Epistemic Game Theory
Prepared for *Handbook of Rationality*

Andrés Perea
Maastricht University

January 16, 2018

Abstract

In this chapter we review some of the most important ideas, concepts and results in *epistemic game theory,* with a focus on the central idea of *common belief in rationality.* We start by showing how belief hierarchies can be encoded by means of epistemic models with types, and how this encoding can be used to formally define common belief in rationality. We next indicate how the induced choices can be characterized by a recursive elimination procedure, and how the concept relates to Nash equilibrium. Finally, we investigate how the idea of common belief in rationality can be extended to dynamic games, by looking at several plausible ways in which players may revise their beliefs.

1 Introduction

Classical game theory, pioneered by the seminal work of von Neumann (1928), von Neumann and Morgenstern (1944) and Nash (1950, 1951), is mainly concerned with the *choices* of players in a game, and often leaves the reasoning preceeding such choices as a black box. The purpose of *epistemic game theory* is to open this black box by explicitly describing, and investigating, the reasoning that players may undertake before making a choice. As such, epistemic game theory is a *descriptive* theory that attempts to model various plausible ways of reasoning, without making any normative statements about the particular type of reasoning or choices that should be employed by people.

The end product of such reasoning can be described by *beliefs* that players have, about the choices of other players, but also about the beliefs that their opponents hold about the choices of others, and so on. These *belief hierarchies* constitute, in a sense, the language of epistemic game theory. Indeed, many contributions in epistemic game theory propose plausible restrictions that may be imposed on such belief hierarchies, or investigate the consequences that these restrictions have for the choices that players make in the game.

As with many disciplines in science, it is difficult to say when epistemic game theory really started off. Morgenstern (1935), more than eighty years ago, already stressed the importance of reasoning and belief hierarchies in economic analysis, but it took a long time before belief hierarchies structurally entered the analysis of human behavior in economic systems and games. A possible reason for this long delay lies in the complexity of a belief hierarchy. Despite being a very natural object, it is quite difficult to work with because it involves *infinitely many* layers.

The purpose of this chapter is to provide an overview of some of the most important ideas and results in epistemic game theory, with a focus on the central reasoning concept of *common belief in rationality.* The outline is as follows. In Section 2 we show how infinite belief hierarchies in static games can conveniently be encoded by means of epistemic models with types, and use it in Section 3 to formally define common belief in rationality. In Section 4 we present a recursive elimination procedure that characterizes the choices that can rationally be made under common belief in rationality. In Section 5 we discuss the epistemic gap between common belief in rationality and the famous notion of Nash equilibrium. In Section 6, finally, we discuss how the idea of common belief in rationality can be extended to dynamic games.

For a more comprehensive overview of epistemic game theory, the reader is referred to the overview paper by Brandenburger (2007), the textbook by Perea (2012), the handbook chapter of Dekel and Siniscalchi (2015), the encyclopedia entry by Pacuit and Roy (2015) and the forthcoming book by Battigalli, Friedenberg and Siniscalchi (2018).

2 Belief Hierarchies and Types

The central idea in epistemic game theory is that of *common belief in rationality.* Informally, it states that you do not only choose rationally yourself, but also believe that your opponents will choose rationally, that your opponents believe that the other players will choose rationally, and so on. Most other reasoning concepts in epistemic game theory may be viewed as refinements, or variants, of common belief in rationality. The intuitive idea of common belief in rationality is already present in Spohn (1982) and in the concept of rationalizability (Bernheim (1984) and Pearce (1984)), although the latter two papers do not formally define the notion.

An important question is how the idea of common belief in rationality can be defined formally. Consider a *finite static game* $G = (C_i, u_i)_{i \in I}$, where I is the finite set of players, C_i the finite set of choices for player i, and $u_i : \times_{j \in I} C_j \to \mathbb{R}$ player i's utility function. When we say that player i believes in the opponents' rationality, we mean that player i believes that every opponent j chooses optimally, given what player i believes that player j believes about his opponents' choices. For this to be formally defined we need to specify i's belief about j's choice – a *first-order* belief – together with i's belief about j's belief about his opponents' choices, which is a *second-order* belief. Similarly, to formally define that player i believes that player j believes in his opponents' rationality, we need to additionally specify the belief that i holds about the belief that j holds about the belief that every opponent k holds about the other players' choices, which is a *third-order* belief. By continuing in this fashion we see that a formal definition of common belief in rationality requires, for a given player i, a first-order belief about

	e	f	g	h
a	$0, 0$	$4, 1$	$4, 4$	$4, 3$
b	$3, 2$	$0, 0$	$3, 4$	$3, 3$
c	$2, 2$	$2, 1$	$0, 0$	$2, 3$
d	$1, 2$	$1, 1$	$1, 4$	$0, 0$

Table 1: A two-player game

the opponents' choices, a second-order belief about the opponents' first-order beliefs, a third-order belief about the opponents' second-order beliefs, and so on. Such infinite strings of beliefs are called *belief hierarchies,* and constitute, in a sense, the language of epistemic game theory.

