Epistemic Foundations of Game Theory

Eric Pacuit Olivier Roy

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 "solutions" 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.

Contents

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 $i, j, ...$ 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 σ, is called a **strategy profile**, or outcome of the game. We write σ_i for i's component in σ, and σ_{-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 1.1 is an example of a game in strategic form. There are two players, Ann and Bob, and each has to choose between two options: $N = \{Ann, Bob\}$, $S_{Ann} = \{t, b\}$ 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 t to the one she would get by choosing b, 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 (t, l) or (b, r)—but they are indifferent about which way they coordinate their choices. 11.5

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.

		Bob	
		l	r
Ann	t	$1,1$	$0,0$
	b	$0,0$	$1,1$

Figure 1: A coordination game

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, for instance iterated elimination of strictly or weakly dominated strategies. In the game above, for instance, there are two Nash equilibria in so-called "pure strategies."[1] These are the two coordination profiles: (t, l) and (b, 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 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 (t, l) or (b, 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; Kadane and Larkey, 1983; de Bruin, 2010) for a discussion.

1.2 From Games to Interactive Decision Problems

The *epistemic view on games* 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*). Once this is specified, a decision-theoretic *choice rule* can be used to make recommendations to (or to predict what) the decision maker about what she should (or will) choose. A standard example of a choice rule is maximazation of (subjective) expected utility, underlying the *Bayesian* view of rationality. It presupposes that the

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

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 3.)

The epistemic view on games is that one can use the same tools to analyze rational decision of individuals in interaction as to analyze rational decision under risk or uncertainty. In a game, each player is faced with their own decision problem, or sequence of decision problems in the case of extensive games. Epistemic game theory then employs the standard tools of decision theory to "solve" these decision problems. That is, the recommendations or predictions in a game come 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 are a number of epistemic analyses with alternative choice rules, for instance *minregret* (more on this below).

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 and Meier, 2010), in which the players have specific knowledge and/or beliefs about each other.

The importance of specifying the interactive decision problem, or context of the game, can be illustrated by a simple political example:

> Formally, a game is defined by its strategy sets and payoff functions. But in real life, many other parameters are relevant; there is a lot more going on. Situations that substantively are vastly different may nevertheless correspond to precisely the same strategic game. For example, in a parliamentary democracy with three parties, the winning coalitions are the same whether the parties each hold a third of the seats in parliament, or, say, 49 percent, 39 percent, and 12 percent, respectively. But the political situations are quite different. The difference lies in *the attitudes of the players, in their expectations about each other, in custom, and in history*, though the rules of the game do not distinguish between the two situations. (Aumann and Dreze, 2008, pg. 72, our emphasis)

Games are played in specific contexts where players have specific "expectations about each other", possibly induced by some background knowledge of

[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.

the "custom, and history". 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." Crucially, the context of the game involves *higher-order information*—i.e., beliefs about what others believe. This is the main ingredient of the epistemic view of games.

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.

2 Information in Games

There are various types of information that an agent 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?).

These four categories are conceptually important, but not necessarily exhaustive nor mutually exclusive. John Harasanyi, for instance, argued that all uncertainty about the structure of the game, that is all possible incompleteness in information, can be reduced to uncertainty over the payoffs (Harsanyi, 1967-68). (This was later formalized and proved by Stuart and Hu, 2002). Along these lines, Kadane and Larkey (1982) argued that all higher-order uncertainty can be reduced to strategic uncertainty. 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.

In this section, we briefly discuss some general issues about imperfect and incomplete information. Higher-order and strategic information are extensively discussed in the remainder of this entry.

2.1 Stages of Information Disclosure

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 degree 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.

Most of the literature we survey in this entry focuses on the *ex interim* stage as it allows for a straightforward assessment of the agents' rationality given their expectations. Philosophically, however, focusing on the *ex interim* stage raises interesting questions regarding possible *correlations* between an agent'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 an agent 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.[3]

2.2 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 John C. Harasanyi'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 and Shoham 2008, Chapter 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 Harasanyi's approach to modeling higher-order beliefs (which we discuss in Section 3.3).

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, Sections 4 and 5).

1. In a Bayesian games, players are uncertain about the payoffs of the game,

[3]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 were players choose sequentially, and will be addressed briefly in Section 5.2.3.

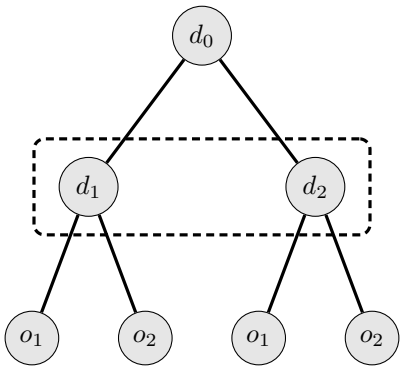
what other players believe are the correct payoffs, what other players believe that the other players believe about the payoffs, and so on, and this is the only source of uncertainty. That is, the players' (higher-order) beliefs about the payoffs in a game completely determine the (higher-order) beliefs 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.[4]

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 Harasanyi used to represent the players' beliefs in a game of incomplete information. Rather, they are better described as conventions followed by Harasanyi and subsequent researchers studying Bayesian games.

2.3 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:



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, Chapters 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.

[4]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.

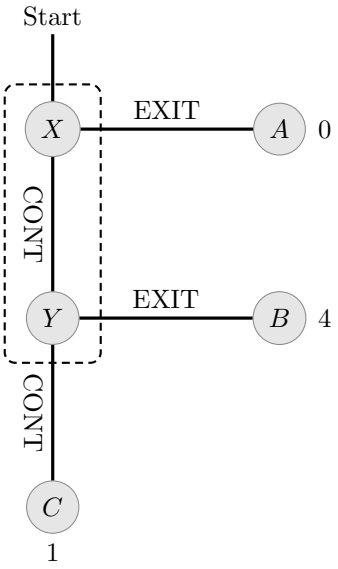
Kuhn (1953) introduced the distinction between *perfect* and *imperfect* recall in games with imperfect information. Roughly, players have perfect recall provided they remember all of their own past moves (see Bonanno, 2004; Kaneko and Kline, 1995, for general discussions of the perfect recall assumption). 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 2.1, there are different stages to the decision making process. Differences between these stages become even more pronounced in extensive games where 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, 2007a, 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 (1997b). 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 and Rubinstein, 1997b, pg. 7)

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



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. 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 et al., 1997; Halpern, 1997; Piccione and Rubinstein, 1997a; Board, 2003).

3 Game Models

3.1 General Issues

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 and Siniscalchi, 2002b) 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 6.3 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.

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*.[5] 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.

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.

Two main types of models have been used in the literature 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 et al., 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. Furthermore, these two different modeling paradigms have given rise to different styles of epistemic analysis (Brandenburger, 2007b): so-called *belief*-based characterizations on the one hand and *knowledge*-based characterizations on the other. In recent years, however, this dichotomy has tended to blur.

[5] A strategy profile is a sequence of actions, one for each player

3.2 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 and 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:

Given a (strategic or extensive) game G, a **strategy profile** is a sequence of strategy choices for each player. Let S be the set of strategy profiles. This set S represents the possible outcomes in a game G.

Definition 3.1 (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[6] over W and $\sigma : W \to S$. ◁

Epistemic models represents 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 3.2 (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 3.3 *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*

[6] 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.

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*[7], *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 and 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 2: 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

[7] 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 $[w]_i = \Pi_i(w)$.

fixed with an additional technical assumption: Suppose S is a set of strategy profiles for some (strategic or extensive) game with players $N = \{1, \ldots, n\}$.

