

Belief Revision and Rationalizability

Oliver J. Board¹
Brasenose College, Oxford
econojb@erml.ne.ox.ac.uk

1 Introduction

The Bayesian approach to non-cooperative game theory, pioneered by Bernheim [6] and Pearce [15], views games as Bayesian decision problems in the sense of Savage, where the uncertainty faced by the players is the strategy choices of their opponents. Accordingly it is assumed that each player has a prior over the strategy sets of the other players. But each player is also uncertain about the others' priors, and so must have a prior over the set of priors, and so on. So we need some representation of this infinite hierarchy of beliefs for the players, and this has been provided by the work of Mertens and Zamir [14]. In their construction of a 'universal type space', each state of the world describes not only the strategy choices but also the (first-order) beliefs of the players. These first-order beliefs over the space in turn generate second-order beliefs, that is, beliefs about others' first-order beliefs, and so on.

In this way, a game can be transformed into a Bayesian decision problem, and solution concepts can be derived axiomatically. There are two directions from which to approach this project. The first is to take existing solution concepts, and search for appropriate restrictions on the players' motivations and beliefs to justify these concepts. For instance, Tan and Werlang [19] proved that players who possessed common knowledge of rationality would play only strategies that survived the iterated deletion of strictly dominated strategies, and conversely, that any strategy surviving this process was compatible with common knowledge of rationality. Their work thus provides epistemic foundations for this solution concept. Similarly, Aumann [1] provided an epistemic characterization of correlated equilibrium, and more recently Blume, Brandenburger and Dekel [7] carried out the same task for perfect and proper equilibrium, and Aumann and Brandenburger [2] for Nash equilibrium.

The other way to approach the project is to start with the restrictions on the players, and examine what is implied about the way they will behave. In this way, new solution concepts may be generated, for example, the iterated deletion procedure of Dekel

¹I would like to thank Michael Bacharach, Matthias Hild, and Hyun Shin for many helpful discussions. Copies of the full version of this paper can be obtained from the author.

and Fudenberg [9] and Börgers [8], and Stalnaker's [17] notion of strong rationalizable equilibrium.

This paper takes the second path. In particular, we consider the implications of assuming that the players possess common belief in rationality in extensive form games. All the papers above consider only static games, and new problems for the formal representation of beliefs are raised when we move to a dynamic setting. We must model not only what the players believe at the start of the game, but also how they will revise these beliefs as the game progresses, their beliefs about this revision process, and so on. Furthermore, some of the paths the game might take will have been assigned zero probability by the players, and so the standard tool used by economists to calculate how a player's beliefs change when she receives new information, Bayesian updating, cannot be used. Both of these problems have to be dealt with if we are to extend the Bayesian approach to extensive form games.

2 Related literature

The infinite hierarchies of beliefs used to model interactive epistemic systems have been extended to the case of conditional probability systems by Battigalli [4] and to the case of lexicographic probability systems by Stalnaker [18]. The former are used to analyze belief revisions in extensive form games, and the latter to express the notion of lexicographic utility maximization, used to characterize normal form refinements such as trembling hand perfect equilibrium. This paper combines the two in a single framework. It shares with Stalnaker's model the advantage over Battigalli's conditional probability systems that it is explicit about the belief revision process, rather than just expressing the outcome of that process. In addition, it is possible to provide a direct link between semantic models of this kind and the syntax.²

Iterated deletion procedures similar to those constructed below have been analyzed by Dekel and Fudenberg [9], Gul [13], and Ben Porath [5], among others. For a detailed survey of this literature, see Dekel and Gul [10]. In most cases, the results obtained are special cases of one of our theorems.