An important practical problem with belief hierarchies is that these are *infinite* strings, making it impossible to write them down explicitly. In order to work with belief hierarchies we must thus find a way to encode these in an easy and compact way. One way to do so is by means of epistemic models with *types* – an idea that goes back to Harsanyi (1967–1968). The main idea is as follows: In a belief hierarchy, player i holds, for every opponent j, a belief about j's choice, j's first-order belief, j's second-order belief, and so on. That is, a belief hierarchy for player i specifies, for every opponent j, a belief about j's choice and j's belief hierarchy. If we replace the word "belief hierarchy" by "type", and formalize beliefs by probability distributions, we obtain the following definition of an epistemic model with types.

Definition 2.1 (Epistemic model with types) *Consider a finite static game $G = (C_i, u_i)_{i \in I}$. A finite epistemic model for G is a tuple $M = (T_i, b_i)_{i \in I}$ where T_i is the finite set of types for player i, and $b_i : T_i \to \Delta(C_{-i} \times T_{-i})$ is i's belief mapping which assigns to every type $t_i \in T_i$ a probabilistic belief $b_i(t_i) \in \Delta(C_{-i} \times T_{-i})$ on the choice-type combinations of i's opponents.*

In this definition we have used the following pieces of notation: For every finite set X, we denote by $\Delta(X)$ the set of probability distributions on X. By $C_{-i} \times T_{-i} := \times_{j \neq i}(C_j \times T_j)$ we denote the set of choice-type combinations for i's opponents.

A finite epistemic model may be viewed as a convenient way to encode belief hierarchies in a finite manner. Indeed, for every type in an epistemic model we may *derive* the full belief hierarchy it induces.

To see how this works, consider the two-player game in Table 1, where player 1's choices are in the rows and player 2's choices are in the columns, together with an epistemic model in Table 2. Please ignore the superscripts of the types for the moment. These will become clear later. The expression $b_1(t_1^c) = (0.6) \cdot (e, t_2^e) + (0.4) \cdot (f, t_2^g)$ means that type t_1^c assigns probability 0.6 to the event that player 2 chooses e and is of type t_2^e, and assigns probability 0.4 to the event that player 2 chooses f and is of type t_2^g.

Consider the type t_1^b. As t_1^b believes that, with probability 1, player 2 chooses e and is of type t_2^e, the induced first-order belief is that player 1 believes that, with probability 1, player

Types	$T_1 = \{t_1^a, t_1^b, t_1^c\}, \quad T_2 = \{t_2^e, t_2^g, t_2^h\}$
Beliefs for player 1	$b_1(t_1^a) = (g, t_2^g)$
	$b_1(t_1^b) = (e, t_2^e)$
	$b_1(t_1^c) = (0.6) \cdot (e, t_2^e) + (0.4) \cdot (f, t_2^g)$
Beliefs for player 2	$b_2(t_2^e) = (0.6) \cdot (c, t_1^c) + (0.4) \cdot (d, t_1^a)$
	$b_2(t_2^g) = (a, t_1^a)$
	$b_2(t_2^h) = (c, t_1^c)$

Table 2: An epistemic model for the game in Table 1

2 chooses e. Moreover, as player 2's type t_2^e has the belief $(0.6) \cdot (c, t_1^c) + (0.4) \cdot (d, t_1^a)$ about player 1, the second-order belief induced by type t_1^b is that player 1 assigns probability 1 to the event that player 2 assigns probability 0.6 to player 1 choosing c and probability 0.4 to player 1 choosing d. In a similar fashion we can derive the higher-order beliefs, and hence the full belief hierarchy, for the type t_1^b, and for all the other types in the epistemic model.

In the game theoretic literature people often use *infinite* instead of finite epistemic models, because they wish to work with models that encode *all possible* belief hierarchies. Such exhaustive models are also called *terminal* type structures. That terminal type structures exist for every finite static game – something that is far from obvious – has been shown by Armbruster and Böge (1979), Böge and Eisele (1979), Mertens and Zamir (1985) and Brandenburger and Dekel (1993). For this chapter we have chosen to work with finite epistemic models instead, for two reasons. First, finite epistemic models are easier to work with than terminal type structures, since no advanced measure theoretic or topological machinery is needed. Moreover, as we will see, this choice does not affect the main results we discuss.

The game theoretic literature also uses alternative ways of encoding belief hierarchies, such as Kripke structures (Kripke (1963)) and Aumann structures (Aumann (1974, 1976)). The first is the predominant model in the logical and philosophical literature, whereas the latter is often used by economists. Both models use *states* instead of types, and assign to every state and every player i a choice for player i, together with a belief for player i about the states. In a similar way as above, one can then derive from such a structure a belief hierarchy for every player at every state. In this chapter we have chosen to encode belief hierarchies by means of types, but the complete chapter could have been written by using Kripke structures or Aumann structures instead.