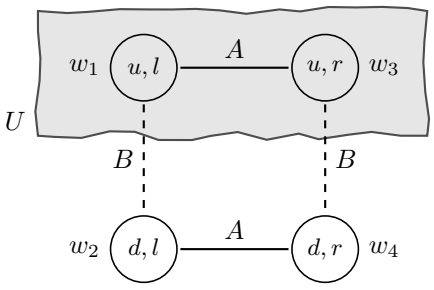
> A model $\mathcal{M} = \langle W, \{\Pi_i\}_{i \in N}, \sigma \rangle$ is said to be an ***ex interim* epistemic model** if for all $i \in N$ and $w, v \in W$, if $v \in \Pi_i(w)$ then $\sigma_i(w) = \sigma_i(v)$

where $\sigma_i(w)$ is the ith 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 1.1 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 (eg., "Ann knows that Bob knows that...") about these decisions.

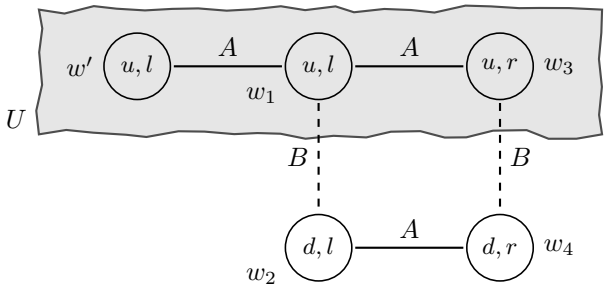
We now look at the epistemic model described above in more detail. We will often use the following diagrammatic representation of the model to ease exposition. States are represented by nodes in a graph where there is a (undirected) edge between states w_i and w_j when w_i and w_j are in the same partition 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:



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:



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 (eg., 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 discussion in Section 7.

3.2.1 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 **??**). 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; 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 et al., 1985), we focus on how to extend an epsitemic models 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 3.4 (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[8], 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 3.5 *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[9]

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

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):

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

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

- *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 (eg., 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 and Leyton-Brown, 2008, Section 13.7).

- *Strong Belief*: $B_i^s(E) = \{w \mid E \cap \Pi_i(w) \neq \emptyset \text{ and } E \cap \Pi_i(w) \preceq -E \cap \Pi_i(w)\}$
  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.

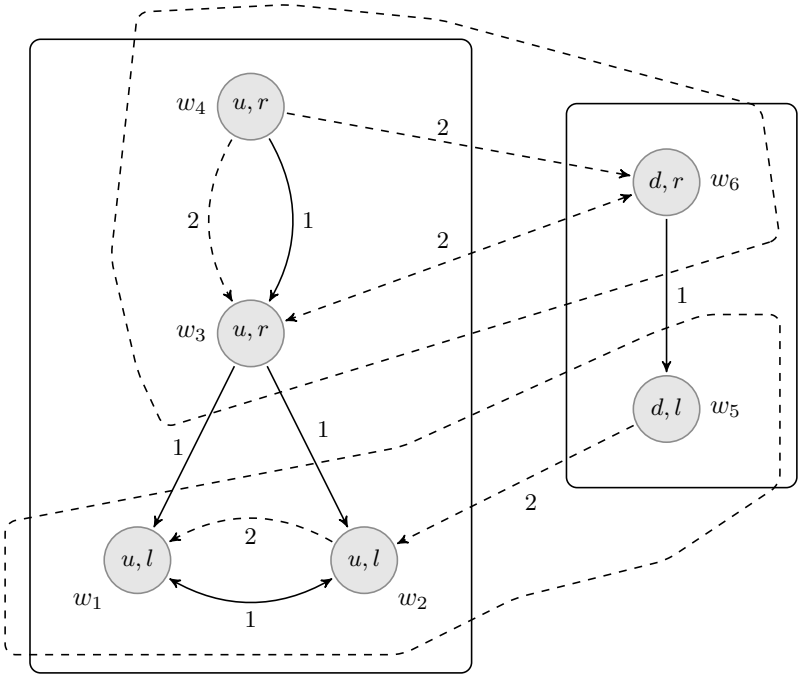
It is not hard to see that if agent i knows that E then i (robustly, strongly) believes that E. However, much more can be said about the logical relationship between these different notions. (The logic of these notions has been extensively studied by Alexandru Baltag and Sonja Smets in a series of articles, see Baltag and Smets 2009 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 3.2: 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.



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 $v \in W$ is 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 3.6 (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 3.4):

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" there 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 cubersome. 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[10]) 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.

[10]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.

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

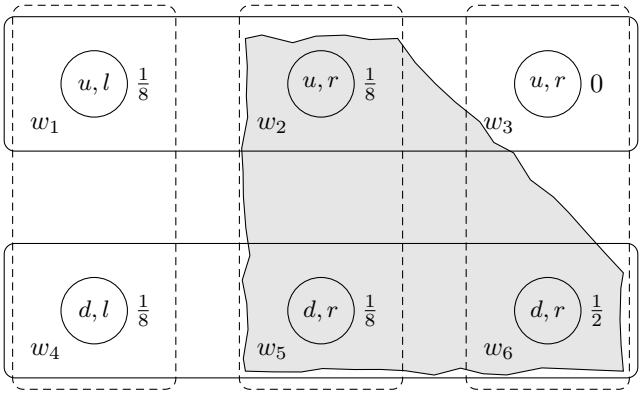
- *Probabilistic belief*: $B_i^r(E) = \{w \mid p_i^w(E) = r\}$
  Here, r can be any real number in the unit interval; however, it is often enough to restrict attention to the rational numbers in the unit interval.

- *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 et al., 1990; Heifetz and Mongin, 2001) complementing the logical frameworks discussed above (cf. Baltag and Smets, 2006).

We conclude this section with an example of an epistemic-probability model. Recall again the coordination game of Figure 3.2: 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.



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 $\Pi_2(w_6) \not\subseteq E$. 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.

3.3 Harasanyi Type Spaces

An alternative approach to modeling beliefs was initiated by Harsanyi in his seminar 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 associate with a specific hierarchy of belief

Definition 3.7 (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

$$\lambda_i : T_i \to \wp(\mathsf{X}_{j \neq i} T_j \times S). \qquad \triangleleft$$

So, each 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 types 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 1.1). 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:

$\lambda_A(t_1^A)$

	l	r
t_1^B	1	0
t_2^B	1	0

$\lambda_A(t_2^A)$

	l	r
t_1^B	0	0
t_2^B	1	0

$\lambda_B(t_1^B)$

	u	d
t_1^A	1	0
t_2^A	0	0

$\lambda_B(t_2^B)$

	u	d
t_1^A	0	0
t_2^A	0	1

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 means 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 $S = \mathsf{X} S_i$ 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 $\mathsf{X}_{j\neq i}T_j \times S$. Again if S is a set of strategy profiles (so $S = \mathsf{X}S_i$), then an event for agent i is a subset of $\mathsf{X}_{j\neq i}(T_j \times S_j)$.
- 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 3.7) allows us to represent *probabilistic* beliefs (we give the full definition here for future reference):

Definition 3.8 (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

$$\lambda_i : T_i \to \Delta(\mathsf{X}_{j\neq i}T_j \times S).$$

where $\Delta(\mathsf{X}_{j\neq i}T_j \times S)$ is the set of probability measures on $\mathsf{X}_{j\neq i}T_j \times S$. ◁

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) as 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

$\lambda_1(t_1)$:

	l	r
t_2	0.5	0
t_2'	0.4	0.1

Figure 3: Ann's beliefs about Bob

$\lambda_2(t_2)$:

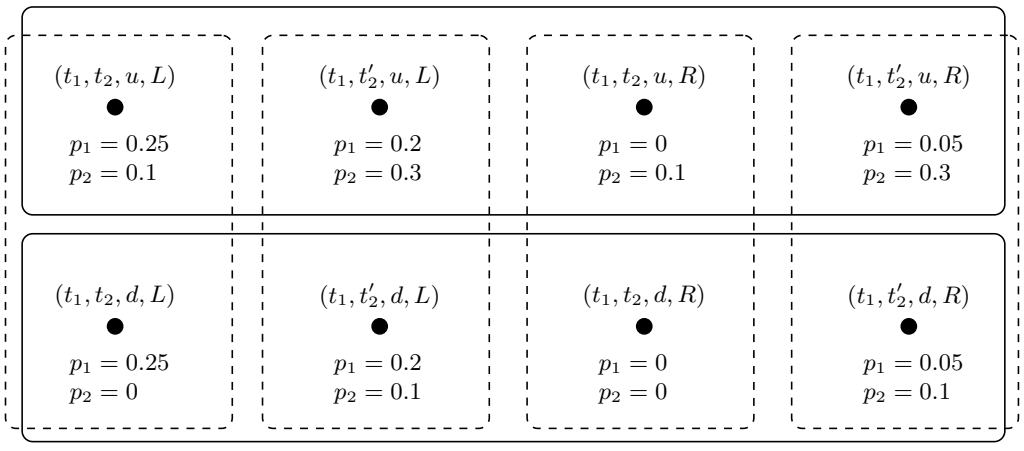
	u	d
t_1	1	0

$\lambda_2(t_2')$:

	u	d
t_1	0.75	0.25

Figure 4: Bob's beliefs about Ann

measures given below: 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):



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

3.4. 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 takes *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 $(p_1, p_2, \ldots, p_n)$ 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 6.2.

3.4 Common Knowledge

States in a game model not only represent the players 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 the 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 and Sillari, 2009) and (Fagin et al., 1995, Chapter 6).

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

$$K_G(E) \quad := \quad \bigcap_{i \in G} K_i(E)$$

where G is any nonempty set of players. If E is common knowledge for the group G, 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 $K_G^n(E)$ for each $n \geq 0$ by induction:

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

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

$$C_G(E) = \bigcap_{n \geq 0} K_G^n(E)$$

Unpacking the definitions, we have

$$C_G(E) = E \cap K_G\varphi(E) \cap K_G(K_G(E)) \cap K_G(K_G(K_G(E))) \cap \cdots$$

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).[11] See (Bonanno, 1996; Lismont and Mongin, 1994, 2003) for more information about the logic of common belief.

4 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

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 and Pass, 2011) or "knowledge-based rationality" (Artemov, 2009). See (Trost, 2009) for an analysis of all choice rules satisfying Savage's *sure-thing principle*. This entry focuses 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

[11] Although we do not discuss it in this entry, a probabilistic variant of common belief was introduced by (Monderer and Samet, 1989).

(c.f. the tripartite notion of *categorization* in Cubitt and Sugden, 2011; Pacuit and 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 remarks in Section 6.2.

4.1 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]

Definition 4.1 (Expected Utility) Suppose that $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ is a strategic game and $\Delta(S_{-i})$ is the set of probability measures over S_{-i}. The **expected utility** of $s_i \in S_i$ with respect to $p \in \Delta(S_{-i})$ 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.

Type spaces. Let $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ 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:

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

[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.

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 3.3 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.

Epistemic-Probability Models. The definition of a rationality event is similar in an epistemic-probability models. For completeness, we give the formal details. Suppose that $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ 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 $\mathsf{Rat} := \bigcap_{i\in N} \mathsf{Rat_i}$.

To illustrate the above definitions, consider the game in Figure 3.2 and the type space in Figures 4. The following calculations show that $((u,l),(t_1,t_2)) \in \mathsf{Rat_1}$ (u is the best response for player 1 given her beliefs defined by t_1):

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

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

4.2 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 4.2 (Strict Dominance) Suppose that $G = \langle N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N} \rangle$ is a strategic game and $X \subseteq S_{-i}$. Let $s_i, s_i' \in S_i$ be two strategies for player i. The strategy s_i **strictly dominates** s_i' **with respect to** X provided

$$\text{for all } s_{-i} \in X,\ u_i(s_i, s_{-i}) > u_i(s_i', s_{-i}).$$

We say s_i is **strictly dominated** provided there is some $s_i' \in S_i$ that strictly dominates s_i. ◁

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

Definition 4.3 (Weak Dominance) Suppose that $G = \langle N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N} \rangle$ is a strategic game and $X \subseteq S_{-i}$. Let $s_i, s_i' \in S_i$ be two strategies for player i. The strategy s_i **weakly dominates** s_i' **with respect to** X provided

$$\begin{gathered}\text{for all } s_{-i} \in X,\ u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i}) \\ \text{and} \\ \text{there is some } s_{-i} \in X \text{ such that } u_i(s_i, s_{-i}) > u_i(s_i, s_{-i}).\end{gathered}$$

We say s_i is **weakly dominated** provided there is some $s_i' \in S_i$ that strictly dominates s_i. ◁

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

Although, we do not focus on mixed strategies in this entry, we briefly mention a few issues concerning strict/weak dominance in the presence of mixed strategies. The above definitions of strict and weak dominance can be easily extended to the *mixed* extensions of strategic games. The key step is to replace the utility function u_i with an expected utility calculation: for $p \in \Delta(S_i)$ and $s_{-i} \in S_{-i}$, let $U_i(p, s_{-i}) = \sum_{s_i \in S_i} p(s_i) u_i(s_i, s_{-i})$. Then, p is strictly dominated by q with respect to $X \subseteq S_{-i}$, provided for all $s_{-i} \in X$, $U_i(q, s_{-i}) > U_i(p, s_{-i})$. Note that we do not extend the definition to probabilities over the opponents' strategies (i.e., replace the previous definition with "p is strictly p-dominated by q with respect to $X \subseteq \Delta(S_{-i})$, provided for all $m \in X$, $U_i(q, m) > U_i(p, m)$"). This is because both definitions are equivalent. Obviously, p-strict dominance implies strict dominance. To see the converse, suppose that p is dominated by q with respect to $X \subseteq S_{-i}$. We show that for all $m \in \Delta(X)$, $U_i(q, m) > U_i(p, m)$ (and so p is p-strictly dominated by q with respect to $\Delta(X)$). Suppose that $m \in \Delta(X)$. Then,

$$U_i(p, m) = \sum_{s_{-i} \in S_{-i}} m(s_{-i}) U_i(p, s_{-i}) > \sum_{s_{-i} \in S_{-i}} q(s_{-i}) U_i(q, s_{-i}) = U_i(q, m)$$

The parameter X in the above definitions is intended to represent the set of strategy profiles that the player i believes are still "live possibilities". Each state

in an epistemic (-plausibility) model is associated with a such a set of strategy profiles. This can be made explicit in epistemic (-plausibility) models as follows:

Epistemic Models. Suppose that $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ 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:

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

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

$\mathsf{Rat_i} := \{w \mid \sigma_i(w)$ is not strictly (weakly) dominated with respect to $S_{-i}(w)\}$.

In addition, we have $\mathsf{Rat} := \bigcap_{i\in N} \mathsf{Rat}_i$.

Epistemic-Plausibility Models. Suppose that $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ 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:

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

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

$\mathsf{Rat_i} := \{w \mid \sigma_i(w)$ is not strictly (weakly) dominated with respect to $S_{-i}(w)\}$.

Further, we again have $\mathsf{Rat} := \bigcap_{i\in N} \mathsf{Rat}_i$.

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 $w \sim_i w'$. This means that i's rationality is assessed on the basis of the result of his *current* choice according to different combinations of actions and information of *the others*.

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 dominate 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 as follows. We start with an easy observation (the proof is immediate from the definition of strict dominance):

Observation 4.4 *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. This is an interesting observation, but much more can be said.

