International Game Theory Review
Vol. 16, No. 1 (2014) 1440001 (22 pages)

DOI: 10.1142/S0219198914400015

FROM CLASSICAL TO EPISTEMIC GAME THEORY*

ANDRÉS PEREA

EpiCenter & Department of Quantitative Economics
Maastricht University, P.O. Box 616
6200 MD Maastricht, The Netherlands
a.perea@maastrichtuniversity.nl

Received 26 October 2012
Revised 22 March 2013
Accepted 16 September 2013
Published 22 November 2013

In this paper, we give a historical overview of the transition from classical game theory to epistemic game theory. To that purpose we will discuss how important notions such as reasoning about the opponents, belief hierarchies, common belief, and the concept of common belief in rationality arose, and gradually entered the game theoretic picture, thereby giving birth to the field of epistemic game theory. We will also address the question why it took game theory so long before it finally incorporated the natural aspect of "reasoning" into its analysis. To answer the latter question we will have a close look at the earliest results in game theory, and see how they shaped our approach to game theory for many years to come.

Keywords: Epistemic game theory; history; reasoning; belief hierarchies; common belief; common belief in rationality.

Subject Classification: C72

1. Introduction

An important characteristic of human beings is that they *reason* before making a decision. Indeed, before we make a choice we typically think about the possible consequences, and we look for the choice that yields — at least in our expectation — the most favorable outcome. This reasoning aspect is even more prominent in *game theoretic* situations, in which the consequence of a choice also depends on the choices made by others. In such situations it is natural to reason about the possible choices

*This paper has been presented at the LOFT conference in Sevilla (2012), and at seminars at Maastricht University, Corvinus University of Budapest and the Institute for Advanced Studies in Vienna. I thank all audiences for their feedback. I would also like to thank two reviewers for their insightful remarks.

that our opponents may make. And in order to reason our way towards *sensible* predictions about the opponents' choices, it may be helpful to also reason about the possible *desires* and *beliefs* of our opponents. This naturally leads to the emergence of *belief hierarchies* which do not only describe what one believes about the others' choices and desires, but also what one believes about the beliefs that others have about their opponents' choices and desires, and so on.

However, it took game theory a very long time before it finally incorporated the aspect of reasoning into its analysis. The question that we wish to answer is *why*?

To answer this question we will first have a close look at the earliest results in game theory, and see how these shaped the classical approach to game theory — an approach that would set the research agenda for many decades to come. Afterwards we discuss how important epistemic notions such as belief hierarchies, common belief, and common belief in rationality, arose and how they slowly but surely provided an alternative to the classical approach. Our historical investigation starts and ends with a discussion of Oskar Morgenstern's view on game theory. He was one of the first persons to argue for models that explicitly deal with the reasoning of people about the choices and beliefs of their opponents. In a sense he was way ahead of his time with these ideas, and it was only many decades later that game theory really offered what he asked for — an approach in which the reasoning of people is at center stage. This approach is nowadays known as *epistemic game theory*.

While carrying out this historical investigation, I benefitted a lot from Brandenburger's (2010) historical overview of the origins of epistemic game theory, especially for the sections on Morgenstern's view and the early results in game theory. I am also grateful to Schwalbe and Walker's (2001) discussion of Zermelo's paper, which provided me with some new insights about Zermelo's work on chess. The overview papers on common belief and common knowledge by van Ditmarsch *et al.* (2009) and Cubitt and Sugden (2003) have been important for the section on common belief. Finally, the book by Leonard (2010) has given me some interesting insights into the lives of John von Neumann and Oskar Morgenstern, and the role they played in the creation of game theory.

2. Morgenstern's View

In an important paper from 1935, called "Perfect foresight and economic equilibrium", Oskar Morgenstern already stresses the importance of reasoning and belief hierarchies for economic analysis. As an illustration, he uses the following story:

> "Sherlock Holmes, pursued by his opponent, Moriarity, leaves London for Dover. The train stops at a station on the way, and he alights there rather than traveling on to Dover. He has seen Moriarity at the railway station, recognizes that he is very clever and expects that Moriarity will take a faster special train in order to catch him in Dover. Holmes' anticipation turns out to be correct. But what if Moriarity had been still more clever,

> had estimated Holmes' mental abilities better and had foreseen his actions accordingly? Then, obviously, he would have traveled to the intermediate station. Holmes, again, would have had to calculate that, and he himself would have decided to go on to Dover. Whereupon, Moriarity would again have "reacted" differently. ... One may be easily convinced that there lies an insoluble paradox." (Morgenstern, 1935)

Indeed, if we assume that both Sherlock Holmes and Moriarity hold *correct* beliefs about the opponent's choice, then no configuration of choices is sustainable. So, at the end one of the two persons must necessarily hold an inaccurate belief about the opponent's choice. Yet, in economic theory up to that time, the traditional assumption was that all agents *do* hold *correct* beliefs about all relevant events, including the choices of other agents. Morgenstern uses the Sherlock Holmes story to show that this assumption may not at all be realistic, and should actually be dropped. Morgenstern writes:

> "Assumptions of this kind, which the analysis of equilibrium must make, are substantially that all persons concerned correctly foresee the relevant events in the future, and this foresight has to include not only the change in objective data but also the behavior of all other persons. ... But can equilibrium really take place with a faulty, heterogenous foresight, however? ... From the whole exposition, it follows that the assumption of perfect foresight is to be cut out from economic theory." (Morgenstern, 1935)

But if the economic model should allow agents to hold *incorrect* beliefs about some events, then it also becomes important to describe what an agent believes about the beliefs of *other* agents about these events, and so on. That is, belief hierarchies come into play. Morgenstern already gives a hint how such belief hierarchies could be constructed:

> "The remedy would lie in analogous employment of the so-called Russell *theory of types* in logistics. This would mean that on the basis of the assumed knowledge by the economic subjects of theoretical tenets of Type I, there can be formulated higher propositions of the theory; thus, at least, of Type II. On the basis of information about tenets of Type II, propositions of Type III, at least, may be set up, etc." (Morgenstern, 1935)

From the above we may conclude that Morgenstern strongly argues in favor of models in which agents *reason* about the possible choices and beliefs of other agents. However, it is precisely this reasoning component that has received very little attention in game theory until recently. A natural question that arises is *why*? I think the answer lies in the nature of the first results in game theory, and how they have shaped our approach to game theory in the decades that followed. Let us therefore go back one century, and have a close look at these pioneering works.