3 Beliefs and belief revision

As we discussed in the introduction, in order to analyze rational play in extensive form games, it is crucial to have a precise model not only of the players' beliefs but also of the way these beliefs are revised as the game proceeds. Traditional theories of belief revision, such as Bayes' rule, have concentrated on modelling how beliefs change when new information is learned that is compatible with one's existing beliefs. But such

²See Dekel et al. [11] for an illustration of the importance of such a link.

a focus is too narrow for our purposes: in order to model counterfactual reasoning in games, we will need to know how beliefs change or would change in the event of surprises, when information is learned that contradicts what is currently believed. In this case, some of these existing beliefs must be given up, and the problem is that there is a multitude of ways to select just how this should be done.

Following Spohn [16], we represent an agent's belief revision theory by an ordinal conditional function (OCF), k ; from a set W of possible worlds into the class of ordinals, such that $k(w) = 0$ for some w . The basic idea is that the function k gives the plausibility of each possible world, with worlds assigned to lower ordinals considered more plausible than worlds assigned to higher ordinals. The set of worlds considered possible by the agent is simply the set of worlds assigned to 0. If new information is learned that is incompatible with what is currently believed, the revised epistemic state is represented by the set of most plausible worlds compatible with the new information.

The reason why we need an ordinal conditional function rather than just an ordering is that, in extensive form games, beliefs may need to be revised more than once: new information may be received at each round of the game. While an ordering can tell us how beliefs will be revised starting from the prior belief state, there is no satisfactory way of preserving the relative plausibility data to determine how they will subsequently be revised from the posterior state³. An ordinal conditional function gives us not just an ordering of the possible worlds, but also a measure of their relative distance, and this extra structure enables us to model iterated belief revision.

This OCF is extended to propositions (sets of possible worlds) as follows: $k(A) = \min \{k(w) : w \in A\}$. Hence, for an arbitrary proposition A ; the agent believes that A if and only if $\exists w : k(w) = 0 \wedge w \in A$; i.e. $k(A) = 0$. Note that $k(A) = 0$ does not necessarily mean that A is believed to be true, only that A is not believed to be false. Using this framework, we can distinguish between two propositions according to epistemic weight: we say that A is more plausible than B if and only if $k(A) < k(B)$ or $k(A) = 0 < k(B)$. The ability to distinguish even between propositions that are all believed (with probability one) will come in useful later when we introduce a refinement of the concept of rationality.

We now show formally how OCFs are used to define the belief revision process. First we define an auxiliary concept:

Definition 1 Let k be an OCF and A be a nonempty set of W : Then the A -part of k is that function $k(\cdot \upharpoonright A)$ defined on A for which for all $w \in A$; $k(w \upharpoonright A) = \min \{k(A) + k(w)\}$. For $B \subseteq W$ with $A \cap B \neq \emptyset$; we also define $k(B \upharpoonright A) = \min \{k(w \upharpoonright A) : w \in A \cap B\} = \min \{k(A) + k(B \setminus A)\}$.

³See Spohn [16] for more on this point.

⁴Left-sided subtraction of ordinals is defined as follows: let α and β be two ordinals with $\alpha \geq \beta$; then $\alpha - \beta$ is that uniquely determined ordinal γ for which $\alpha = \beta + \gamma$.

We are now in a position to describe how an agent's OCF changes when she receives new information.

Definition 2 Let k be an OCF, A be a nonempty set of W ; and \otimes an ordinal. Then $k_{A;\otimes}$ is the OCF defined by

$$k_{A;\otimes}(w) = \begin{cases} \frac{1}{2} k(w \mid A); & \text{if } w \in A \\ \otimes + k(w \mid A); & \text{if } w \notin A \end{cases}$$

We call $k_{A;\otimes}$ the $(A;\otimes)$ -conditionalization of k ; and it represents agent i 's revised OCF on receiving information A with firmness \otimes (a measure of the reliability of the information source). It is easy to show that $k_{A;\otimes}(A) = 0$ and $k_{A;\otimes}(\neg A) = \otimes$; hence A is believed in the revised state with firmness \otimes : It can be verified this process defined in this way satisfies the standard (AGM) axioms of belief revision.