The more general fact is that a strategy is strictly dominated if, and only if, the strategy is never a best response to any probabilistic belief the player might have about the strategies of her opponents. The following Lemma is well-known:

Lemma 4.1 *Suppose that $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ 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 Nuemann:

Theorem 4.5 (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*

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

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

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 $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ 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 $\overline{U}_1(b(p_2^*), p_2^*) > \overline{U}_1(\alpha, p_2^*)$:

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

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

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

$$= \sum_{x\in S_1} b(p_2^*)(x)U_1(\alpha, p_2^*) = U_1(\alpha, p_2^*) \sum_{x\in S_1} b(p_2^*)(x) = U_1(\alpha, p_2^*)$$

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

The general conclusion is that no dominated strategy can maximize expected utility at a given state, but not the other way around: some strategies that are not maximizing in a specific context might be in others, and are thus not necessarily strictly dominated.

Similar facts hold about *weak dominance*, though the situation is more subtle. A strategy s_i is weakly dominated or *inadmissible* provided there is another strategy s_i' such that player i's utility from s_i' is at least as good as s_i no matter what the other players do *and* there is a situation where s_i' gives player i strictly more utility than s_i. Of course, all strictly dominated strategies are weakly dominated, but not vice-versa. In many games, there will be more weakly dominated strategies than strictly dominated ones. The existential part of the definition means that the analogue of Observation 4.4 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 $X' \subseteq X$.

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 choice of her opponents. The following analogue of Lemma 4.1 is also well-known:

Lemma 4.2 *Suppose that $G = \langle N, \{S_i\}_{i \in N}, \{u_i\}_{i \in N} \rangle$ 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 no 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. Again, comparing the two Lemmas, we see that strict dominance implies weak dominance, but not necessarily vice versa. A strategy might not be a best response against any full-support probability distribution while being a best response to some beliefs, assigning say probability one to a state where the agents is indifferent between the outcome of its present action and the potentially inadmissible one.

5 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 rules of choice. This is a bottom-up, as opposed to the classical top-down, approach. Early work in this paradigm include Bernheim (Bernheim, 1984) and Pearce's (Pearce, 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 provided by classic solution concepts? Providing such conditions is known as an *epistemic characterization* of a given 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 backwards induction. These epistemic characterizations 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.

5.1 Characterization of Iterated Removal of Strictly Dominated Strategies

The most fundamental result of epistemic game theory is probably that "rationality and common belief in rationality implies iterated elimination of strictly dominated strategies." This result is already covered in (Vanderschraaf and 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. The result shows the importance of *higher-order information*.

5.1.1 The Result

Iterated elimination of strictly dominated strategies (ISDS) is a solution concept that runs as follow. 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:

		2		
		l	c	r
1	t	$3,3$	$1,1$	$0,0$
	m	$1,1$	$3,3$	$1,0$
	b	$0,4$	$0,0$	$4,0$

Note that r is strictly dominated for player 2 with respect to $\{t, m, b\}$. Once b 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:

	l	c	r
t	$3,3$	$1,1$	$0,0$
m	$1,1$	$3,3$	$1,0$
b	$0,4$	$0,0$	$4,0$

$\longmapsto$

	l	c
t	$3,3$	$1,1$
m	$1,1$	$3,3$
b	$0,4$	$0,0$

$\longmapsto$

	l	c
t	$3,3$	$1,1$
m	$1,1$	$3,3$

For arbitrary large (finite) strategic game, 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 usually credited to Bernheim (1984) and Pearce (1984). See (Sophn, 1982) for an early version, and (Brandenburger and 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

We then consider the pairs (s,t) where $s \in S_i$ and $t \in T_i$ and identify the 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 $R_i^1 = \mathsf{Rat_i}$. 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 general *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 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 in the set of strategies that survive iteratively removing strictly dominated strategies.

Theorem 5.1 *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 and Dekel 1987 (cf. also Tan and Werlang 1988):

Theorem 5.2 *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 (Apt and Zvesper, 2010) for a general discussion of these results and proofs of corresponding theorems in for game models that do not describe the players' beliefs as probabilities.

5.1.2 Philosophical Issues

Many authors have pointed out the strength of the common belief assumption in this result (see, eg., Gintis, 2009; de Bruin, 2010). It requires that the agents not only believe that the others are not choosing an irrational strategy, but also to believe that everybody believes that everybody believes that...the others are not choosing irrational strategies. 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 an model for it, a finite iteration of "everybody knows 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 of rationality has long been used to provide an informal explanation of the idealizations underlying classical game-theoretical analyses (Myerson, 1991). The result above shows 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 below.

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 agents' expectation 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; van Benthem and Gheerbrant, 2010; van Benthem et al., 2011; Pacuit and Roy, 2011).

Higher-order information is the key element in the above result:

"Bayesian rationality" alone, i.e. maximization of expected utility, is not sufficient to ensure ISDS in the general case.

In general, first-order belief of rationality will not do either. Strategic reasoning in games involves higher-order information.

When there is more than two players, common belief of rationality can force a player into the belief that the actions or beliefs of the other are *correlated* (Brandenburger and Dekel, 1987; Brandenburger and Friedenberg, 2008). In other words, in some games with more than two players, there is no model rationality is common belief and all players believe that the actions and/or beliefs of the others are *independent*.

The following example from (Brandenburger and 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 $S_C = \{x, y, z\}$:

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

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

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

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 $3(1-a)b + 3(1-a)(1-b) = 3(1-a)(b+(1-b)) = 3(1-a)$.

There are three cases:

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

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

   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,

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

   Hence, y is not a best response.

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

5.2 Epistemic Characterization of Backwards Induction

The second fundamental result that we discuss 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 *backwards induction*. The epistemic characterization of this solution concept is in terms of "substantive rationality" and common belief in substantive rationality (cf. also Vanderschraaf and Sillari, 2009, Section 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 backwards 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 backwards induction path.

5.2.1 Extensive games: basic definitions

Extensive games make explicit the sequential structure of choices in a game. They can be represented as tree-like structures, as for instance in Figure 5.2.1, which is an instance of the well-known *centipede game*. This game is an extensive

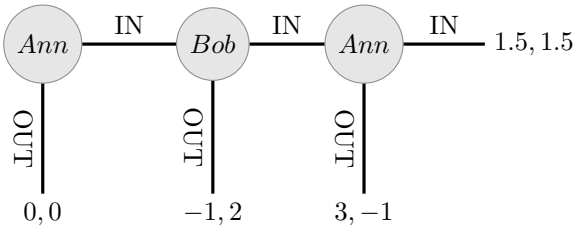


Figure 5: An extensive game

game of *perfect information*.

Definition 5.3 (Extensive form games) A *game in extensive form* $\mathcal{T}$ is a tuple $\langle I, T, \tau, \{\pi_i\}_{i \in I} \rangle$ such that:

- I is a finite set of players.
- T is a finite set of finite sequences of *actions*, called *histories*, such that:
  - The empty sequence $\emptyset$, the *root* of the tree, is in T.
  - T is closed under sub-sequences: if $(a_1, \ldots, a_n, a_{n+1}) \in T$ then $(a_1, \ldots, a_n) \in T$.

  Given a history $h = (a_1, \ldots, a_n)$, the history $(a_1, \ldots, a_n, a)$, h followed by the action a, is denoted ha. We write $A(h)$ for the set of actions a such that $ha \in T$. A history h is *terminal in* T whenever it is the sub-sequence of no other history $h' \in T$. Z denotes the set of terminal histories in T.
- $\tau : (T - Z) \to I$ is a *turn function* which assigns to every non-terminal history h the player whose turn it is to play at h.
- $\pi_i : Z \to \mathbb{B}$ is a real-valued *payoff function* on the leaves or terminal histories of the tree.

◁

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 5.4 (Strategies) A *strategy* σ_i for agent i is a function that gives, for every history h such that $i = \tau(h)$, an action $a \in A(h)$. S_i is the set of strategies for agent i.

- A *strategy profile* $\sigma \in \Pi_{i \in I} S_i$ is a combination of strategies, one for each agent. Agent i's component of σ is denoted σ_i, and $\sigma(h)$ is shorthand for the action a such that $a = \sigma_i(h)$ for the agent i whose turn it is at h.