3 Common Belief in Rationality

In the previous section we have seen that belief hierarchies, which are fundamental for the idea of common belief in rationality, can be encoded by means of epistemic models with types. This now enables us to provide a formal definition of common belief in rationality. We will do so step by step, starting from the first layer of common belief in rationality which states that player i believes that every opponent chooses rationally.

To express this event within the formalism of epistemic models with types, we must first define what it means for a choice to be optimal for a type. Consider an epistemic model $M = (T_i, b_i)_{i \in I}$ for a static game $G = (C_i, u_i)_{i \in I}$, a type $t_i \in T_i$, and a choice $c_i \in C_i$. Then,

$$u_i(c_i, t_i) := \sum_{(c_{-i}, t_{-i}) \in C_{-i} \times T_{-i}} b_i(t_i)(c_{-i}, t_{-i}) \cdot u_i(c_i, c_{-i})$$

denotes the expected utility for type t_i of choosing c_i. We say that choice c_i is *optimal* for type t_i if $u_i(c_i, t_i) \geq u_i(c_i', t_i)$ for all $c_i' \in C_i$. In the epistemic model of Table 2, it can be verified that a is optimal for the type t_1^a, b is optimal for the type t_1^b and c is optimal for the type t_1^c. Similarly, e is optimal for player 2's type t_2^e, g is optimal for the type t_2^g and h is optimal for the type t_2^h.

Remember that a type t_i holds a probabilistic belief $b_i(t_i)$ on the opponents' choice-type combinations. For a type t_i to believe in the opponents' rationality means that $b_i(t_i)$ must only assign positive probability to opponents' choice-type pairs where the choice is optimal for the type.

Definition 3.1 (Belief in the opponents' rationality) *Consider a finite epistemic model $M = (T_i, b_i)_{i \in I}$ for a finite static game $G = (C_i, u_i)_{i \in I}$. A type $t_i \in T_i$ believes in the opponents' rationality if $b_i(t_i)((c_j, t_j)_{j \neq i}) > 0$ only if, for every opponent $j \neq i$, choice c_j is optimal for type t_j.*

In the epistemic model of Table 2 it can be verified that types t_1^a, t_1^b, t_2^g and t_2^h believe in the opponent's rationality, but the other two types do not. Indeed, the type t_1^c for player 1 assigns positive probability to player 2's choice-type pair (f, t_2^g) where f is not optimal for the type t_2^g, and hence t_1^c does not believe in 2's rationality. Similarly, player 2's type t_2^e assigns positive probability to player 1's choice-type pair (d, t_1^a) where d is not optimal for t_1^a, and hence t_2^e does not believe in player 1's rationality.

With the definition of belief in the opponents' rationality at hand we can now recursively define k-fold belief in rationality for all $k \geq 1$, which finally enables us to formalize common belief in rationality.

Definition 3.2 (Common belief in rationality) *Consider a finite epistemic model $M = (T_i, b_i)_{i \in I}$ for a finite static game $G = (C_i, u_i)_{i \in I}$.*

(Induction start) A type $t_i \in T_i$ expresses 1-fold belief in rationality if t_i believes in the opponents' rationality.

(Induction step) For $k > 1$, a type $t_i \in T_i$ expresses k-fold belief in rationality if $b_i(t_i)((c_j, t_j)_{j \neq i}) > 0$ only if, for every opponent $j \neq i$, type t_j expresses $(k-1)$-fold belief in rationality.

A type $t_i \in T_i$ expresses common belief in rationality if t_i expresses k-fold belief in rationality for every $k \geq 1$.

Hence, 2-fold belief in rationality entails that a type only assigns positive probability to opponents' types that express 1-fold belief in rationality. In other words, the player believes that every opponent believes in his opponents' rationality. Similarly, 3-fold belief in rationality corresponds to the event that a player believes that his opponents believe that their opponents believe in their opponents' rationality, and so on.

Within a finite static game $G = (C_i, u_i)_{i \in I}$, we say that player i can *rationally choose* $c_i \in C_i$ *under common belief in rationality* if there is a finite epistemic model $M = (T_i, b_i)_{i \in I}$ and a type $t_i \in T_i$ such that t_i expresses common belief in rationality and c_i is optimal for t_i. That is, choice c_i can be supported by some belief hierarchy that expresses common belief in rationality.

In the epistemic model of Table 2 it can be verified that types t_1^c and t_2^e do not express 1-fold belief in rationality, that types t_1^b and t_2^h express 1-fold but not 2-fold belief in rationality, and that types t_1^a and t_2^g express common belief in rationality. Consequently, player 1 can rationally choose a and player 2 can rationally choose g under common belief in rationality.

4 Recursive Procedure

Suppose that in a given static game we are interested in the choices that the players can rationally make under common belief in rationality. Is there an easy method to find these choices, without having to resort to epistemic models with types? That is the question that will be addressed in this section.