3.1 Models of interactive belief revision

An extensive form game is a representation of dynamic strategic interaction, and as we noted in the introduction, a formal model of rational behavior in such a situation should specify players' beliefs at every stage in the game, and beliefs at every stage about beliefs at every stage, and so on. That is, it should specify how beliefs of every order are revised as the game progresses. To this end, we extend Spohn's model to the multi-agent setting.

Definition 3 An OCF-structure M is a tuple $\langle N; W; \{k_i^w\}_{i \in N, w \in W} \rangle$; where:

N is the set of agents;

W is a set of possible worlds;

k_i^w is an ordinal conditional function for agent i at world w :

An OCF-structure determines each agent's beliefs, and beliefs about others' beliefs, in much the same way as a Kripke structure, but in addition to the belief sets of each agent at every possible world, it gives us the agent's belief revision policy as well. So, at world w ; an agent i believes that A if and only if $\{x \in W : k_i^w(x) = 0\} \subseteq A$; and letting $K_i(A)$ denote the set of all worlds in which agent i believes that A ; at world w agent i believes that agent j believes that A if and only if $w \in K_i(K_j(A))$: To determine what agent i believes about what agent j would believe if he were to receive information $(B; \neg)$; we simply replace k_j^x for every world x by the $(B; \neg)$ -conditionalization of k_j^x ; and re-compute the operator K_j : We define a further OCF for each world as follows: for all $x \in W$; $k^w(x) = \min \{k_i^w(x)\}$: Then, the operator K represents the predicate "everyone believes that". Common belief, represented by the operator C , is then defined in the standard way (see e.g. [12]).

4 Models of extensive form games

The analysis of this paper is restricted to finite extensive form games with perfect recall and without chance moves. The formal description of extensive form games is well known, and we do not repeat it here. In contrast to the traditional definition, however, we assume that a player's strategy specifies what that player will do only at nodes that are consistent with her previous moves, rather than at all of her information sets. In a sense it is of no importance which definition we work with, in that it will not affect any of the results, but the broader definition leads to unnecessary duplication and untidier proofs.

For an extensive form game Γ played by a set N of players, we denote by S_i the set of pure strategies of player i ; and by S the set of pure strategy profiles. Similarly, we let H_i denote the set of information sets at which i is on move, and H the set of all information sets. The set of strategies of player i and the set of strategy profiles consistent with an information set $h \in H$ being reached we label $S_i(h)$ and $S(h)$ respectively. Conversely, the set of information sets consistent with a strategy s_i we denote by $H(s_i)$; and let $H_j(s_i) = H(s_i) \setminus H_j$. Given perfect recall, each player's information sets can be partially ordered without ambiguities according to the precedence relation between the respective nodes. We denote by $h^-(h)$ the information set for the player on move at h that is immediately prior to h ; if such an information set exists.

Finally, we let $G(\Gamma) = \{S_i; u_i; g_{i \in N}\}$ denote the normal form of Γ ; where u_i is i 's expected utility function, defined on the set of strategy profiles. Further, for every information set h ; we define the (normal form) subgame associated with h as follows:

Definition 4 For every $h \in H$; let the subgame associated with h ; denoted $G^h(\Gamma)$; be the tuple $\{S_i(h); u_i(h); g_{i \in N}\}$; where $u_i(h)$ is the restriction of u_i to $S(h)$:

We shall drop the Γ in parentheses when there is no risk of ambiguity.

We now describe our epistemic model. A model M for an extensive form game Γ is a tuple

$$D \quad n \quad o \quad E \\ W; a; \{k_i^w\}_{i \in N}^w; p_i; f_i$$

where

W is a nonempty finite set, the set of possible worlds;

$a \in W$ represents the actual world;

k_i^w is the ordinal conditional function of player i at world w ;

p_i is a measure on the set of possible worlds, which will be used to derive i 's probabilistic beliefs at each possible world;

$f_i : W \rightarrow S_i$ is a decision function, with the interpretation that $f_i(w)$ is the strategy that player i will carry out at world w :

$$\text{Let } f(w) = \langle f_i(w) \rangle_{i \in N}$$

Note that none of the facts about the structure of the game are included in the model. The implication is that all these facts are true at every possible world in the model, and hence are common knowledge among the players. So we are assuming that there is common knowledge that the game is one of complete information.