A history h' is *reachable* or *not excluded* by the profile σ from h if $h' = (h, \sigma(h), \sigma(h, \sigma(h)), ...)$ for some finite number of applications of σ. A history h' is *reachable* from h by playing the strategy σ_i, for $i = \tau(h)$ if there is a combination of strategies for the other players $\sigma_{j \neq i}$ such that h' is reachable from h by the profile $(\sigma_i, \sigma_{j \neq i})$. ◁

In the game displayed in Figure 5.2.1, for instance, (Out, In) is a strategy for *Ann*, even though by going down at the first node she ends the game and thus never gets to play her second turn.

5.2.2 Epistemic Characterization of Backward Induction

In extensive games, the sequential decision structure allows for at least two *loci* for the assessment of the rationality of an agent: "local" rationality *at a given history* of the tree, and "global" rationality *at a state* of an epistemic frame. We first define the former.

Definition 5.5 (Rationality at state-history pair) Agent i is *irrational* at history h in state w, written $w \in IRR_i^h$, whenever there is a strategy $\sigma_i' \neq \sigma_i(w)$, such that $h'' \succ_i^Z h'$ for all terminal histories h', h'' reachable from h by $\sigma(w')$ and $(\sigma_i', \sigma_{j \neq i}(w'))$, respectively. Agent i is *rational* at node h in state w, written $w \in \neg IRR_i^h$, whenever she is not irrational at w. ◁

Intuitively, an agent is irrational at a history h given a certain state w whenever the strategy she is playing at w is strictly dominated in the sub-game starting at h, given her information. All possible outcomes that she considers possible by playing such a strategy from that history are strictly less preferred than those she considers would have been possible to achieve had she chosen another strategy. In Figure 5.2.1, for instance, choosing In at her second move is never rational for Ann according to this definition.

This notion of rationality at a history h 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 h, i.e. which

choices have or could have lead to a given history h. We shall return to this in the discussion below.

A consequence of the local definition of rationality is that the (ir-)rationality of a certain choice is independent of an agent's information at histories that are only followed by terminal ones. At such histories, which we call pre-terminal, it does not matter what the agent whose turn it is to move expects of the other. Her choices entirely determine the outcome of the game, within those that are reachable from the current history, of course. Again, at history (In, In) of Figure 5.2.1 choosing In is never rational for Ann. Whatever she believes, or rather believed about what Bob would do, if the game reaches this point the outcome of opting Out is always better then the one she gets by playing In.

How to lift this local definition of rationality at a state-history pair to a global statement about rationality at a state is the main point of divergence between the two analysis of common knowledge of rationality in extensive games that we survey here. We start with the "classical" one, proposed in (Aumann, 1995).

Definition 5.6 (Substantive Rationality) A player is *substantively rational* at a state w whenever she is locally rational at all histories-pairs (w, h). $\triangleleft$

Information about rationality at pre-terminal histories can be of great importance to agents choosing earlier in the game. If, at a given state w of a model of the game in Figure 5.2.1, Bob knows that Ann is substantively rational, then he knows that she will choose Out. But then the only rational choice for him is to opt Out himself, leading to the terminal history (In, Out) instead of the strictly less preferred (In, In, Out). If Ann knows that Bob knows that she is substantively rational, and that he is rational himself, she will go through a similar reasoning and opt Out at the first move.

The strategy profile $((Out, Out), Out)$ is the unique pure-strategy *sub-game perfect equilibrium* (Selten, 1975) of that game, and the reasoning that we went through in the previous paragraph is close to the algorithm known as *backward induction* for solving extensive games of perfect information. This algorithm tells to start by identifying the most preferred choices at the histories which are one step before the terminal ones, then proceed further inwards in the game tree by identifying the most preferred choices on the assumption that the choices thereafter will be those identified previously. In games of perfect information the strategy profile(s) computed by the backward induction algorithm coincide with the sub-game perfect equilibrium profiles. See (Osborne and Rubinstein, 1994) for a precise definition of this algorithm.

The reasoning used above, i.e., moving from the pre-terminal histories to earlier ones by adding levels of knowledge of local rationality, indeed generalizes to any finite extensive game with perfect information. Common knowledge of susbstantive rationality implies sub-game perfect equilibrium play.

Theorem 5.7 *(Aumann, 1995) The following are equivalent, for any game in extensive form $\mathcal{T}$ and strategy profile s of it:*

1. *There is a state w of an epistemic frame for $\mathcal{T}$ such that $f(w) = s$ and $w \in C_I(\bigcap_{i \in I} \neg IRR_i)$.*

2. *s is a sub-game perfect equilibrium of $\mathcal{T}$.*

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 Bob do at history (In) under common knowledge of rationality, and we concluded that he would opt Out. But if the game ever reaches that state then by the theorem above Bob has to conclude that either Ann is not rational, or that she doesn't believe 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 (de 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.

It is by now widely acknowledged in epistemic game theory that the crux of the backward induction solution, from an epistemic perspective, is that the players should keep their beliefs in other players' rationality at future nodes as long as this belief is consistent with the hard information they have (which includes the moves they observed previously). This is certainly the case when substantive rationality is common *knowledge*. In fact, one might argue, if they really have knowledge of that fact then they will not observe any deviation from the backward induction solution. Knowledge is truthful, after all. More generally, however, such deviation will *not* be inconsistent with future rationality, except in very specific contexts where the players have precise beliefs regarding the meaning of a deviation from the backward induction path. This form of *strong belief* (see Section 3.4 above) in rationality has been shown to be sufficient for backward induction (Battigalli and Siniscalchi, 2002a; Baltag et al., 2009; Perea, 2012).

Theorem 5.8 *(Battigalli and Siniscalchi, 2002a) The following are equivalent, for any game in extensive form $\mathcal{T}$ and strategy profile s of it:*

1. *There is a state (σ, t) in a type structure for $\mathcal{T}$ such that at which all types are rational and rationality is common strong belief.*

2. *σ is a sub-game perfect equilibrium of $\mathcal{T}$.*

In a nutshell:

One does not need common knowledge of rationality to play backward induction: rationality and common strong belief in rationality suffice.

The "dynamic" approaches to common knowledge of rationality in strategic game have also been generalized characterization of common knowledge and

common strong belief of rationality in extensive games (van Benthem and Gheerbrant, 2010; van Benthem et al., 2011)

5.2.3 Common Knowledge of Rationality without Backward Induction

We just saw that if the players in an extensive game believe in each others' future rationality, and hold this belief as long as it is inconsistent with the hard information they have, then they will play according to the backward induction solution. However:

With an explicit account of the players' counter-factual beliefs, common knowledge of rationality does not imply backward induction.

The most well-known example of this is due to Stalnaker (1996). The key to his analysis is to endow the player the players with an explicit *belief revision policy*, that tells which informational state they would revert to in case they were to observe moves that are inconsistent with the hard information they have. The counter-factual character (would... in case were) is important here. The agent's hard information, what they "know", in Stalnaker's models is modeled just as before: by an epistemic accessibility relation. But rationality statements at a given node are evaluated in the *closest, counter-factual* doxastic state the agent would be in if that node was reached.

Formally, these counter-factual beliefs are is encoded by enriching epistemic models[13] with a selection function which assigns to each state and history pair (w, h) the state w' which is closest to w according to i where h is reached. Such selection functions are common tools in models for *conditionals* or *belief revision*. Formally, the value of such function ρ_i at a pair (w, h) is constrained by three postulates:

1. The history h is reached in $\rho_i(w, h)$.
2. If history h is reached at w then $\rho_i(w, h) = w$.
3. If $w' = \rho_i(w, h)$ then for all $h' \supseteq h$, $\sigma(w)(h') = \sigma(w')(h')$.