3. Early Days

Zermelo (1913) published a theoretical paper on chess which is generally regarded as the first contribution to game theory. He was primarily interested in two questions: (a) Can we give a mathematical definition of a winning position for White? and (b) Can we determine an upper bound for the number of moves that White needs to win whenever he is in a winning position? Of course, the same questions can be posed for Black. As a by-result of this analysis, he showed that every position is either a winning position for White, or a winning position for Black, or a nonlosing position for both White and Black. That is, from every position either White can guarantee a win within finitely many moves, independently of Black's strategy, or Black can guarantee a win within finitely many moves, independently of White's strategy, or both White and Black can guarantee at least a draw, independently of the opponent's strategy. Within game theory, this latter result has become known as Zermelo's theorem.

However, contrary to what many game theorists believe (including myself until recently), Zermelo did not use backward induction of any kind to prove his results on chess! In fact he could not, as he assumed no stopping rule for chess. That is, within his model the game of chess could potentially go on forever as long as no player reaches a win, and therefore there is no place to start the backward induction procedure from. This confusion probably arose because von Neumann and Morgenstern (1953), in the third edition of their book, *did* assume a stopping rule for chess and *did* use backward induction to prove Zermelo's theorem (see their Secs. 15.6 and 15.7). I think many people assumed that von Neumann and Morgenstern based their proof on Zermelo's proof, but that is not true. In fact, Schwalbe and Walker (2001) already warned us that Zermelo did not use backward induction, but somehow their warning did not receive the attention it deserved.

Some years later, Borel (1921, 1924, 1927) started to investigate a different class of recreational games, namely *symmetric two-person zero-sum games* involving chance. These are games where (a) the outcome depends both on chance and the skills of the players, (b) at every outcome the sum of the payoffs among the two players is zero, and (c) both players can choose from the same set of strategies, and their roles in the game are identical. Borel starts by iteratedly removing *bad* strategies from the game, which are strategies that yield an expected payoff of at most 0, no matter what the opponent does. So, we first eliminate all bad strategies for both players in the original game, which possibly yields a reduced game with fewer strategies. Within that reduced game we again remove all bad strategies for both players, and so on, until no bad strategies remain.

In Borel (1921) he shows that, if the final reduced game contains exactly three strategies for each player, then both players have a randomization over strategies that yields them an expected payoff of exactly zero, no matter what the opponent does in the reduced game. That is, both players can randomize in such a way that they are guaranteed to break even in expectation. In Borel (1924) he proves the

same result for the case where the final reduced game contains exactly five strategies for both players, whereas in Borel (1927) he conjectures the result to be true also for the case of seven remaining strategies. In these three works, Borel was the first to introduce the concept of a *randomized*, or *mixed* strategy, and to use the idea of iterated elimination of "unreasonable" strategies — an idea that later would play a prominent role within *epistemic game theory*.

Shortly after, von Neumann (1928) generalized Borel's results to the class of *all* two-person zero-sum games, including those that are not symmetric. More precisely, von Neumann showed that for every two-person zero-sum game, there is a unique number v such that (a) player 1 has a randomized strategy that guarantees him an expected payoff of at least v, no matter what player 2 does, and (b) player 2 has a randomized strategy that yields him an expected payoff of at least $-v$, no matter what player 1 does. In a symmetric game it is clear that v must be zero, and hence Borel's results can be derived from von Neumann's theorem. The number v above is called the *value* of the game, and the strategies for players 1 and 2 that guarantee them an expected payoff of at least v and $-v$, respectively, are called *maxmin* strategies for these players. The term *maxmin* is chosen because such a strategy *maximizes* the *minimum* expected payoff that can be achieved by playing this strategy.

If we look at the results by Zermelo, Borel and von Neumann we see a common pattern, namely that all these results focus on strategies that *guarantee* a player a certain minimum outcome, irrespective of what the opponent does. We call this the *maxmin approach* to games. Note that this approach is basically free of any reasoning about the opponent, because it is interested in outcomes that can be guaranteed by a player even if he has no clue about the opponent's choice. Indeed, the maxmin-criterion makes no distinction between *more reasonable* and *less reasonable* choices by the opponent, but simply looks at the "worst" strategy that the opponent could choose for you, no matter whether this strategy is plausible or not.

In 1944, when von Neumann and Morgenstern published their seminal book on game theory, it was von Neumann's maxmin approach that dominated the book, and not Morgenstern's concern for reasoning and belief hierarchies. Also in the first decade after the publication of their book it was the maxmin approach that would set the research agenda in game theory for many years to come. We can only speculate about the reasons for this phenomenon, but very likely game theory would have evolved in a rather different direction if Morgenstern's views on reasoning and belief hierarchies were more prominently present at that time.

Also Nash's concept of *equilibrium*, presented in Nash (1950, 1951), can be seen as a product of the maxmin approach to games, as it yields precisely von Neumann's maxmin strategies when applied to two-person zero-sum games. Its original definition — stating that a player's strategy must be optimal *given* the opponents' strategies — suggests that players are somehow able to correctly foresee the strategies by their opponents. This makes it hard to place the concept of equilibrium

within a model of reasoning, because in such models it seems natural to allow players to have *incorrect* beliefs about the opponents' choices.

A more recent interpretation of Nash equilibrium is that the components in a Nash equilibrium do not represent the choices by the players, but rather the *beliefs* that players hold about their opponents' choices. See, for instance, Aumann and Brandenburger (1995). But also with this interpretation there is still a problem, namely that Nash equilibrium assumes that players can somehow correctly foresee the *beliefs* that the opponents hold about their opponents' choices. See Tan and Werlang (1988), Brandenburger and Dekel (1989), Aumann and Brandenburger (1995), Polak (1999), Perea (2007) and Bach and Tsakas (2012) for papers that basically make this point.

Despite these problems, there is probably no concept that has dominated game theory so strongly, and for such a long period, as Nash equilibrium. Indeed, after its invention Nash equilibrium has been at the very center of game theoretic research for many decades. The downside of it all is that the reasoning component did not really enter the game theory picture during all these decades, despite the early message by Oskar Morgenstern. In 1953, Maurice Fréchet — when commenting on the papers by Émile Borel — already points at the absence of this reasoning part from game theory by saying:

> "One can imagine other theories, directed to the same end, that would take into account, in each individual decision, the presumptions (of the individual who makes the decision) concerning the decisions of the other individuals." (Fréchet, 1953)

Summarizing, we can conclude that the *maxmin approach* advocated by von Neumann — to be continued by Nash's *equilibrium approach* — have strongly shaped our view on game theory for many decades, thereby preventing the reasoning approach from entering the picture during all those years. This, at least to me, is the reason why it took game theory such a long time to take the reasoning aspect seriously, and to explicitly incorporate it into its models.

