*)\\ &= U_1(\alpha,p_2^*)\sum_{x\in S_1} b(p_2^*)(x)\\ &= U_1(\alpha,p_2^*)\end{aligned}$$

Hence, for all $m \in \Delta(S_2)$ we have $\overline{U}_1(s_1^*, m) > 0$ which implies for all $m \in \Delta(S_2)$, $U_1(s_1^*, m) > U_1(\alpha, m)$, and so α is strictly dominated by p_1^*. QED

2. Proof of Theorem 6.1

Theorem 6.1 (Brandenburger & Keisler 2006: Theorem 5.4) There is no assumption-complete type structure for the set of conjectures that contain the first-order definable subsets.

To prove this theorem, we follow an idea recently discussed in Abramsky & Zvesper (2010). Suppose that $\mathcal{C}_A \subseteq \wp(T_A)$ is a set of *conjectures* about Ann states (similarly, let $\mathcal{C}_B \subseteq \wp(T_B)$ be a set of conjectures about Bob states). We start with the flowing assumption:

For all $X \in \mathcal{C}_A$ there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: "in state x_0, Ann has consistent beliefs"
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: "in state x_0, Ann believes that Bob assumes X"

To prove the theorem, we need the following lemma.

> **Lemma S1** Under the above assumption, for each $X \in \mathcal{C}_A$ there is an x_0 such that
>
> $x_0 \in X$ iff there is a $y \in T_B$ such that $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$

> *Proof of Lemma S1:* Suppose that $X \in \mathcal{C}_A$. Then there is an $x_0 \in T_A$ satisfying 1 and 2.
>
> Suppose that $x_0 \in X$. By 1., $\lambda_A(x_0) \neq \emptyset$ so there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$. We show that $x_0 \in \lambda_B(y_0)$. By 2., we have $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in X = \lambda_B(y_0)$, as desired.
>
> Suppose that there is a $y_0 \in T_B$ such that $y_0 \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y_0)$. By 2., $y_0 \in \lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$. Hence, $x_0 \in \lambda_B(y_0) = X$, as desired.
>
> Consider a first-order language $\mathcal{L}$ whose signature contains binary relational symbols $R_A(x,y)$ and $R_B(x,y)$ defining λ_A and λ_B respectively. The language $\mathcal{L}$ is interpreted over qualitative type structures where the interpretation of R_A is the set $\{(t,s) \mid t \in T_A, s \in T_B, \text{ and } s \in \lambda_A(t)\}$. QED

Proof of Theorem 6.1. Consider the formula ϕ in $\mathcal{L}$:

$$\phi(x) := \exists y(R_A(x,y) \wedge R_B(y,x))$$

Then, the negation of ϕ is:

> $\neg\phi(x) := \forall y(R_A(x,y) \rightarrow \neg R_B(y,x))$: "all states x where any state y that Ann considers possible is such that Bob does not consider x possible at y." That is, this formulas says that "Ann believes that Bob's assumption is *wrong*."

Suppose that $X \in \mathcal{C}_A$ is defined by the formula $\neg\phi(x)$.

Suppose that there is a $x_0 \in T_A$ such that

1. $\lambda_A(x_0) \neq \emptyset$: Ann's beliefs at x_0 are consistent.
2. $\lambda_A(x_0) \subseteq \{y \mid \lambda_B(y) = X\}$: At x_0, Ann believes that Bob assumes $X = \{x \mid \phi(x)\}$ (i.e., Ann believes that Bob assumes that Ann believes that Bob's assumption is wrong.)

We have

$$\begin{array}{lll} \neg\phi(x_0) \text{ is true iff} & x_0 \in X & (\text{defn of } X) \\ \text{iff} & \text{there is a } y \in T_B \text{ with } y \in \lambda_A(x_0) & \\ & \quad \text{and } x_0 \in \lambda_B(y) & (\text{Lemma S1}) \\ \text{iff} & \phi(x_0) \text{ is true.} & (\text{defn of } \phi(x)) \end{array}$$

QED

Notes to Epistemic Foundations of Game Theory

1. We are bracketing cases where the players can flip a coin or, more generally, randomize between a number of strategies.

2. Not all choice rules presuppose these representations of preferences and beliefs. Minmax, for instance, makes recommendations or predictions in cases where decision makers have no probabilistic beliefs about the states of the environment.

3. This does not mean that the player will know exactly what the other players will do in the game. There may be more than one "rational choice" or the other players may randomize.

4. A variant of this problem is the well-known *sleeping beauty problem*, which has bee extensively discussed in the philosophy literature. Of course, much of the discussion of the sleeping beauty problem found in the philosophy literature is relevant here; however, the issues surrounding the sleeping beauty problem is typically framed differently than we have done in this section. See Titelbaum (2013) for a survey and pointers to the relevant literature.

5. Recall that I am restricting attention to finite strategic games.

6. A strategy profile is a sequence of actions, one for each player

7. A partition of W is a pairwise disjoint collection of subsets of W whose union is all of W. Elements of a partition Π on W are called **cells**, and for $w \in W$, let $\Pi(w)$ denote the cell of Π containing w.

8. Given an equivalence relation $\sim_i$ on W, the collection

$$\Pi_i = \{[w]_i \mid w \in W\}$$

is a partition. Furthermore, given any partition Π_i on W,

$$\sim_i = \{(w, v) \mid v \in \Pi_i(w)\}$$

is an equivalence relation with $[w]_i = \Pi_i(w)$.

9. Well-foundedness is only needed to ensure that for any set X, $Min_{\leq_i}(X)$ is nonempty. This is important only when W is infinite.

10. This is only one of many possible choices here, but it is the most natural in this setting (cf. Liu 2011).

11. Some care needs to be taken when W is infinite, but these technical issues are not important for us at this point, so we restrict attention to finite sets of states.

12. The weighed component of maximization of expected utility makes it difficult to capture in relational structures or plausibility models. In this entry, maximization of expected utility is always referring to type spaces or epistemic-probability models.

13. The uniqueness of the payoffs at each outcome is only needed to ensure that there is a unique backward induction solution.

14. The models used by Samuelson differ from the ones presented in Section 2. In his model, each state is assigned a *set* of actions for each agent (rather than a single action). This formal detail is important for Samuelson's main results, but is not crucial for the main point we are making here.

15. Recall the well-known distinction between "picking" and "choosing" from the seminal paper by Edna Ullmann-Margalit and Sidney Morgenbesser (1977).

16. Wlodeck Rabinovich (1992) takes this idea even further and argues that from the principle of indifference, players must assign equal probability to all choice-worthy options.

17. This same analysis applies to the other models discussed in Section 2.

18. The reasoning is that if there was a human intruder then the dog would bark and if a dog intruded then the cat would have howled.

19. Of course, one could move to different classes of models where monotonicity does not hold, for instance neighborhood models.