Constraint (1) should be seen as a success postulate. The function brings the agent to a informational state where reaching history h is actually reached, and at the very least not excluded given the agents' hard information. If h is already reached at the current state, then learning that won't change the agents' hard information. This is the content of Postulate 2, analogous to the centering assumption in belief revision. Finally, the third postulate states that what the agent's would believe after history h should not be altered by this belief revision. This minimality postulate ensures that the agents change as little as possible their beliefs about future actions—allowing, crucially, *some* change

[13] Here we only focus on hard information and the way it would be revised.

in their beliefs about the information, and thus potentially the rationality, of others.

With this in hand, a simple change in the notion of global rationality is enough to break the connection between common knowledge of rationality and the backward induction solution.

Definition 5.9 (Counter-factual rationality) A player is *rational** at a state w whenever she is locally rational at $(\rho_i(w,h),h)$ for all histories-pairs (w,h). $\triangleleft$

The intuition behind this change is simple: an agent is rational at a state w if for all histories and pairs (w,h), she is rational at that history, *in the (informational) state she would be in if that history was reached.* If h is actually already reached in w then this boils down to what we had before, but otherwise the substantive rationality and rationality* might diverge. A simple example, devised by Halpern (2001a) shows that.

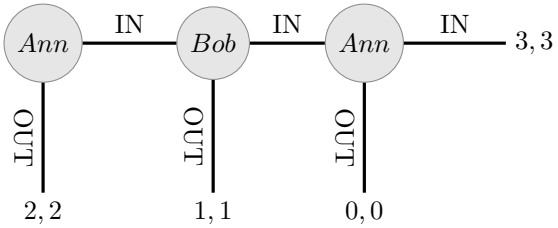


Figure 6: A game where common knowledge of rationality with belief revision does not imply backward induction.

Consider the game in Figure 5.2.3, with the epistemic model in Figure 5.2.3, and the value of the selection function as specified in Table 5.2.3.[14]

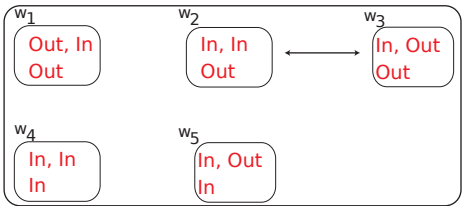


Figure 7: A model for the game in Figure 5.2.3. A each state w_i we have that $[w_i]_{Ann} = \{w_i\}$. Same for Bob, except for w_2 and w_3, with his hard information illustrated by the double arrow. The value of $f(w_i)$ is written in each state, with Ann's strategy on top and Bob below.

Now observe that Bob plays In at his turn in state w_1 while knowing that Ann plays Out after. He is thus locally irrational at that state-history pair,

[14] The reader can check that these are the only value of ρ_i that meet postulate 1 to 3 above.

ρ	$\emptyset$	(In)	(In, In)
w_1	w_1	w_2	w_4
w_2	w_2	w_2	w_4
w_3	w_3	w_3	w_5
w_4	w_4	w_4	w_4
w_5	w_5	w_5	w_5

Table 1: The value of $\rho_i(w, h)$ for both players in the model of Figure 5.2.3.

and so substantive rationality is not common knowledge. And, indeed, at that state Ann and Bob do not play according to the backward induction solution, which tells to both of them to play In always. But rationality*, *is* common knowledge at that state. To see that, given that w_1 is the only state compatible with Ann and Bob's hard information, all we have to do is to check that they are globally rational at that state, which boils down to checking that at each node both Ann and Bob are rational at the information state she/he would be if that node was reached. Observe that In, In is not reached at w_1, so to evaluate rationality at that node we have to move to a different information state, namely $w_4 = \rho(w_1, (In, In))$. At that state Ann plays In, and she is rational there. Node In is not reached either. Here the closest state according to ρ is w_2. In that state Bob is opting out but, unlike at w_1, *this is rational for him* at that state because he considers it possible (w_3) that Ann opts out at her next move. Finally, Ann is rational at the root ($\emptyset$) because she knows that Bob is would opt out at his turn. So both players are globally rational* at w_1, despite the fact that they do not play the backward induction solution, and this is common knowledge.

6 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.

6.1 Nash Equilibrium

6.1.1 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 1.1 above) for instance, all profiles, can be played under rationality and common belief of rationality.

Looking again at Figure 1.1, one can ask what would happen if Bob *knew* (that is had correct beliefs about) Ann's strategy choice? Intuitively, it is quite

clear what 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 and Hart, 1994) and philosophy (eg., famously in (Lewis, 1969)).

The seminal result of Aumann and 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, (Sophn, 1982) for an early statement. Before stating the theorem, we discuss an example from (Aumann and 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

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

$\lambda_A(a_3)$

	u	d
a_1	0.5	0
a_2	0	0.5
a_3	0	0

$\lambda_B(b_1)$

	u	d
a_1	0.5	0
a_2	0	0
a_3	0	0.5

$\lambda_B(b_2)$

	u	d
a_1	0	0
a_2	0	0.5
a_3	0	0.5

$\lambda_B(b_3)$

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 and 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 6.1 (Aumann and 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 and Brandenburger, 1995, Theorem B) for precise formulation of the result, and, again, (Sophn, 1982) for an early version. See, also, (Perea, 2007) and (Tan and Werlang, 1988) for similar results about the Nash equilibrium.

6.1.2 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; de 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 2.1, 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 6.3). 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.

6.1.3 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 (van der Hoek and 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, 2011; Lorini and Schwarzentruber, 2010)).

6.2 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 **??**, 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 and Dufwenberg, 2003; Samuelson, 1992; Cubitt and Sugden, 1994; Brandenburger et al., 2008). For example, (Samuelson, 1992) showed (among other things) that the analogue of Theorem 5.1 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	$1,0$	$0,1$

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 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 et al., 1995, pg. 240)

The extent of this conflict is nicely illustrated in (Samuelson, 1992). In particular, Samuelson (1992) shows that there is no epistemic-probability model[15] 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 beliefs means that a player cannot *know* which options her opponent(s) will *pick*[16] 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, pg. 8) and "no extraneous beliefs" by Asheim and Dufwenberg (2003).[17] For an extended discussion of the above assumption see (Cubitt and 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

[15] The models used by Samuelson differ from the ones presented in Section 3. 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.

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

[17] 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.

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 and Dufwenberg, 2003; Halpern and 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 $(p_1, p_2, \ldots, p_n)$ with supports (The **support** of a probability measures 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 $(p_1, \ldots, p_n)$ 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 has 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, pg. 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** and 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).

6.3 Incorporating Unawareness

The game models introduced in Section 3 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 3 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 3 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' story *Silver Blaze* (pgs. 346-7):

> ...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? [18]

[18] This same analysis applies to the other models discussed in Section 3.

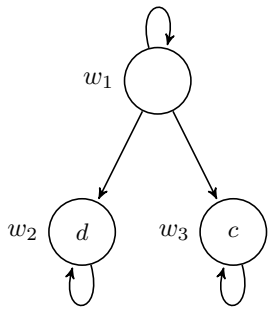
Suppose that $U_i(E)$ is the event that the player i is unaware of the event E. Of course, if i is unaware of E then i does not know that E is true ($U_i(E) \subseteq \overline{K_i(E)}$, where $\overline{X}$ denotes the complement of the event X). Recall that in epistemic models (where the players' information 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 $P_i : W \to \wp(W)$. 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[19]). 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:



Consider the following calculations:

[19]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.

- $K(E) = \{w_2\}$ (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. The website `http://www.econ.ucdavis.edu/faculty/schipper/unaw.htm` has a up-to-date list of references. The key references are (Board et al., 2011; Dekel et al., 1998; Halpern, 2001b; Heifetz et al., 2006; Halpern and Rego, 2008; Chen et al., 2012).