4. Belief Hierarchies: The Type-Based Approach

As we saw, Morgenstern (1935) already stressed the importance of *belief hierarchies* for a more realistic economic analysis. Remember that a belief hierarchy does not only describe the belief a person has about the relevant parameters in the model and the opponents' choices, but also the belief he holds about the beliefs his opponents hold about these objects, and so on. A full description of a belief hierarchy thus involves *infinitely* many belief levels, which makes it a complicated construct to work with from a practical point of view. Morgenstern already gave a verbal description of how to model such an infinite belief hierarchy — see the third quote in Sec. 2 — but many years elapsed before a first formal definition of such a belief hierarchy was given in game theory.

To the best of my knowledge, the paper by Harsanyi in 1962 is the first to formally define an infinite belief hierarchy, although it does so for a very special setting. Harsanyi (1962) focuses on bargaining situations between two persons who may face uncertainty about the opponent's utility function. He informally introduces a belief hierarchy as follows:

> "In bargaining, and more generally in all nontrivial game situations, the behavior of a rational individual will depend on what he expects the *other* party will do. Party 1 will ask for the best terms he expects *party* 2 to accept. But party 1 will know that the terms party 2 will accept in turn depend on what terms party 2 expects *party* 1 to accept. Thus, party 1's behavior will depend on what may be called his *second-order* expectations, i.e., on party 1's expectations concerning party 2's *expectations* about party 1's behavior. These again will depend on party 1's *third-order* expectations, i.e., on his expectations concerning party 2's *second-order* expectations, etc." (Harsanyi, 1962)

Later on in that paper, Harsanyi formalizes such belief hierarchies. However, he restricts to a special type of belief hierarchies that contains *nonprobabilistic* beliefs only. That is, player 1 assigns probability 1 to one specific concession point by player 2, assigns probability 1 to one specific nonprobabilistic estimate that player 2 can hold about player 1's concession point, and so on. Here, a concession point represents the least favorable outcome that a player is willing to accept.

A few years later, Harsanyi (1967–1968) applied the idea of an infinite belief hierarchy to a much broader setting, namely to analyze the reasoning and behavior of players in the general class of games with *incomplete information*. These are games in which players may have uncertainty about some of the relevant characteristics of the game, such as the physical outcomes of the game, the opponents' utility functions or the opponents' sets of available choices. However, Harsanyi showed that each of these three types of uncertainty may eventually be reduced to uncertainty about the opponents' utility functions alone. Hence, the belief hierarchies that Harsanyi uses contain the belief that a player has about the opponents' utility functions, the belief he has about the opponents' beliefs about their opponents' utility functions, and so on, *ad infinitum*.

A major problem that needs to be tackled in order to put these belief hierarchies to work is how to *represent* such infinite belief hierarchies. Writing these down *explicitly* is not really an option, because it would require writing down *infinitely* many levels — an impossible task. According to Harsanyi, this problem has obstructed the research on games with incomplete information in an important way:

> "It seems to me that the basic reason why the theory of games with incomplete information has made so little progress so far lies in the fact that these games give rise, or at least appear to give rise, to an infinite regress

> in reciprocal expectations on the part of the players." (Harsanyi, 1967, Part I)

One of the most beautiful contributions of Harsanyi's (1967–1968) work is to show how such infinite belief hierarchies can be *encoded* in a simple and compact way. Here is the main idea. Every player i in Harsanyi's model is characterized by (a) a utility function a_i, and (b) an infinite belief hierarchy b_i. Here, we stick to Harsanyi's original notation. The pair (a_i, b_i) is called player i's *information vector*, or *attribute vector*, or *type*. In fact, Harsany uses all three terms at different places in the paper, but nowadays we mostly use the term *type*. The key insight is to see that a belief hierarchy for player i induces a belief about (a) the opponents' utility functions, and (b) the opponents' *belief hierarchies*. Indeed, a belief hierarchy specifies a belief about the opponents' utility functions, a belief about the opponents' first-order beliefs about their opponents' utility functions, a belief about the opponents' second-order beliefs, and so on. In short, it yields a belief about the opponents' utilities and the opponents' belief hierarchies. Remember that a *type* is combination (a_i, b_i) of player i's utility function and player i's belief hierarchy. Hence, we conclude that every type (a_i, b_i) of player i specifies a belief about the opponents' utility functions and belief hierarchies, and hence specifies a belief about the opponents' *types*!

So, Harsanyi proposes to specify for every player a number of types, and to specify for every type a utility function and some probabilistic belief about the opponents' types. Within this construction we can then *derive* for every type the *complete infinite belief hierarchy* it induces. Harsanyi's model thereby provides an extremely simple, yet beautiful, way to *encode* such complicated objects as infinite belief hierarchies. This construction turned out to be a milestone in the development of epistemic game theory, since game theorists could now work with infinite belief hierarchies in an easy and convenient way, without having to write them down explicitly.

The original model by Harsanyi has been designed to encode belief hierarchies that only involve beliefs about the opponents' utilities, beliefs about these beliefs, and so on. In subsequent years, Harsanyi's construction has been extended to allow for more *general* belief hierarchies, which also involve beliefs about the opponents' *choices*. Pioneering work in this direction has been done by Werner Böge and his colleagues at the University of Heidelberg in the seventies. In Böge and Eisele (1979), for instance, they define the notion of *systems of complete reflections* in Definition 2, which is very similar to Harsanyi's type model. Indeed, the set R in their definition plays the same role as the set of types in Harsanyi's construction. However, the crucial difference is that Böge and Eisele assign to every type not only a utility function and a probabilistic belief about the opponents' types — like Harsanyi does — but also a choice for every such type. Since a type holds a belief about the opponents' types, and every opponent's type is associated with a choice, a type in Böge and Eisele's model holds in particular a belief about the opponents'

choices. By iterating this argument, we see that a type in Böge and Eisele also holds a belief about the opponents' beliefs about their opponents' choices, and so on. As such, the model by Böge and Eisele is more general than Harsanyi's original construction as it allows us to model belief hierarchies about the opponents' utilities *and choices* — not only about the opponents' utilities. It should be noted, however, that the way of encoding infinite belief hierarchies in Böge and Eisele is essentially Harsanyi's. In that sense, it is quite surprising to see that Böge and Eisele do not refer to Harsanyi's work. We do not know the reasons for this. In a related work, Armbruster and Böge (1979) define the notion of an *oracle system* in Definition 4.1, which plays exactly the same role as a *system of complete reflections* in Böge and Eisele (1979). This paper, contrary to Böge and Eisele (1979), *does* refer to Harsanyi's work.