In addition, we assume that all models satisfy the following conditions:

(A) For all $w, x, y \in W$; if $f_i(w) = f_i(x)$ and $f_i(w) \neq f_i(y)$; and $f_{-i}(x) = f_{-i}(y)$; then $k_i^w(x) < k_i^w(y)$

(C) For each $s \in S$; there exists a possible world w such that $f(w) = s$:

The first condition ensures that each player knows her own action at every information set at which she is on move, and the second that we can model players (counterfactual) beliefs at every node in the game. The second condition is in fact important even in static games if we are to consider certain refinements of the standard definition of rationality as expected utility maximization. For instance, if we are to model the idea that the players assign strictly positive probability to every possible strategy of their opponents, (C) is essential, and a weaker version of (C) is required to capture Pearce's [15] notion of 'caution'.

The k_i^w functions represent the beliefs and belief revision policies of the players. Beliefs at every information set of the game for the player to move at that information set are derived from her previous beliefs by conditionalizing on the set of worlds consistent with that information set being reached. So $k_{i,h}^w$ is the $(f_w : f(w) = S(h))$ -conditionalization of k_i^w if $S(h)$ exists, or the $(f_w : f(w) = S(h))$ -conditionalization of k_i^w otherwise, where \otimes is some ordinal such that $\otimes > \max_{x \in W} k_{i,h}^w(x)$ or $\otimes > \max_{x \in W} k_i^w(x)$ respectively. This is to ensure that the actual observation of a move is believed with greater firmness than the prior belief about what the move would be, and that that belief will be retained unless subsequent contradictory evidence is received.

The measure functions, p_i ; do the work of prior probability functions, and from them each player's probability assignment at a given information set over the set of possible worlds (and hence over her opponents' strategy choices and beliefs) is obtained by conditionalizing on the set of worlds considered possible at that set. Note that we do not assume that players have a 'common prior'. In fact, the p_i functions should be seen merely as a mathematical tool for calculating probabilistic beliefs at each possible world, and not as representing prior beliefs in any literal sense.

We can now give a formal definition of rationality in extensive form games, and consider the implications of common belief in rationality.

5 Rationalizability and refinements

We start by defining rationality as expected utility maximization, and subsequently introduce two refinements, the first a decision-theoretic version of Selten's notion of perfection, and the second based on Pearce's concept of extensive-form rationalizability. For each definition, we consider what restrictions are placed on players' strategy choices by the assumption that there is common belief that they are rational in that sense.

5.1 Rationalizability

Rationality is most commonly defined by economists as expected utility maximization, so in the context of an extensive form game, we say that a player is rational if, at every information set consistent with her chosen strategy at which she is on move, that strategy maximizes expected utility given her beliefs at that information set. The formal definition is analogous to that employed by Aumann [1], extended to extensive form games. We let R_i denote the set of worlds in which player i is rational, R the set of worlds in which all the players are rational, and CR the set of worlds in which there is common belief in rationality.

We are now in a position to consider the impact of common belief in rationality. For any game Γ ; we say a strategy $s_i \in S_i$ is rationalizable if it is compatible with common belief in rationality. Formally,

Definition 5 For any game Γ ; a strategy $s_i \in S_i$ is rationalizable if there is some model of Γ such that $\omega \in CR$ and $f_i(\omega) = s_i$:

Note that this is a purely model-theoretic concept, and not an game-theoretic equilibrium concept, as originally defined by Bernheim [6]. Nevertheless, we intend our concept to provide an epistemic foundation for Bernheim's concept⁵, generalized to extensive form games.