7 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 3. As we noted already in Section 3.2, 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 (eg., history, conventions, etc.) and this "context" causes the players to rule out certain beliefs." (Brandenburger and Friedenberg, 2010, pg. 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 propositions or basic facts, 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 3, where it is not common knowledge that the players believe all the logical consequences of their beliefs.[20]

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 the how the players' informational attitudes are represented—eg., 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 and Keisler, 2006a, Section 11) and (Siniscalchi, 2008, Section 3) for discussion and pointers to the relevant results). One approach considers the space of all (Harasanyi 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 and Zamir, 1985; Brandenburger and Dekel, 1993). For Kripke structures the question as been answered in the negative (Heifetz and Samet, 1998; Fagin et al., 1999; Meier, 2005), with some qualifications regarding the language that is used to describe them (Heifetz, 1999; Roy and Pacuit, 201X).

A second approach takes and internal perspective asking whether, *for a fixed set of states or types*, the agents are making any substantive assumptions about

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

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 (2006b) 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 an very strong impossibility results 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 3.7). A

qualitative type space for two players (cf. Definition 3.7. 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 results is:

Theorem 7.1 *(Brandenburger and Keisler, 2006b, Theorem 5.4) There is no assumption-complete type structure for the set of conjectures that contain the first-order definable subsets.*

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

[[**Put in a supplemental page**]] To prove this theorem, we follow an idea recently discussed in (Abramsky and 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"

Lemma 7.1 *Under the above assumption, for each $X \in \mathcal{C}_A$ there is an x_0 such that*

$$x_0 \in X \text{ iff there is a } y \in T_B \text{ such that } y \in \lambda_A(x_0) \text{ and } x_0 \in \lambda_B(y)$$

Proof. 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. QED

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 $\{(t,s) \mid t \in T_A, s \in T_B, \text{ and } s \in \lambda_A(t)\}$.

Consider the formula φ in $\mathcal{L}$:

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

Then, the negation of φ is:

$\neg\varphi(x) := \forall y(R_A(x,y) \rightarrow \neg R_B(y,x))$: "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*."

Proof of Theorem 7.1. Suppose that $X \in \mathcal{C}_A$ is defined by the formula $\neg\varphi(x)$.

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 $X = \{x \mid \varphi(x)\}$ (i.e., Ann believes that Bob assumes that Ann believes that Bob's assumption is wrong.)

We have

$\neg\varphi(x_0)$ is true	iff (definition of X)	$x_0 \in X$
	iff (Lemma 7.1)	there is a $y \in T_B$ with $y \in \lambda_A(x_0)$ and $x_0 \in \lambda_B(y)$
	iff (definition of $\varphi(x)$)	$\varphi(x_0)$ is true.

8 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.

9 Bibliography

References

Abramsky, S. and Zvesper, J. A., 2010, "From Lawvere to Brandenburger-Keisler: interactive forms of diagonalization and self-reference", *CoRR*, abs/1006.0992.

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

Apt, Krzysztof and Zvesper, Jonathan, 2010, "The role of monotonicity in the epistemic analysis of strategic games", *Games*, 1(4): 381–394, URL `http://www.mdpi.com/2073-4336/1/4/381`.

Artemov, S., 2009, "Knowledge-based rational decisions", Tech. rep., CUNY Ph.D. Program in Computer Science. Technical Report TR-2009011.

Asheim, Geir and Dufwenberg, Martin, 2003, "Admissibility and common belief", *Game and Economic Behavior*, 42: 208 – 234.

Aumann, R., 2010, "Interview on epistemic logic", in *Epistemic Logic: Five Questions*, Vincent F. Hendricks and Olivier Roy, eds., Automatic Press, 21–35.

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

Aumann, R.J. and Hart, S., 1994, *Handbook of game theory with economic applications*, vol. 2, North Holland.

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

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

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

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

Aumann, Robert and Brandenburger, Adam, 1995, "Epistemic conditions for nash equilibrium", *Econometrica*, 63(5): 1161–1180.

Aumann, Robert and Dreze, Jacques, 2008, "Rational expectations in games", *American Economic Review*, 98(1): 72 – 86.

Aumann, Robert J., Hart, Sergiu, and Perry, Motty, 1997, "The absent-minded driver", *Games and Economic Behavior*, 20(1): 102 – 116, URL `http://www.sciencedirect.com/science/article/pii/S0899825697905777`.

Baltag, A. and Smets, S., 2006, "Conditioanl doxastic models: A qualitative approach to dynamic belief revision", in *Electornic notes in theoretical computer science*, Springer, vol. 165, 5 – 21.

Baltag, A. and Smets, S., 2009, "ESSLLI 2009 course: Dynamic logics for interactive belief revision", Slides available online at http://alexandru.tiddlyspot.com/#%5B%5BESSLLI09%20COURSE%5D%5D.

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

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

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

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

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

Board, Oliver, 2003, "The not-so-absent-minded driver", *Research in Economics*, 57(3): 189 – 200, URL http://www.sciencedirect.com/science/article/pii/S1090944303000346.

Bonanno, Giacomo, 1996, "On the logic of common belief", *Mathematical Logical Quarterly*, 42: 305 – 311.

Bonanno, Giacomo, 2004, "Memory and perfect recall in extensive games", *Games and Economic Behavior*, 47(2): 237 – 256, URL http://www.sciencedirect.com/science/article/pii/S0899825603001933.

Brandenburger, A., 2003, "On the existence of a 'complete' possibility structure", in *in Cognitive Processes and Economic Behavior*, M. Basili, N. Dimitri, and I. Gilboa, eds., Routledge, 30–34.

Brandenburger, A. and Dekel, E., 1993, "Hierarchies of beliefs and common knowledge", *Journal of Economic Theory*, 59.

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

Brandenburger, Adam, 2007a, "A note on Kuhn's theorem", in *Interactive Logic, Proceedings of the 7th Augstus de Morgan Workshop, London*, J. van Benthem, D. Gabbay, and B. Loewe, eds., Texts in Logic and Games, Amsterdam University Press, 71 – 88.

Brandenburger, Adam, 2007b, "The power of paradox: some recent developments in interactive epistemology", *International Journal of Game Theory*, 35(4): 465–492.

Brandenburger, Adam, 2010, "Origins of epsitemic game theory", in *Epistemic Logic: Five Questions*, Vincent F. Hendricks and Olivier Roy, eds., Automatic Press, 59–69.

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

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

Brandenburger, Adam and Friedenberg, Amanda, 2010, "Self-admissible sets", *Journal of Economic Theory*, 145: 785–811.

Brandenburger, Adam, Friedenberg, Amanda, and Keisler, H. Jerome, 2008, "Admissibility in games", *Econometrica*, 76(2): 307–352.

Brandenburger, Adam and Keisler, H. Jerome, 2006b, "An impossibility theorem on belies in games", *Studia Logica*, 84(2): 211–240.

Chen, Y. C., Ely, J., and Luo, X., 2012, "Note on unawareness", *International Journal of Game Theory*, forthcoming.

Cubitt, Robin and Sugden, Robert, 2011, "Common reasoning in games: A Lewisian analysis of common knowledge of rationality", CeDEx Discussion Paper.

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

de Bruin, B., 2010, *Explaining Games : The Epistemic Programme in Game Theory*, Synthese Library, Springer.

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

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

Fagin, R., Halpern, J., Moses, Y., and Vardi, M., 1995, *Reasoning about Knowledge*, The MIT Press.

Fagin, Ron, Halpern, Joe, and Megiddo, Nimrod, 1990, "A logic for reasoning about probabilities", *Information and Computation*, 87(1-2): 78 – 128.

Finetti, B., 1974, *Theory of Probability, Vols. 1 and 2*, Wiley, New York.

Friedenberg, Amanda and Meier, Martin, 2010, "The context of a game", Manuscript, URL `http://www.public.asu.edu/~afrieden/cog-08-26-10.pdf`.

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

Halpern, J. Y., 2001a, "Substantive rationality and backward induction", *Games and Economic Behavior*, 37(2): 425–435.