5. Belief Hierarchies: The State-Based Approach

Next to Harsanyi's *type-based approach* there is another prominent model in the literature that can be used to describe belief hierarchies, namely the *state-based model* developed by Kripke (1963) and Aumann (1976). The philosopher and logician Kripke (1963) presented his model as a semantics for modal logic, which describes what a person deems possible, what he deems possible about what other people deem possible, and so on, about some relevant states of affair. The economist and game theorist Aumann takes a purely set theoretic approach in his model, without any explicit reference to logic. Moreover, Aumann's model was intended to model what people *know*, what they know about what others know, and so on, about some relevant states of affair. Since knowledge can be viewed as a special case of Kripke's "deeming possible" operator, Kripke's model is a little bit more general than Aumann's. This statement has been made precise in Samet (1990), for instance, who shows that Aumann's model is obtained from Kripke's model by imposing some suitable restrictions on the "deeming possible" operator. We will come back to this issue below.

The main idea in Kripke and Aumann's approach is that we model the possible states of affair by a set of *states of the world*. Every state of the world describes a possible way how the states of affair could be, but a player may not know exactly what the real state of the world is. This uncertainty is modeled as follows: At every state of the world ω, we specify for each player the set of states he deems possible at ω. This set may or may not contain the true state ω. In fact, Kripke allows this set not to contain the true state of the world — and hence the person to be wrong about the state of affairs — whereas Aumann's model requires this set to contain the true state ω. More precisely, Samet (1990) shows that if we require the "deeming possible" operator to satisfy *positive introspection, negative introspection* and the *truth axiom*, then we obtain precisely Aumann's model. A similar result is shown in Bacharach (1985) who, unlike Samet, does not explicitly refer to Kripke's work.

Now, if we assume that every player is informed about the above ingredients of the model — including the "deeming possible" operator for each of the players — we can derive at every state ω, for every player, a complete hierarchy which describes (a) which states this person deems possible at ω, (b) what he deems possible at ω about the sets of states that his opponents deem possible, and so on. Namely, at ω he deems possible some set of states, and at each of these states every opponent deems possible a certain set of states as well, and hence he holds a certain belief about the sets of states that his opponents may deem possible. In this way, we can at every state ω derive, for every player i, an infinite hierarchy of beliefs, similarly to how we derived an infinite belief hierarchy for every *type* from Harsanyi's type based model.

Note, however, that the original models by Kripke and Aumann are non-probabilistic in nature, as they specify only which states a player deems possible, without assigning probabilities to these states. But their models can easily be extended by assigning at every state ω, for every player i, a probability distribution over the states he deems possible at ω. If we do so, then we can derive at every state an infinite probabilistic belief hierarchy for every player.

Like Harsanyi's model, also the models by Kripke and Aumann are very flexible in that they are capable to describe belief hierarchies about basically anything we want. Suppose, for instance, that we wish to describe belief hierarchies concerning the players' *utilities* in a game, like Harsanyi did. Then, we can assign to every state of the world a utility function for every player, just like Harsanyi assigned a utility function to every *type*. If we want to model belief hierarchies concerning the players utilities *and choices*, then the only thing we have to change is to assign to every state of the world a utility function *and choice* for every player. This is comparable to the type-based model by Böge and Eisele (1979) and Armbruster and Böge (1979) described above, which also models belief hierarchies concerning the players utilities and choices, by assigning to every *type* a utility function and a choice.

So we see that the type-based model by Harsanyi, and the state-based model by Kripke and Aumann, are very similar in nature, and the types in Harsanyi's model play essentially the same role as the states of the world in Kripke and Aumann's model. This statement has been made precise in Brandenburger and Dekel (1993), Tan and Werlang (1992) and Tsakas (2012) who show that every encoding of a belief hierarchy in one model can be mimicked by an equivalent encoding in the other model. In the game theoretic literature both the type-based approach and the state-based approach have been extensively used, but eventually they do exactly the same thing — namely to provide an easy and convenient encoding of infinite belief hierarchies.

To conclude this section we wish to mention that Aumann, in an interview published in "Epistemic Logic, 5 Questions", has said that Harsanyi played an important role in developing his model for representing knowledge (Hendricks and Roy, 2010).

6. Common Belief

With the introduction of belief hierarchies it became possible to make formal statements about what a player believes the others will do, about the beliefs that a player has about the beliefs that others have about what their opponents will do, and so on. A major task of epistemic game theory is to put some *plausible restrictions* on such belief hierarchies, as to distinguish *reasonable* from *less reasonable* belief hierarchies. Indeed, all concepts in epistemic game theory can be viewed as a collection of *conditions* on belief hierarchies, which eventually selects for every player a family of belief hierarchies he could plausibly hold if he were to reason in accordance with that concept. Most of these concepts assume *common belief*, or *common knowledge*, of a particular pattern of reasoning about the opponents, which means that every player reasons in this particular way about his opponents, every player believes that every player reasons in this way, and so on, *ad infinitum*.

A prominent example of such a concept is *common belief in rationality*, which will be discussed in more detail in the next section. Common belief in rationality states that every player believes that all of his opponents will choose rationally, that every player believes that every player believes that all of his opponents will choose rationally, and so on. Hence, it assumes *common belief* in the event that "players believe that their opponents choose rationally".

In order to define such concepts formally we must first have a precise definition of what we mean by *common belief* and *common knowledge*. It seems that the sociologist Friedell (1967, 1969) and the philosopher Lewis (1969) were the first to give a definition of common belief and common knowledge, followed by the economist and game theorist Aumann (1976). It is interesting to see that the same idea has been presented by people from three different disciplines. As all three authors use a rather different approach to common belief and common knowledge, we will now briefly discuss these approaches separately and see what the main differences and similarities are.

Friedell works within a syntactic framework of modal logic, and uses the term "common opinion" rather than "common belief". For a person A and an event x he defines Ax to be the event that "A believes x". For two persons A and B, the event ABx means that A believes that B believes x. Beliefs of higher order can be generated in a similar fashion. Friedell then defines the event $Co_{A,B}$, meaning that "x is a matter of *common opinion* between persons A and B", by

$$Co_{A,B} = \left(\bigcap_{i=1}^{\infty} (A \cap B)^i \right) x.$$

That is, both A and B believe x, both A and B believe that both A and B believe x, and so on, *ad infinitum*. This is what we usually call *common belief* in the event x.

In his papers, Friedell makes a distinction between *belief* and *knowledge*. A person may believe an event x which is in fact not true, but *knowing* an event x implies that x must be true. In this spirit, he defines *common knowledge* among

persons A and B of the event x to be the event where (a) x is a matter of *common opinion* between persons A and B, and (b) the event x is true.

However, Friedell goes much beyond merely *formalizing* the notion of common opinion, as he is also interested in real life situations that could *generate* common opinion between two persons of a certain event x. One such situation could be that person A believes x, and believes that the other person B is in the same cognitive position as A himself (Friedell, 1969, p. 31). Then, A will also believe that B believes x. But then, as A expects B to reason in precisely this way about A, person A will believe that B believes that A believes x. By continuing this argument, we eventually arrive at the event where there is common opinion between A and B of x. One could also think of a situation where A and B are in face-to-face contact and hear from a voice whose authority is not in question that "You both believe x" (Friedell, 1969, p. 32). This event would also lead to common opinion between A and B of x. Or, one could imagine a situation where there is eye-contact between A and B, which leads to common knowledge of the presence of both persons (Friedell, 1969, p. 34).