The following theorem provides a characterization of the set of rationalizable strategies. First we define for each player i a set $D_i \subseteq S_i$ of strategies that survives a certain iterated elimination procedure, a generalization of iterated deletion of strictly dominated strategies to extensive form games: every strategy that is strictly dominated in any subgame for the player on move at that subgame is deleted in the first round, and the standard procedure is applied to what is left of the whole game. Recall that G^h denotes the subgame of Γ at h formed by restricting the set of strategy profiles to $S(h)$; those consistent with h being reached.

$$D_i^1 = \{s_i \in S_i : \text{there is no } h \in H_i(s_i) \text{ such that } s_i \text{ is strictly dominated in } G^h\};$$

For all m ; let $D^m = \bigcap_{i \in N} D_i^m$;

⁵Except that we permit the players to have correlated beliefs over their opponents' strategies.

For all m ; $D_i^{m+1} = \bigcap_{s_i \in D_i^m} s_i$: s_i is not strictly dominated given that $s_i \in D_i^m$;
 $D_i = \bigcap_{m=1}^{\infty} D_i^m$; and $D = \bigcap_{m=1}^{\infty} D^m$:

Theorem 1 For any game Γ ; a strategy s_i is rationalizable if and only if $s_i \in D_i$:

The intuition behind the theorem is straightforward. It follows from the definition of rationality that no rational player will play a strategy that is strictly dominated at any information set at which she is on move. This accounts for the first round of deletion. But we cannot apply iterated deletion in any of these subgames: unless all the players believed with positive probability at the start of the game that a particular subgame would be reached, there may no longer be common belief in rationality if that subgame is reached. And it is common belief that drives iterated deletion. Nevertheless, there is common belief at the start of the game, so we can apply iterated deletion to the game as a whole.

The results established by Tan and Werlang [19] for static games and Ben-Porath [5] for generic games of perfect information are special cases of Theorem 1.

5.2 Perfect rationalizability

The first refinement we consider makes use of stronger rationality requirement, called perfect rationality, which is based on lexicographic utility maximization. This concept was first presented by Blume et al. [7] to provide a decision-theoretic justification of Selten's trembling hand perfect equilibrium, and was also analyzed by Stalnaker [18]. Both these models, however, were static. Here we analyze the implications of common belief in perfect rationality in extensive form games.

The basic idea behind perfect rationality is that in cases where two or more strategies maximize expected utility, the agent should consider in choosing between them how she should act if she learned that she were in error about something. And if two actions are still tied, the tie-breaking procedure is repeated. To formalize this concept, we first note that the belief revision structure generated by our model can be used to analyze agents' hypothetical beliefs if they are in error at any particular point in time, as well as if they receive information in the future which shows them to have been in error, and hence to derive lexicographic probability systems similar to those used by Blume et al. The agent's 'primary theory' is given simply by her original epistemic state; her secondary theory is then obtained by conditionalizing on the event that the primary theory is wrong (i.e. none of the worlds considered possible is true), her tertiary theory by conditionalizing on the event that both primary and secondary theories are wrong, and so on. Hence, an agent is perfectly rational if she chooses one of those strategies which maximizes expected utility according to her primary theory, and among all those strategies, also maximizes expected utility according to her secondary theory, and so on.

We denote by PR_i the set of worlds in which agent i is perfectly rational, PR the set of worlds in which all the players are perfectly rational, and CPR the set of worlds in which there is common belief in perfect rationality. Perfect rationalizability is then defined in the same way as rationalizability.

The next task is to provide a characterization theorem for the set of perfectly rationalizable strategies, analogous to theorem 1. It turns out that only a slight modification of the iteration procedure is needed. First we eliminate every strategy that is weakly dominated in any subgame for the player on move at that subgame, and then we apply iterated deletion of strictly dominated strategies to what is left of the whole game. We label the set of strategies for i which survive this process P_i :

Theorem 2 For any game Γ ; a strategy s_i is perfectly rationalizable if and only if $s_i \in P_i$:

The reason why we delete weakly dominated strategies in the first round is that such strategies can never yield higher expected utility than the dominating strategy, and since every strategy profile of one's opponents consistent with the current information set is played with positive probability according to some theory⁶, there is a theory for which it will yield strictly lower expected utility. The intuition behind the rest of the deletion procedure is the same as for theorem 1.