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

Halpern, Joe, 2001b, "Alternative semantics for unawareness", *Game and Economic Behavior*, 37: 321 – 339.

Halpern, Joe, 2003, *Reasoning about Uncertainty*, The MIT Press.

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

Halpern, Joe and Pass, Rafael, 2009, "A logical characterization of iterated admissibility", in *Proceedings of the Twelfth Conference on Theoretical Aspects of Rationality and Knoweldge*, Aviad Heifetz, ed., 146 – 155.

Halpern, Joe and Rego, L. C., 2008, "Interactive unawareness revisited", *Game and Economic Behavior*, 62: 232 – 262.

Halpern, Joseph, 1991, "The relationship between knowledge, belief, and certainty", *Annals of Mathematics and Artificial Intelligence*, 4(3): 301 – 322, URL `http://dx.doi.org/10.1007/BF01531062`.

Halpern, Joseph, 1997, "On ambiguities in the interpretation of game trees", *Games and Economic Behavior*, 20(1): 66 – 96, URL `http://www.sciencedirect.com/science/article/pii/S0899825697905571`.

Halpern, J.Y. and Pass, R., 2011, "Iterated regret minimization: A new solution concept", *Games and Economic Behavior*.

Harsanyi, J.C., 1967-68, "Games with incomplete information played by 'bayesian' players, I - III", *Management Science*, 14: 159–182, 320–334, 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. and Samet, D., 1998, "Knowledge spaces with arbitrarily high rank", *Games and Economic Behavior*, 22(2): 260–273.

Heifetz, Aviad, Meier, Martin, and Schipper, Bernhard, 2006, "Interactive unawareness", *Journal of Economic Theory*, 130: 78 – 94.

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

Hendricks, Vincent and Symons, John, 2009, "Epistemic logic", in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed., spring 2009 ed.

Huber, Franz, 2009, "Formal representations of belief", in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed., summer 2009 ed.

Joyce, J.M., 2004, "Bayesianism", in *The Oxford Handbook of Rationality*, A.R. Mele and P. Rawling, eds., Oxford University Press.

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

Kadane, J.B. and Larkey, P.D., 1983, "The confusion of is and ought in game theoretic contexts", *Management Science*, 1365–1379.

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

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

Kuhn, Harold, 1953, "Extensive games and the problemof informaiton", in *Contributions to the Theory of Games, Vol. II*, H. Kuhn and A. Tucker, eds., Princeton University Press.

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

Leyton-Brown, Kevin and Shoham, Yoav, 2008, *Essentials of Game Theory: A Consice, Multidisciplinary Introduction*, Morgan & Claypool.

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

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

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

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

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

Mas-Colell, Andreu, Winston, Michael, and Green, Jerry, 1995, *Microeconomic Theory*, Oxford University Press.

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

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

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

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

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

Morris, Stephen, 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, URL `http://econpapers.repec.org/RePEc:uto:cesmep:200903`.

Myerson, R.B., 1991, *Game Theory: Analysis of Conflict*, Harvard University Press, 1997 ed.

Osborne, Martin, 2003, *An Intrdocution to Game Theory*, Oxford University Press.

Osborne, M.J. and Rubinstein, A., 1994, *A Course in Game Theory*, MIT Press.

Pacuit, E. and Roy, O., 2011, "A dynamic analysis of interactive rationality", in *Proceedings of the Third International Workshop on Logic, Rationality and Interaction*, H. van Ditmarsch, J. Lang, and Shier Ju, eds., vol. 6953 of *Lecture Notes in AI*, 244–258.

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

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

Perea, Andres, 2012, *Epistemic Game Theory: Reasoning and Choice*, Cambridge UP.

Peterson, Martin, 2009, *An Introduction to Decision Theory*, Cambridge University Press.

Piccione, Michele and Rubinstein, Ariel, 1997a, "The absent-minded driver's paradox: Synthesis and responses", *Games and Economic Behavior*, 20(1): 121 – 130, URL `http://www.sciencedirect.com/science/article/pii/S0899825697905790`.

Piccione, Michele and Rubinstein, Ariel, 1997b, "On the interpretation of decision problems with imperfect recall", *Games and Economic Behavior*, 20(1): 3–24, URL `http://ideas.repec.org/a/eee/gamebe/v20y1997i1p3-24.html`.

Rabinowicz, Wlodzimierz, 1992, "Tortous labyrinth: Noncooperative normal-form games between hyperrational players", in *Knowledge, Belief and Strategic Interaction*, Cristina Bicchieri and Maria L. D. Chiara, eds., 107 – 125.

Ross, Don, 2010, "Game theory", in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed., fall 2010 ed.

Roy, O. and Pacuit, E., 201X, "Substantive assumptions in interaction: A logical perspective", To appear in *Synthese*.

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

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

Schwitzgebel, Eric, 2010, "Belief", in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed., winter 2010 ed.

Selten, R., 1975, "Reexamination of the perfectness concept for equilibrium points in extensive games", *International Journal of Game Theory*, V4(1): 25–55, URL `http://dx.doi.org/10.1007/BF01766400`.

Shoham, Yoav and Leyton-Brown, Kevin, 2008, *Multiagent Systems*, Cambridge University Press.

Siniscalchi, Marciano, 2008, "Epistemic game theory: Beliefs and types", in *The New Palgrave Dictionary of Economics*, S. Durlauf and L. Blume, eds., Basingstoke: Palgrave Macmillan.

Sophn, Wolfgang, 1982, "How to make sense of game theory", in *Philosophy of Economics: Proceedings*, W. Stegmueller et al., ed., Munich, 239–270.

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

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

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

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

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

Stuart Jr., Harborne W. and Hu, Hong, 2002, "An epistemic analysis of the harasanyi transformation", *International Journal of Game Theory*, 30(4): 517 – 525.

Tan, Tommy Chin-Chiu and Werlang, Sérgio Ribeiro da Costa, 1988, "The bayesian foundations of solution concepts of games", *Journal of Economic Theory*, 45(2): 370 – 391, URL `http://www.sciencedirect.com/science/article/pii/0022053188902761`.

Trost, Michael, 2009, "Solutions of strategic games under common belief of sure-thing principle", in *Proceedings of TARK XII (2009)*, Aviad Heifetz, ed.

Ullmann-Margalit, Edna and Morgenbesser, Sydney, 1977, "Picking and choosing", *Social Research*, 44: 757 – 785.

van Benthem, J., 2003, "Rational dynamic and epistemic logic in games", in *Logic, Game Theory and Social Choice III*, S. Vannucci, ed., University of Siena, department of political economy. An updated version of this paper is now available on `http://staff.science.uva.nl/~johan/RatDyn.2006.pdf`. The page numbering comes from this version.

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

van Benthem, Johan, 2010, *Modal Logic for Open Minds*, CSLI Publications.

van Benthem, Johan, 2011, *Logical Dynamics of Information and Interaction*, Cambridge University Press.

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

van Benthem, Johan, Girard, Patrick, and Roy, Olivier, 2009, "Everything else being equal: A modal logic for *ceteris paribus* preferences", *Journal of Philosophical Logic*, 38: 83–125.

van der Hoek, Wiebe and Pauly, Marc, 2007, "Modal logic for games and information", in *Handbook of Modal Logic*, P. Blackburn, J. van Benthem, and F. Wolter, eds., Elsivier, vol. 3 of *Studies in Logic*.

Vanderschraaf, Peter and Sillari, Giacomo, 2009, "Common knowledge", in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, ed., spring 2009 ed.

Zvesper, Jonathan, 2010, *Playing with Information*, Ph.D. thesis, ILLC, University of Amsterdam. ILLC DS-2010-03.

10 Other Internet Resources

- Put URL for website 1 here
- Put URL for website 2 here

11 Related Entries

Game Theory — Bayesian Epistemology — Common Knowledge — Game Theory and Ethics — Prisoner's Dilemma — Formal Representations of Belief