The key to answering this question is Lemma 3 in Pearce (1984), which we will reproduce below. To state the lemma formally, we need the following definitions. Consider a finite static game $G = (C_i, u_i)_{i \in I}$, a choice c_i, and a belief $b_i \in \Delta(C_{-i})$ about the opponents' choices. Then,

$$u_i(c_i, b_i) := \sum_{c_{-i} \in C_{-i}} b_i(c_{-i}) \cdot u_i(c_i, c_{-i})$$

denotes the expected utility of choice c_i under the belief b_i. Choice c_i is called *optimal in* G for the belief b_i if $u_i(c_i, b_i) \geq u_i(c_i', b_i)$ for all $c_i' \in C_i$. Choice c_i is called *strictly dominated* in G if there is some randomization $r_i \in \Delta(C_i)$ such that

$$u_i(c_i, c_{-i}) < \sum_{c_i' \in C_i} r_i(c_i') \cdot u_i(c_i', c_{-i}) \text{ for all } c_{-i} \in C_{-i}.$$

In the literature, such randomizations $r_i \in \Delta(C_i)$ are typically called *mixed strategies* or *randomized choices,* and are often interpreted as real objects of choice for player i. In this chapter, however, we assume that players do not randomize when making a decision, and these randomizations r_i are merely used as an auxiliary device to characterize choices that are optimal for some belief. The reason is that players are assumed to be expected utility maximizers, and hence a player can never increase his expected utility by randomizing over his choices.

Lemma 3 in Pearce (1984) can now be stated as follows.

Lemma 4.1 (Pearce (1984)) *Consider a finite static game $G = (C_i, u_i)_{i \in I}$ and a choice $c_i \in C_i$. Then, there is a belief $b_i \in \Delta(C_{-i})$ such that c_i is optimal in G for b_i, if and only if, c_i is not strictly dominated in G.*

This lemma can be used to characterize the choices a player can rationally make if he believes in his opponents' rationality. Let G^1 be the reduced game that remains if we eliminate, for every player, the choices that are strictly dominated in G. For a player to believe in the opponents' rationality thus means, by Lemma 4.1, that his belief is fully concentrated on opponents' choices in G^1. By applying Lemma 4.1 to the reduced game G^1 we thus conclude that, for every player, the choices he can rationally make if he believes in the opponents' rationality are exactly the choices in G^1 that are not strictly dominated in G^1. That is, these are the choices that survive *two* rounds of elimination of strictly dominated choices. In a similar vein it can be shown that the choices that can rationally be made if a player believes in his opponents' rationality, and believes that his opponents believe in their opponents' rationality (that is, if he expresses up to *two*-fold belief in rationality) are exactly the choices that survive *three* rounds of elimination of strictly dominated choices. By continuing in this fashion we arrive at the following elimination procedure.

Definition 4.1 (Iterated elimination of strictly dominated choices) *Consider a finite static game $G = (C_i, u_i)_{i \in I}$.*

(Induction start) Let $G^0 := G$ be the full game.

(Induction step) For every $k \geq 1$ let G^k be the reduced game that remains if we eliminate from G^{k-1} all choices that are strictly dominated in G^{k-1}.

A choice $c_i \in C_i$ survives iterated elimination of strictly dominated choices if c_i is in G^k for all $k \geq 1$.

By the argument above, we thus see that G^2 contains exactly those choices that can rationally be made if a player believes in the opponents' rationality. By iterating this argument, we conclude that, for every $k \geq 2$, the k-fold reduced game G^k contains exactly those choices that can rationally be made under some belief hierarchy that expresses up to $(k-1)$-fold belief in rationality. This argument already appears in Spohn (1982). In particular, the choices that survive the full procedure will be exactly those choices that can rationally be made under common belief in rationality. This leads to the following central result, which is based on Theorems 5.2 and 5.3 in Tan and Werlang (1988), and which Brandenburger (2014) has called the "fundamental theorem of epistemic game theory". Brandenburger and Dekel (1987) offer in Proposition 2.1 a similar result, characterizing common belief in rationality by "best reply sets" instead of an elimination procedure.

Theorem 4.1 (Fundamental theorem of epistemic game theory) *Consider a finite static game $G = (C_i, u_i)_{i \in I}$ and a choice $c_i \in C_i$. Then, c_i can rationally be made under common belief in rationality, if and only if, c_i survives iterated elimination of strictly dominated choices.*

The fundamental theorem would remain unaffected if we would use terminal type structures (hence, *infinite* epistemic models) instead of *finite* epistemic models to define common belief in rationality. To illustrate the procedure of iterated elimination of strictly dominated choices and the theorem above, consider the game G from Table 1. In the full game G, it is easily verified that player 1's choice d is strictly dominated by the randomization that assigns probability 0.5 to his choices a and $b,$ and that player 2's choice f is strictly dominated by the randomization that assigns probability 0.5 to his choices g and $h.$ No other choices are strictly dominated. Hence, G^1 is the game obtained after eliminating the choices d and f. Within the 1-fold reduced game G^1, player 1's choice c is strictly dominated by b (or rather, the randomization that assigns probability 1 to b), and player 2's choice e is strictly dominated by $h.$ Hence, G^2 is the game obtained from G^1 after eliminating the choices c and e. Finally, within G^2 player 1's choice b is strictly dominated by $a,$ and player 2's choice h is strictly dominated by $g.$ As such, only the choices a and g survive iterated elimination of strictly dominated choices, and hence, by Theorem 4.1, these are the only choices that can rationally be made under common belief in rationality.