Similarly to Friedell, also Lewis (1969), in his book *Convention*, is interested in plausible situations that could *generate* common knowledge. In fact, Lewis uses one such situation as his *definition* for common knowledge! Indeed, stated within Friedell's terminology, Lewis' definition of common knowledge runs as follows: There is *common knowledge* between persons A and B of the event x if some other event y holds such that (1) both A and B have reason to believe that y holds, (2) event y indicates to A and B that both A and B have reason to believe that y holds, and (3) event y indicates to both A and B that event x holds (see Lewis, 1969, p. 56).

If this is the case, then by combining (1) and (3) it follows that both A and B have reason to believe that x holds. Moreover, by combining (1) and (2) we obtain that both A and B have reason to believe that A and B have reason to believe that y holds. This, together with (3), implies that A and B have reason to believe that A and B have reason to believe that x holds. Now, if we apply (2) to itself, then we conclude that y indicates to A and B that both A and B have reason to believe that A and B have reason to believe that y holds. This, in combination with (1) and (3), leads to the conclusion that both A and B have reason to believe that A and B have reason to believe that A and B have reason to believe that x holds. By continuing this argument, we see that the conditions (1), (2) and (3) finally lead to the event where A and B have reason to believe x, A and B have reason to believe that A and B have reason to believe x, and so on, *ad infinitum*.

The latter event, that follows from Lewis' definition of common knowledge, is similar to what Friedell has called *common opinion* between A and B of the event x. However, there is one important difference, namely that Lewis talks about the persons' *reasons to believe*, not their real beliefs, whereas Friedell speaks about the persons' *real* beliefs. But apart from this difference, one could argue that Lewis' definition of common knowledge provides a set of *sufficient conditions* for Friedell's notion of common opinion.

Unlike Friedell and Lewis, who defined their notions in a syntactic logical framework, Aumann (1976) used a *semantical* framework to give his definition of common knowledge. More precisely, Aumann used the Kripke–Aumann structure discussed in Sec. 5, with a set of possible states of the world describing all the relevant states of affair that could possibly be. An event corresponds to a set E of states in Aumann's model, namely the set of those states where the event is true. Remember that in a Kripke–Aumann structure, a person A deems at every state ω some set of states possible. Aumann says that at a given state ω there is *common knowledge* of an event E among two persons A and B, if at ω both A and B only deem possible states in E, if at ω both A and B only deem possible states at which A and B only deem possible states in E, and so on, *ad infinitum*. This corresponds exactly to Friedell's definition of common knowledge. However, in contrast with Friedell and Lewis, Aumann does not describe situations that would *generate* common knowledge among a group of persons.

In spite of the differences described above, the three definitions of common belief (and common knowledge) given in Friedell, Lewis and Aumann all share the same flavor – they describe situations in which all members of a group believe a certain event x, all members believe that all members believe x, and so on. This definition proved to be crucial for the development of concepts in epistemic game theory, as we shall see in the following section.

7. Common Belief in Rationality

In my opinion, the concept of *common belief in rationality* really constitutes the central idea in epistemic game theory. It states that all players in a game believe that their opponents choose rationally, that all players believe that all players believe that their opponents choose rationally, and so on, *ad infinitum*. So, in terms of *common belief* as discussed in the previous section, it can be expressed as the event that (a) players believe that their opponents choose rationally, together with (b) common belief in the event that "players believe that their opponents choose rationally".

This concept thus imposes restrictions on the belief hierarchy of a player in a game. However, it does not necessarily tell us *how* a player reasons his way towards such a belief hierarchy. In that sense, the concept only imposes restrictions on the *output* of the reasoning procedure by a player, not necessarily on the reasoning process itself.

Since most other concepts in epistemic game theory can be seen as some sharpening, or variant, of common belief in rationality, we may indeed say that the concept of common belief in rationality is the cornerstone of epistemic game theory. See Perea (2012) for a detailed overview of such concepts in epistemic game theory that are based on the idea of common belief in rationality.

The idea behind common belief in rationality certainly has some intuitive appeal, and it is therefore not surprising that this idea has been floating around in

the game theory literature for quite some time — starting from the late sixties — although it did so in different disguises as we shall see.

Friedell (1969), in his section on economics, already discusses the idea of common knowledge of rationality in games, although he is not very precise about the notion of rationality. Moreover, he restricts attention to a very specific setting, namely that of two-player zero-sum games where both players face uncertainty about the utilities in the game. But it is the first paper I am aware of that explicitly deals with the idea of common knowledge (or belief) of rationality.

The first papers to provide a *formal* definition of common belief in rationality in a general setting are, to the best of my knowledge, Böge and Eisele (1979) and Armbruster and Böge (1979). Indeed, it can be shown that the notion of a *system of complete reflections over* R^1 in Böge and Eisele (1979) (see their Definition 2) corresponds precisely to the idea of common belief in rationality. Similarly for the concept of an *oracle system* in Armbruster and Böge (1979) (see their Definition 4.1). Nevertheless, these papers have been largely overlooked in the literature when it comes to discussing common belief in rationality. The main reason, I believe, is that the authors use completely different names for the concept, and — even more importantly — that it is not so easy to deduce that their definitions actually correspond to common belief in rationality. Let us therefore scrutinize their definitions in some detail, and explain why they represent common belief in rationality — although in a somewhat different disguise than we are used to. For the sake of brevity, we restrict our attention to the notion of a *system of complete reflections over* R^1 in Böge and Eisele (1979). A similar story could be told for the notion of an *oracle system* in Armbruster and Böge (1979).

In Böge and Eisele's terminology, a *system of complete reflections over* R^1 consists of (a) a set R, representing the set of possible type combinations for the players (in the sense of Harsanyi), (b) a set R^0, containing the possible choices and utility functions for the players, and (c) a mapping ρ that assigns to every type combination in R a combination of choices and utility functions for the players in R^0, and that additionally assigns to every type combination in R, for every player, a probabilistic belief about the type combinations in R. So, the mapping ρ assigns to every type combination $r \in R$, and for every player i, some choice $c_i(r)$, some utility function $u_i(r)$, and some probabilistic belief $b_i(r)$ about the players' type combinations. In particular, the belief $b_i(r)$ induces a belief for player i about the opponents' choices, as every type combination yields a combination of choices for the opponents.