Stalnaker's [18] characterization of normal form perfect rationalizability is an immediate corollary of theorem 2.

5.3 The Best Rationalization Principle

The second refinement is based on an idea, originally due to Pearce [15], that one's opponents' strategy choices should be interpreted as rational whenever possible. This idea has been formalized and analyzed extensively by Battigalli (see e.g. Battigalli [3]), and is summed up in his 'best rationalization principle':

A player should always believe that her opponents are implementing one of the 'most rational' (or 'least irrational') strategy profiles which are consistent with her information.

We represent this principle as a restriction on the players' belief revision schemes. Say that a player strongly believes something if she continues to believe it unless it becomes incompatible with the evidence. Then our assumption is that the players strongly believe that everyone is rational; subject to this constraint, they strongly believe that everyone believes that everyone is rational, and so on. If the players' beliefs meet this assumption,

⁶This is implied by assumption (C); that for each strategy profile, there is some possible world in which that profile is played.

then we say they satisfy the best rationalization principle. To formalize this concept, we first define the set of worlds which are compatible with the rationality of all the players, and, for all $j > 1$; the set of worlds which are compatible with j th-degree belief in mutual rationality. We can then say that agent i satisfies the best rationalization principle if, for she considers every world compatible with rationality strictly more plausible than every world not, and, for all $j > 1$; she considers every world compatible with rationality together with 1st- through j th-degree belief in rationality strictly more plausible than every world not.

Let the BRP_i denote the set of worlds in which i satisfies the best rationalization principle. Note that $BRP_i \subseteq R_i$; given our assumption (A) of own-act knowledge, so there is no need to impose an additional assumption of rationality. Finally, we let BRP be the set of worlds in which everyone satisfies the best rationalization principle, and $CBRP$ the set of worlds in which this is common belief.

The iterated deletion procedure that characterizes the set of best rationalizable strategies is somewhat different from the two previous algorithms. In each round we delete in every subgame every strategy that is strictly dominated in any subgame for the player on move at that subgame:

$$B_i^1 = \{s_i \in S_i : \text{there is no } h \in H_i(s_i) \text{ such that } s_i \text{ is strictly dominated in } G^h\};$$

$$\text{For all } m; \text{ let } B^m = \bigcap_{i \in N} B_i^m;$$

$$\text{For all } m; B_i^{m+1} = \{s_i \in B_i^m : \text{there is no } h \in H_i(s_i) \text{ such that } s_i \text{ is strictly dominated in } G^h \text{ given that } s_{-i} \in B_{-i}^m\};$$

$$B_i = \bigcap_{m=1}^{\infty} B_i^m; \text{ and } B = \bigcap_{m=1}^{\infty} B^m;$$

Theorem 3 For any game Γ ; a strategy s_i is best rationalizable if and only if $s_i \in B_i$:

The reason why we can apply iterated deletion in the subgames is that the best rationalization principle implies that there will be 'enough' rationality left, unless that degree of rationality is incompatible with the subgame being reached, in which case the entire subgame will have been deleted and an extra round of deletion makes no difference.

It is worth noting that in static games, every subgame is identical to the whole game, and best rationalizability collapses into rationalizability. In perfect information games, however, best rationalizability is a very powerful concept. The following corollary implies that it yields a unique prediction in a wide class of such games.

Corollary 1 In generic games of perfect information, each of the components of a strategy profile is best rationalizable if and only if that profile is path equivalent to a subgame perfect Nash equilibrium strategy profile.

The reason for this close connection between best rationalizability and subgame perfection is that they both correspond closely to backward induction. In non-generic games, however, they may fail to coincide.