5 Nash Equilibrium

For many decades, the concept of Nash equilibrium (Nash (1950, 1951)) has dominated the classical approach to game theory, inspiring many refinements such as perfect equilibrium (Selten (1975)) and proper equilibrium (Myerson (1978)) for static games, and subgame perfect equilibrium (Selten (1965)) and sequential equilibrium (Kreps and Wilson (1982)) for dynamic games. However, for a long time it remained unclear what epistemic conditions are needed for players to choose in accordance with Nash equilibrium. The purpose of this section is to investigate Nash equilibrium from an epistemic perspective, and to link it to the conditions of common belief in rationality that we have explored so far. Let us start by giving the definition of Nash equilibrium.

Definition 5.1 (Nash equilibrium) *Consider a finite static game $G = (C_i, u_i)_{i \in I}$. A Nash equilibrium in G is a tuple $(\sigma_i)_{i \in I}$, where $\sigma_i \in \Delta(C_i)$ for every player i, such that $\sigma_i(c_i) > 0$*

only if

$$\sum_{c_{-i}=(c_j)_{j\neq i}\in C_{-i}} \left(\prod_{j\neq i}\sigma_j(c_j)\right)\cdot u_i(c_i,c_{-i}) \geq \sum_{c_{-i}=(c_j)_{j\neq i}\in C_{-i}} \left(\prod_{j\neq i}\sigma_j(c_j)\right)\cdot u_i(c_i',c_{-i})$$

for all $c_i' \in C_i$.

In other words, a Nash equilibrium is a tuple of probability distributions on choices such that a choice only receives positive probability if it is optimal against the probability distributions on the opponents' choices. Traditionally, these probability distributions σ_i have been interpreted as conscious randomizations, or *mixed strategies,* by the players. A more recent approach, adopted by Spohn (1982), Aumann and Brandenburger (1995) and other authors, is to interpret σ_i as the (common) probabilistic belief that i's opponents have about i's choice, and this is also the interpretation we use here.

A Nash equilibrium $(\sigma_i)_{i\in I}$ induces, in a natural way, a belief hierarchy for player i in which his (first-order) belief about the opponents' choices is given by $(\sigma_j)_{j\neq i}$, his (second-order) belief about j's belief about his opponents' choices is given by $(\sigma_k)_{k\neq j}$, and so on. Such belief hierarchies are called *simple* in Perea (2012). Moreover, this belief hierarchy can be shown to express common belief in rationality, relying on the optimality conditions in a Nash equilibrium. To see this, consider the belief hierarchy for player i induced by a Nash equilibrium $(\sigma_i)_{i\in I}$. Then, player i only assigns positive probability to a choice c_j of player j if $\sigma_j(c_j) > 0$. By the optimality condition of Nash equilibrium, this is only the case if c_j is optimal against $(\sigma_k)_{k\neq j}$, which is what player i believes that player j believes about his opponents' choices. Altogether, we see that player i only assigns positive probability to c_j if c_j is optimal for player j, given what player i believes that player j believes about his opponents' choices. That is, with this belief hierarchy player i believes in j's rationality. In a similar vein it can be shown that with this belief hierarchy, induced by a Nash equilibrium, player i also believes that every opponent j believes in his opponents' rationality, and so on. Hence, every Nash equilibrium induces, for every player, a belief hierarchy that expresses common belief in rationality. We can thus say that Nash equilibrium implies common belief in rationality.

But is the other direction also true? Does common belief in rationality necessarily lead to Nash equilibrium? The answer, as we will see, is "no". Consider the two-player game in Table 3. It may be verified that all three choices can rationally be made under common belief in rationality. However, there is only one Nash equilibrium (σ_1, σ_2) in this game, where σ_1 assigns probability 1 to c and σ_2 assigns probability 1 to f. Hence, in this example Nash equilibrium imposes more restrictions than just common belief in rationality. But what are these extra restrictions?

To see this most clearly, consider the epistemic model, together with its graphical representation, in Figure 1. It may be verified that all types in the epistemic model express common belief in rationality. Moreover, the superscript of the types indicate the choice that is optimal

	d	e	f
a	$0,3$	$3,0$	$0,2$
b	$3,0$	$0,3$	$0,2$
c	$2,0$	$2,0$	$2,2$

Table 3: A two-player game

Types	$T_1 = \{t_1^a, t_1^b, t_1^c\},\ T_2 = \{t_2^d, t_2^e, t_2^f\}$
Beliefs for player 1	$b_1(t_1^a) = (e, t_2^e)$ $b_1(t_1^b) = (d, t_2^d)$ $b_1(t_1^c) = (f, t_2^f)$
Beliefs for player 2	$b_2(t_2^d) = (a, t_1^a)$ $b_2(t_2^e) = (b, t_1^b)$ $b_2(t_2^f) = (c, t_1^c)$

$$\begin{tikzcd}
\textbf{Player 1} & \textbf{Player 2} & \textbf{Player 1} \\
(a, t_1^a) \arrow[dr] & (d, t_2^d) \arrow[r] & (a, t_1^a) \\
(b, t_1^b) \arrow[ur] & (e, t_2^e) \arrow[r] & (b, t_1^b) \\
(c, t_1^c) \arrow[r] & (f, t_2^f) \arrow[r] & (c, t_1^c)
\end{tikzcd}$$

Figure 1: Epistemic model for the game in Table 3, and a graphical representation

for that type. Remember that only the choices c and f are supported by a Nash equilibrium in this game.