The key condition that Böge and Eisele impose is that for every type combination r, and every player i, the induced choice $c_i(r)$ must be optimal for player i, given his induced utility function $u_i(r)$, and given his belief about the opponents' choices induced by $b_i(r)$. Let us call this condition $(\star)$. (This corresponds to their condition (4) in Definition 2.) If condition $(\star)$ holds for every type combination r, then the construct above is called a *system of complete reflections over* R^1.

It can now be shown that condition $(\star)$ implies that every type in R expresses *common belief in rationality*. Take, namely, some type combination r in R and some player i. Then, by construction, player i's type in r only assigns positive probability to type combinations $\hat{r}$ in R. By condition $(\star)$ we have for every such type combination $\hat{r}$, and every opponent j, that the induced choice $c_j(\hat{r})$ is optimal for player j, given his induced utility function $u_j(\hat{r})$, and given his belief about the opponents' choices induced by $b_j(\hat{r})$. In other words, player i's type in r only assigns positive probability to opponents' types for which the induced choice is optimal, given his induced utility function and belief. That is, player i's type in r believes in his opponents' rationality. So, we see that at every type combination in R, every player believes in his opponents' rationality.

Since at every type combination r in R, every player only assigns positive probability to type combinations that are in R, it follows from our insight above that at every type combination in R, every player only assigns positive probability to type combinations where all players believe in their opponents' rationality. That is, at every type combination in R, every player believes that his opponents believe in their opponents' rationality. By continuing in this fashion, we conclude that at every type combination in R, every player expresses *common belief in rationality*. So, indeed, condition $(\star)$ — and hence Böge and Eisele's definition of a *system of complete reflections over* R^1 — implies common belief in rationality.

The main difficulty, however, is to see that our condition $(\star)$ is equivalent to Böge and Eisele's condition (4) in their definition of a *system of complete reflections over* R^1. This is far from easy, and this may have contributed to the fact that the literature has largely overlooked the paper by Böge and Eisele when it comes to discussing common belief in rationality.

Böge and Eisele (1979) and Armbruster and Böge (1979) are not only the first ones to give a formal definition of common belief in rationality, they also provide in their papers a *recursive procedure* that yields *all* choices that the players can rationally make under common belief in rationality. See Theorem 2 in Böge and Eisele (1979) and Example 6.2 in Armbruster and Böge (1979). Their procedure may be summarized as follows: In round 1 we start with the full set of choices for every player. At every further round k we select those choices for player i that are optimal for some probabilistic belief on the opponents' choices that have survived up to this round. In Theorem 2, Böge and Eisele (1979) prove that this procedure yields precisely those choices that the players can rationally make under common belief in rationality.

Some years after Böge and Eisele (1979) and Armbruster and Böge (1979), the papers by Bernheim (1984) and Pearce (1984) independently developed the concept of *rationalizability* which is equivalent to the idea of common belief in rationality. Perhaps somewhat surprisingly, both Bernheim and Pearce do not refer to the works by Böge and Eisele (1979) and Armbruster and Böge (1979). Moreover, both Bernheim and Pearce had a rather different motivation for their concept of

rationalizability than Böge and Eisele (1979) and Armbruster and Böge (1979), as they viewed it as a more basic and natural alternative to Nash equilibrium — a concept they heavily critize in their respective papers. Here are just a few quotes which illustrate this:

> "While analyses of Nash equilibria have unquestionably contributed to our understanding of economic behavior, it would be unreasonably optimistic to maintain that Nash "solved" the problem of noncooperative strategic choice. ... The notion of an equilibrium has little intrinsic appeal within a strategic context. ... The economist's predilection for equilibria frequently arises from the belief that some underlying dynamic process (often surpressed in formal models) moves a system to a point from which it moves no further. However, where there are no equilibrating forces, equilibrium in this sense is not a relevant concept." (Bernheim, 1984)

> "... as a criterion for judging a profile of strategies to be "reasonable" choices for players in a game, the Nash equilibrium property is neither necessary nor sufficient. ... The standard justifications for considering only Nash profiles are circular in nature, and make gratuitous assumptions about players' decision criteria or beliefs." (Pearce, 1984)

They then introduce the notion of *rationalizability* as a constructive answer to this critique about Nash equilibrium. Although the idea of *common belief in rationality* is very much present in their papers, both Bernheim and Pearce do not formalize this notion explicitly. But some quotes in these papers show that they really had the idea of common belief in rationality in mind as a motivation for rationalizability:

> "For strategic games in normal form, it is natural to proceed on the basis of two premises: (1) agents view their opponents' choices as uncertain events, and (2) all agents abide by Savage's axioms of individual rationality, and this fact is common knowledge (in the sense of Aumann). Rationalizability is the logical consequence of these two premises." (Bernheim, 1984)

> "The purpose of this section is to develop a solution concept for finite normal form games, based on three assumptions:
> Assumption (A1): When a player lacks an objective probability distribution over another player's choice of strategy, he forms a subjective prior that does not contradict any of the information at his disposal.
> Assumption (A2): Each player maximizes his expected utility relative to his subjective priors regarding the strategic choices of others.
> Assumption (A3): The structure of the game (including all participants' strategies and payoffs, and the fact that each player satisfies Assumptions (A1) and (A2)) is common knowledge." (Pearce, 1984)

Nevertheless, the definitions that Bernheim and Pearce give for rationalizability do not explicitly refer to common belief in rationality. In fact, Bernheim and Pearce both give a *different* definition of rationalizability, but it can be shown that these two definitions give rise to the same set of choices for every player, and can thus be viewed as equivalent.

Of these two definitions for rationalizability, Bernheim's is perhaps the one that comes closest to our modern formulation of common belief in rationality, since it uses belief hierarchies — although in a different way than we are used to today. His key concept is that of a *consistent system of beliefs*, which may be defined as follows.

A *system of beliefs* Σ consists of objects $(i, i_1, \ldots, i_k, D_{i_k})$ where $(i, i_1, \ldots, i_k)$ is a sequence of players, and D_{i_k} is a subset of choices for the last player, i_k, in this sequence. The interpretation is that i believes that i_1 believes that i_2 believes that ... that i_k chooses from D_{i_k}. We call this a kth-order belief for player i. Moreover, a system of beliefs Σ must have the property that for every sequence of players $(i, i_1, \ldots, i_k)$ there is exactly one subset of choices D_{i_k} such that $(i, i_1, \ldots, i_k, D_{i_k})$ is in Σ.