Consider the type t_1^a that supports the choice a – a choice that is not supported by a Nash equilibrium. The induced belief hierarchy states that, on the one hand, player 1 believes that player 2 chooses e but, on the other hand, believes that 2 believes that 1 believes that 2 chooses $d.$ That is, player 1 believes that player 2 is *incorrect* about 1's first-order belief. The same can be said about his type t_1^b. In contrast, the type t_1^c that supports the Nash equilibrium choice c induces a belief hierarchy in which 1 believes that 2 is *correct* about 1's first-order belief.

It turns out that in two-player games, this *correct beliefs assumption* – that is, that a player believes that his opponent is correct about his first-order belief – is exactly what separates common belief in rationality from Nash equilibrium. This is reflected in Spohn's (1982) theorem on page 253, and Aumann and Brandenburger's (1995) Theorem A, which both state that in two-player games, mutual belief in rationality, together with mutual belief in the actual first-order beliefs, leads to Nash equilibrium. Here, mutual belief in rationality means that player 1 believes in 2's rationality, and player 2 believes in 1's rationality. Similarly, mutual belief in the actual first-order beliefs means that player 1 is correct about 2's first-order belief, and player 2 is correct about player 1's first-order belief. From a one-person perspective (in which conditions are imposed on the belief hierarchy of a *single* player i) the Spohn-Aumann-Brandenburger conditions thus state that player i believes that j is rational, believes that j believes that i is rational, that i believes that j is correct about i's first-order belief, and that i believes that j believes that i is correct about j's first-order belief. In particular, Spohn, Aumann and Brandenburger show that the *first two layers* of common belief in rationality, together with the correct beliefs assumptions above, are enough to imply Nash equilibrium. Not all layers of common belief in rationality are needed. Polak (1999) shows, however, that if mutual belief in the actual first-order beliefs is strengthened to *common* belief in the actual first-order beliefs, then the Spohn-Aumann-Brandenburger conditions would imply common belief in rationality. Other epistemic foundations for Nash equilibrium in two-player games, which in some way or another involve the correct beliefs assumptions above, can be found in Tan and Werlang (1988), Brandenburger and Dekel (1989), Asheim (2006) and Perea (2007a). As the reasonability of the correct beliefs assumption can be debated – after all, why should an opponent be correct about your first-order belief? – these papers implicitly point at the problematic assumptions underlying Nash equilibrium.

For more than two players the above conditions are no longer enough to characterize Nash equilibrium. For such games, Nash equilibrium additionally implies that i's belief about j's choice must be stochastically independent from i's belief about k's choice, and that i's belief about j's belief about k's choice must be the same as i's belief about k's choice. The first property follows from the fact that in a Nash equilibrium $(\sigma_i)_{i\in I}$, the belief of i about the opponents' choices is given by the independent probability distributions $(\sigma_j)_{j\neq i}$, whereas the second condition is implied by the property that i's belief about j's belief about k's choice and i's belief about k's choice are both given by σ_k. These two conditions are not implied by

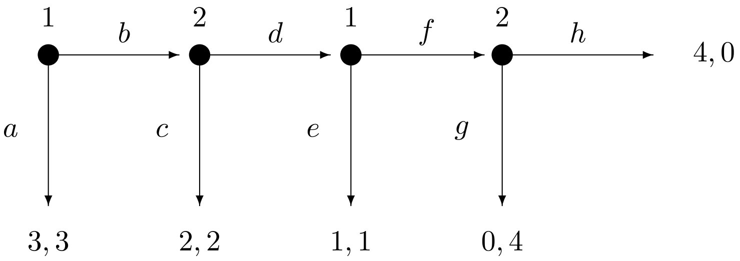

Figure 2: Reny's game

common belief in rationality, and hence the gap between Nash equilibrium and common belief in rationality is even bigger in games with more than two players. Epistemic foundations for Nash equilibrium in games with more than two players can be found in Brandenburger and Dekel (1987), Aumann and Brandenburger (1995), Perea (2007a), Barelli (2009) and Bach and Tsakas (2014).

6 Dynamic Games

So far we have been exploring *static* games, where all players only make one choice, and players choose in complete ignorance of the opponents' choices. We now investigate how the idea of common belief in rationality can be translated to *dynamic games*. In a dynamic game, players may choose one after the other, may choose more than once, and may fully or partially observe what the opponents have done in the past when it is their turn to move. As a consequence, a player may need to *revise* his belief about the opponents when he discovers that his previous belief has been contradicted by some of the opponents' past choices. As an illustration, consider the game from Figure 2 which is based on Reny (1992).

If player 1 believes that player 2 would rationally choose g at his last move, then he would choose a at the beginning. Common belief in rationality thus seems to suggest that player 2 should initially believe that player 1 chooses a. However, when it is player 2's turn to move, this initial belief has been contradicted by player 1's past play, and hence player 2 must revise his belief about player 1. But how? As we will see, there are at least two plausible ways for player 2 to revise his belief.

One option is to interpret player 1's past move b as a *mistake*, yet at the same time maintain the belief that player 1 would choose rationally at his second move, and maintain the belief that player 1 believes that would player 2 would rationally choose g at his second move. In that case, player 2 would believe, upon observing b, that player 1 would choose e at this second move, and therefore player 2 would choose c. This type of reasoning, in which the players are free to interpret "surprising" past moves as mistakes, but believe that the opponents will choose rationally in the future, believe that the opponents always believe that their opponents will choose rationally in the future, and so on, is called *backward induction reasoning,* and is formally captured by the concept of *common belief in future rationality* (Perea (2014)). Similar lines of reasoning are present in Penta (2015), Baltag, Smets and Zvesper (2006) and the concept of *sequential rationalizability* (Dekel, Fudenberg and Levine (1999, 2002) and Asheim and Perea (2005)). Backward induction reasoning is also implicitly present in the backward induction procedure (see Perea (2007b) for a survey on the various epistemic foundations for backward induction) and the equilibrium concepts of subgame perfect equilibrium (Selten (1965)) and sequential equilibrium (Kreps and Wilson (1982)) (see Perea and Predtetchinski (2017) for a formal statement).

Another option for player 2, after observing the "surprising" move b, is to interpret b as a conscious, optimal choice for player 1. However, this is only possible if player 2 believes that player 1 would choose f afterwards, and if player 2 believes that player 1 assigns a high probability to player 2 making the suboptimal choice h at his second move. Consequently, player 2 would choose d and, in case he is asked to make a second move, choose g. This type of reasoning, where a player, whenever possible, tries to interpret "surprising" past choices as conscious, optimal choices, is called *forward induction reasoning.* The concepts that most closely implement this type of reasoning are *extensive form rationalizability* (Pearce (1984), Battigalli (1997)), epistemically characterized by *common strong belief in rationality* in Battigalli and Siniscalchi (2002), and *explicable equilibrium* (Reny (1992)). See also Battigalli and Friedenberg (2012) who study forward induction with exogenous restrictions on the players' beliefs.

As the example above illustrates, backward induction and forward induction reasoning may lead to different strategy choices. Indeed, player 2 chooses c under backward induction reasoning, but would choose (d, g) under forward induction reasoning. However, both types of reasoning lead to the same outcome, which is the terminal history following a. Battigalli (1997) has shown that the latter is *always* true in dynamic games with perfect information without relevent ties, by proving that in every such game, the forward induction concept of extensive form rationalizability always uniquely leads to the backward induction *outcome.* This result is remarkable, as forward induction and backward induction represent two completely different lines of reasoning. The connection between these two lines of reasoning in general dynamic games is one of the many intriguing problems in epistemic game theory that need further exploration.

References

[1] Armbruster, W. and W. Böge, (1979), Bayesian game theory, in O. Moeschlin and D. Pallaschke (eds.), *Game Theory and Related Topics*, North-Holland, Amsterdam.

[2] Asheim, G.B. (2006), *The consistent preferences approach to deductive reasoning in games,* Theory and Decision Library, Springer, Dordrecht, The Netherlands.

[3] Asheim, G.B. and A. Perea (2005), Sequential and quasi-perfect rationalizability in extensive games, *Games and Economic Behavior* **53**, 15–42.

[4] Aumann, R.J. (1974), Subjectivity and correlation in randomized strategies, *Journal of Mathematical Economics* **1**, 67–96.

[5] Aumann, R.J. (1976), Agreeing to disagree, *Annals of Statistics* **4**, 1236–1239.

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

[7] Bach, C.W. and E. Tsakas (2014), Pairwise epistemic conditions for Nash equilibrium, *Games and Economic Behavior* **85,** 48–59.

[8] Baltag, A., Smets, S. and J.A. Zvesper (2009), Keep 'hoping' for rationality: a solution to the backward induction paradox, *Synthese* **169,** 301–333 (*Knowledge, Rationality and Action* 705–737).

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

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

[11] Battigalli, P. and A. Friedenberg (2012), Forward induction reasoning revisited, *Theoretical Economics* **7,** 57–98.

[12] Battigalli, P., Friedenberg. A. and M. Siniscalchi (2018), *Epistemic Game Theory: Reasoning about Strategic Uncertainty,* In progress.