Take a kth-order belief $(i, i_1, \ldots, i_k, D_{i_k})$ for player i. For every player $j \neq i_k$, extend this belief to a $(k+1)$th-order belief $(i, i_1, \ldots, i_k, j, D_j)$. We say that the kth-order belief $(i, i_1, \ldots, i_k, D_{i_k})$ is *justified* by the $(k+1)$th-order beliefs $(i, i_1, \ldots, i_k, j, D_j)$ for every player $j \neq i_k$, if every choice c_{i_k} in D_{i_k} is optimal for some belief of player i_k about the opponents' choices which, for every opponent j, only assigns positive probability to choices in D_j. Bernheim calls a system Σ of beliefs *consistent*, if every kth-order belief in Σ is justified by $(k+1)$th-order beliefs in Σ, for every k.

Now, consider some consistent system of beliefs Σ. To say that a 1st-order belief (i, i_1, D_{i_1}) in Σ is justified by 2nd-order beliefs (i, i_1, i_2, D_{i_2}) in Σ, actually means that player i believes that opponent i_1 chooses rationally. Namely, player i only deems possible choices for i_1 in D_{i_1}, and each of these choices c_{i_1} in D_{i_1} is optimal for some belief of player i_1 that, for each of his opponents i_2, only assigns positive probability to choices in D_{i_2}. Moreover, every 2nd-order belief (i, i_1, i_2, D_{i_2}) in Σ that was used to justify the 1st-order belief (i, i_1, D_{i_1}), is in turn justified by 3rd-order beliefs in Σ. By using a similar argument as above, this implies that i also believes that i_1 believes that each of his opponents i_2 chooses rationally and so on. Hence, by continuing in this way, we see that a consistent system of beliefs, in the sense of Bernheim, actually gives rise to belief hierarchies that express *common belief in rationality*.

Bernheim then calls a choice c_j for player j *rationalizable* if there is a consistent system of beliefs Σ, and some belief $(i, i_1, \ldots, i_{k-1}, j, D_j)$ in Σ, such that c_j is in D_j. Actually, Bernheim's original definition is slightly different, but it can easily be seen that our definition is equivalent to Bernheim's.

Pearce (1984) defines rationalizability in a completely different, yet equivalent, way. He introduces a recursive elimination procedure, which is almost identical to

the procedure used by Böge and Eisele (1979) and Armbruster and Böge (1979) (see our discussion above), and calls a choice *rationalizable* if it survives this procedure. The only difference with the procedure by Böge and Eisele (1979) and Armbruster and Böge (1979) is that Pearce assumes that a player's belief about the opponents' choices must be *independent* across the opponents, in case there are more than two players, whereas Böge and Eisele (1979) and Armbruster and Böge (1979) do not impose this independence assumption. This independence condition is also assumed by Bernheim (1984). If we leave out the independence condition from rationalizability, then the resulting concept is often called *correlated* rationalizability.

So, what Böge and Eisele (1979) actually showed in their Theorem 2 is that the choices that can rationally be made under common belief in rationality are exactly the choices that correspond to correlated rationalizability. This also indicates that Bernheim's formulation of rationalizability — which basically resembles the idea of common belief in rationality — is actually equivalent to Pearce's *algorithmic* definition of rationalizability — which corresponds to the recursive procedure by Böge and Eisele and Armbruster and Böge.

A few years later, Aumann (1987) and Brandenburger and Dekel (1987) provide foundations for the concepts of *correlated equilibrium* and *correlated rationalizability*, respectively, by using a notion that is slightly *stronger* than common belief in rationality. Both papers model the players' belief hierarchies about choices through the state-based model by Kripke and Aumann, as discussed in Sec. 5. To every state of the world they assign a choice for every player, and a probabilistic belief for every player about the states he deems possible. The condition they impose is that at *every* state of the world ω, the choice prescribed for every player must be optimal, given his belief at ω about the opponents' choices. As a consequence, at every state a player only deems possible states at which his opponents choose optimally given their beliefs, only deems possible states at which their opponents only deem possible states at which their opponents choose optimally, and so on. That is, the conditions in Aumann (1987) and Brandenburger and Dekel (1987) *imply* common belief in rationality. But their conditions are actually a bit stronger than this, as they require optimality of the players' choices to hold at *every* state in the model, and not only at those states that are relevant for a specific belief hierarchy of a player. Sometimes this condition is called *universal rationality*, but Aumann, Brandenburger and Dekel use it to formalize the idea of *common belief in rationality*, which is really what they had in mind.

Around the same time, Tan and Werlang (1988) provide a definition of common belief in rationality — they actually call it common *knowledge* of rationality — by using Harsanyi's type-based model in Sec. 4. More precisely, they assume for every player a set of types, and assign to every type a probabilistic belief about the opponents' choices *and types*. In a similar way as explained in Sec. 4, we can then *derive* for every type a complete belief hierarchy about the players' choices.

Tan and Werlang define the concept of common belief in rationality *recursively*. They start by defining for every player i the set K_i^1, which is the set of types for player i which only assign positive probability to opponents' choice-type pairs (c_j, t_j) where the choice c_j is optimal for the type t_j, given the belief that the type t_j holds about the other players' choices. Intuitively, the set K_i^1 contains those types that *believe in the opponents' rationality*. For every $m \geq 2$, they recursively define K_i^m to be set of types for player i that only assign positive probability to opponents' types t_j that are in K_j^{m-1}. So, K_i^2 contains those types which only assign positive probability to opponents' types that believe in their opponents' rationality. That is, all types in K_i^2 believe that the opponents believe in their opponents' rationality, and so on. Finally, common belief in rationality is said to hold at a given type t_i for player i if type t_i belongs to K_i^m for all m. Within the literature that uses Harsanyi's type-based models, the definition by Tan and Werlang has become the standard definition of common belief in rationality.

So we see that the literature has offered various different formulations of the same idea of common belief in rationality, ranging from the definitions of a *system of complete reflections* and an *oracle system* in Böge and Eisele and Armbruster and Böge, through Bernheim's *consistent system of beliefs* to the more standard definitions by Aumann, Brandenburger and Dekel, and Tan and Werlang. The first three formulations are often overlooked in the literature, as they are not that easily recognizable as being representations of common belief in rationality, but this does not make these contributions less important.

Moreover, the various formulations of common belief in rationality have also given rise to some of the first theorems in epistemic game theory. As we already mentioned, Böge and Eisele (1979) showed in their Theorem 2 that the choices that can rationally be made under their formulation of common belief in rationality are exactly the choices that are given by their recursive elimination procedure. This is perhaps the first real theorem in epistemic game theory, characterizing the behavioral consequences of the fundamental epistemic concept of common belief in rationality. Similar results have been shown in Brandenburger and Dekel (1987) and Tan and Werlang (1988). Brandenburger and Dekel show, in their Proposition 2.1, that their formulation of common belief in rationality yields precisely those choices that are correlated rationalizable, which, we have seen, are precisely the choices that survive the Böge–Eisele recursive procedure. Tan and Werlang additionally assume independence, and show in their Theorems 5.4 and 5.5 that their definition of common belief in rationality, together with independence, gives precisely the original (uncorrelated) rationalizable choices by Bernheim and Pearce.