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

[14] Bernheim, B.D. (1984), Rationalizable strategic behavior, *Econometrica* **52,** 1007–1028.

[15] Böge, W. and T.H. Eisele (1979), On solutions of bayesian games, *International Journal of Game Theory* **8**, 193–215.

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

[17] Brandenburger, A. (2014), *The Language of Game Theory: Putting Epistemics into the Mathematics of Games,* World Scientific Series in Economic Theory, Volume 5.

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

[19] Brandenburger, A. and E. Dekel (1989), The role of common knowledge assumptions in game theory, in *The Economics of Missing Markets, Information and Games,* ed. by Frank Hahn. Oxford: Oxford University Press, pp. 46–61.

[20] Brandenburger, A. and E. Dekel (1993), Hierarchies of beliefs and common knowledge, *Journal of Economic Theory* **59**, 189–198.

[21] Dekel, E., Fudenberg, D. and D.K. Levine (1999), Payoff information and self-confirming equilibrium, *Journal of Economic Theory* **89**, 165–185.

[22] Dekel, E., Fudenberg, D. and D.K. Levine (2002), Subjective uncertainty over behavior strategies: A correction, *Journal of Economic Theory* **104,** 473–478.

[23] Dekel, E. and M. Siniscalchi (2015), Epistemic game theory, in P. Young and S. Zamir (eds.), *Handbook of Game Theory,* Volume 4, North-Holland.

[24] Harsanyi, J.C. (1967–1968), Games with incomplete information played by "bayesian" players, I–III', *Management Science* **14**, 159–182, 320–334, 486–502.

[25] Kreps, D.M. and R. Wilson (1982), Sequential equilibria, *Econometrica* **50**, 863–894.

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

[27] Mertens, J.-F. and S. Zamir (1985), Formulation of bayesian analysis for games with incomplete information, *International Journal of Game Theory* **14**, 1–29.

[28] Morgenstern, O. (1935), Vollkommene Voraussicht und wirtschaftliches Gleichgewicht, *Zeitschrift für Nationalökonomie* **6**, 337–357. (Reprinted as "Perfect foresight and economic equilibrium" in A. Schotter (ed.) (1976), *Selected Economic Writings of Oskar Morgenstern,* New York University Press, pp. 169–183).

[29] Myerson, R.B. (1978), Refinements of the Nash equilibrium concept, *International Journal of Game Theory* **7,** 73–80.

[30] Nash, J.F. (1950), Equilibrium points in N-person games, *Proceedings of the National Academy of Sciences of the United States of America* **36,** 48–49.

[31] Nash, J.F. (1951), Non-cooperative games, *Annals of Mathematics* **54**, 286–295.

[32] Pacuit, E. and O. Roy (2015), Epistemic foundations of game theory, in Ed Zalta (ed.), *Stanford Encyclopedia of Philosophy.*

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

[34] Penta, A. (2015), Robust dynamic implementation, *Journal of Economic Theory* **160,** 280–316.

[35] Perea, A. (2007a), A one-person doxastic characterization of Nash strategies, *Synthese* **158**, 251–271 (*Knowledge, Rationality and Action* 341–361).

[36] Perea, A. (2007b), Epistemic foundations for backward induction: An overview, in J. van Benthem, D. Gabbay and B. Löwe (eds.), *Interactive Logic Proceedings of the 7th Augustus de Morgan Workshop, London. Texts in Logic and Games 1*, Amsterdam University Press, pp. 159–193.

[37] Perea, A. (2012), *Epistemic Game Theory: Reasoning and Choice,* Cambridge University Press.

[38] Perea, A. (2014), Belief in the opponents' future rationality, *Games and Economic Behavior* **83,** 231–254.

[39] Perea, A. and A. Predtetchinski (2017), An epistemic approach to stochastic games, Working paper.

[40] Polak, B. (1995), Epistemic conditions for Nash equilibrium, and common knowledge of rationality, *Econometrica* **67,** 673–676.

[41] Reny, P.J. (1992), Backward induction, normal form perfection and explicable equilibria, *Econometrica* **60,** 627–649.

[42] Selten, R. (1965), Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragezeit, *Zeitschrift für die Gesammte Staatswissenschaft* **121,** 301–324, 667–689.

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

[44] Spohn, W. (1982), How to make sense of game theory, in W. Stegmüller, W. Balzer and W. Spohn (eds.), *Philosophy of Economics,* Springer Verlag, pp. 239–270.

[45] Tan, T. and S.R.C. Werlang (1988), The bayesian foundations of solution concepts of games, *Journal of Economic Theory* **45,** 370–391.

[46] von Neumann, J. (1928), Zur Theorie der Gesellschaftsspiele, *Mathematische Annalen* **100**, 295–320. (Translated by Sonya Bargmann as "On the theory of games of strategy" in A.W. Tucker and R.D. Luce (eds.) (1959), *Contributions to the Theory of Games*, Volume IV, Princeton University Press, Princeton, NJ, pp. 13–43 (*Annals of Mathematics Studies* **40**)).

[47] von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.