We may thus conclude that the various definitions of common belief in rationality have been fundamental for the development of epistemic game theory, as they did not only formalize a natural way of reasoning upon which most other concepts have been based, but also triggered some of the first theorems in epistemic game theory.

8. Morgenstern's View

Remember from the beginning of our story that Oskar Morgenstern, in the thirties of the previous century, already asked for models in which players may be wrong in their beliefs, and in which these players may also hold beliefs about the beliefs of their opponents. As natural as this idea may be, it took game theory many decades before it eventually developed, and integrated, such models. But after a very long and gradual process — in which notions such as belief hierarchies, common belief, common belief in rationality, and other epistemic notions slowly but surely entered the game theoretic picture — it seems that Morgenstern's view has now received the attention that it deserved. The field that arose from this process — *epistemic game theory* — has finally put Morgenstern's view at the place where it belongs — right at the center of game theory.

One important contribution of epistemic game theory is that it has established solid epistemic foundations for existing game-theoretic solution concepts such as rationalizability, Nash equilibrium, iterated elimination of weakly dominated choices, and many other concepts. Indeed, for each of these concepts, authors have singled out collections of epistemic assumptions that characterize the concept at hand in the following sense: if a player reasons in accordance with these epistemic assumptions, then the possible beliefs he can have, or the possible choices he can make, are precisely those given by the concept.

In my opinion, an important task for epistemic game theory in the future is to develop *new* game-theoretic concepts by first presenting new, natural collections of epistemic assumptions, and subsequently characterize the choices — or beliefs — that result from these. Some work has already been done in this direction, but there is still plenty of room for more. In particular, epistemic game theory could help to develop concepts that rely on elements of bounded rationality, or even irrationality, as to better accomodate the game-theoretic concepts to the observed behavior of people in laboratory experiments. The good news is that with epistemic game theory we finally have the tools that are necessary to successfully tackle these problems.

References

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

Aumann, R. J. [1976] Agreeing to disagree, *Ann. Stat.* **4**, 1236–1239.

Aumann, R. J. [1987] Correlated equilibrium as an expression of Bayesian rationality, *Econometrica* **55**, 1–18.

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

Bach, C. W. and Tsakas, E. [2012] Pairwise interactive knowledge and Nash equilibrium, Working paper, Maastricht University.

Bacharach, M. [1985] Some extensions of a claim of Aumann in an axiomatic model of knowledge, *J. Econ. Theor.* **37**, 167–190.

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

Böge, W. and Eisele, T. H. [1979] On solutions of bayesian games, *Int. J. Game Theor.* **8**, 193–215.

Borel, É. [1921] La théorie du jeu et les equations intégrales à noyau symétrique, *Comptes Rendus Hebdomadaire des Séances de l'Académie des Sciences* (*Paris*) **173**, 1304–1308 (in French). (Translated by Savage, L. J. as The theory of play and integral equations with skew symmetric kernels, *Econometrica* **21** (1953) 97–100).

Borel, É. [1924] *Eléments de la Théorie des Probabilités*, 3rd edn. (Hermann, Paris) (in French). (Pages 204–221 translated by Savage, L. J. as On games that involve chance and the skill of players, *Econometrica* **21**, 101–115).

Borel, É. [1927] Sur les systèmes de formes linéaires à déterminant symétrique gauche et la théorie générale du jeu, *Comptes Rendus Hebdomadaire des Séances de l'Académie des Sciences* (*Paris*) **184**, 52–54 (in French). (Translated by Savage, L. J. as On systems of linear forms of skew symmetric determinant and the general theory of play, *Econometrica* **21**, 116–117).

Brandenburger, A. [2010] Origins of epistemic game theory, in *Epistemic Logic: 5 Questions*, eds. Hendricks, V. F. & Roy, O. (Automatic Press, VIP), pp. 59–69.

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

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

Brandenburger, A. and Dekel, E. [1993] Hierarchies of beliefs and common knowledge, *J. Econ. Theor.* **59**, 189–198.

Cubitt, R. P. and Sugden, R. [2003] Common knowledge, salience and convention: A reconstruction of David Lewis' game theory, *Econ. Phil.* **19**, 175–210.

Fréchet, M. [1953] Commentary on the three notes of Émile Borel, *Econometrica* **21**, 118–124.

Friedell, M. F. [1967] On the structure of shared awareness, Working paper, University of Michigan.

Friedell, M. F. [1969] On the structure of shared awareness, *Behav. Sci.* **14**, 28–39.

Harsanyi, J. C. [1962] Bargaining in ignorance of the opponent's utility function, *J. Conflict Resolution* **6**, 29–38.

Harsanyi, J. C. [1967–1968] Games with incomplete information played by "bayesian" players, I–III, *Manage. Sci.* **14**, 159–182, 320–334, 486–502.

Hendricks, V. F. and Roy, O. (eds.) [2010] *Epistemic Logic: 5 Questions* (Automatic Press, VIP).

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 (in German).

Leonard, R. [2010] *Von Neumann, Morgenstern, and the Creation of Game Theory* (Cambridge University Press).

Lewis, D. K. [1969] *Convention* (Harvard University Press, Cambridge).

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

Nash, J. F. [1950] Equilibrium points in N-person games, *Proc. Nat. Acad. Sci. U.S.A* **36**, 48–49.

Nash, J. F. [1951] Non-cooperative games, *Ann. Math.* **54**, 286–295.

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

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

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

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

Samet, D. [1990] Ignoring ignorance and agreeing to disagree, *J. Econ. Theor.* **52**, 190–207.

Schwalbe, U. and Walker, P. [2001] Zermelo and the early history of game theory, *Games Econ. Behav.* **34**, 123–137.

Tan, T. and Werlang, S. R. C. [1988] The Bayesian foundations of solution concepts of games, *J. Econ. Theor.* **45**, 370–391.

Tan, T. and Werlang, S. R. C. [1992] On Aumann's notion of common knowledge: An alternative approach, *Revista Brasileira de Economia* **64**, 151–166 (in Portuguese).

Tsakas, E. [2012] Epistemic equivalence of lexicographic belief representations, Working paper, Maastricht University.

van Ditmarsch, H., van Eijck, J. and Verbrugge, R. [2009] Common knowledge and common belief, in *Discourses on Social Software, Texts in Logic and Games*, eds. van Eijck, J. & Verbrugge, R. (Amsterdam University Press), pp. 99–122.

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

von Neumann, J. and Morgenstern, O. [1944, 1953] *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ).

Zermelo, E. [1913] Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels, *Proc. Fifth Int. Congress of Mathematicians*, Vol. 2, pp. 501–504 (in German).