6 Conclusions

We constructed a formal model of player's beliefs and the way they revise these beliefs in extensive form games by combining the standard (static) model of interactive epistemology with a belief revision structure. The model gives us a unified framework which can express both the lexicographic probability systems of Blume et al. [7] and the conditional probability systems of Battigalli [4], and hence we are able to generalize their results and those in related papers, providing epistemic characterizations of several new iterated deletion procedures. These procedures give us a surprising amount of predictive power in certain classes of games. This makes them easy to test empirically.

In addition to its generality, our model has the advantage that it can be extended using well-known techniques into a propositional model, in which the syntax and axioms governing that syntax are made explicit. This would help us to identify precisely what is being assumed about the beliefs, belief revision processes, reasoning power, and rationality of the players. One immediate result would be to show that we have not used the axioms of knowledge or positive and negative introspection (all implicit in the partitional model of differential information) to derive any of our results. To this extent, our model is one of bounded rationality.

There are several tasks for further research. The most obvious is to consider the impact of further refinements. In particular, one might combine perfect rationality with the best rationalization principle. We conjecture that this would give an algorithm identical to that used in theorem 3, except that weakly dominated strategies are deleted in each round. When applied to static games, this would provide an epistemic foundation for iterated deletion of weakly dominated strategies, a procedure which has often been considered difficult to justify. Another task is to extend the model to games of incomplete information. This could be done simply by introducing variables for the parameters of the game, and allowing them to vary across the state space (a direct formalization of the 'Harsanyi transformation'). The model could then be used to extend the results of Aumann and Brandenburger [2] and provide epistemic foundations for equilibrium concepts such as sequential equilibrium, and to justify refinements such as the intuitive criterion.

References

- [1] Aumann, R. J. (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55, 1-18.
- [2] Aumann, R. J. and A. Brandenburger (1995): "Epistemic Conditions for Nash Equilibrium", *Econometrica* 63, 1161-1180.
- [3] Battigalli, P. (1996): "Strategic Rationality Orderings and the Best Rationalization Principle", *Games and Economic Behavior* 13, 178-200.
- [4] Battigalli, P. (1997): "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games", mimeo, Princeton University.
- [5] Ben-Porath, E. (1997): "Rationality, Nash Equilibrium and Backwards Induction in Perfect-Information Games", *Review of Economic Studies* 64, 23-46.
- [6] Bernheim, B. D. (1984): "Rationalizable Strategic Behavior", *Econometrica* 52, 1007-1028.
- [7] Blume, L., A. Brandenburger, and E. Dekel (1991): "Lexicographic Probabilities and Equilibrium Re...nements", *Econometrica* 59, 81-98.
- [8] Börgers, T. (1994): "Weak Dominance and Approximate Common Knowledge", *Journal of Economic Theory* 64, 265-276.
- [9] Dekel, E. and D. Fudenberg (1990): "Rational Behavior with Payoffs Uncertainty", *Journal of Economic Theory* 52, 243-267.
- [10] Dekel, E. and F. Gul (1997): "Rationality and Knowledge in Game Theory", in *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Vol. 1*, ed. by D. M. Kreps and K. W. Wallis. Cambridge University Press, 87-172.
- [11] Dekel, E., B. Lipman, and A. Rustichini (1998): "Standard State-Space Models preclude Unawareness", *Econometrica* 66, 159-173.
- [12] Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995): *Reasoning about Knowledge*. Cambridge, MA: The MIT Press.
- [13] Gul, F. (1996): "Rationality and Coherent Theories of Strategic Behavior", *Journal of Economic Theory* 70, 1-31.
- [14] Mertens, J. F. and S. Zamir, (1985): "Formalization of Harsanyi's notion of 'type' and 'consistency' in games with incomplete information", *International Journal of Game Theory* 14, 1-29.

- [15] Pearce, D. G. (1984): "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica* 52, 1029–1050.
- [16] Spohn, W. (1987): "Ordinal Conditional Functions: A Dynamic Theory of Epistemic States", in *Causation in Decision, Belief Change, and Statistics*, Vol. 2, ed. by W. L. Harper and B. Skyrms. Dordrecht: Reidel, 105–134.
- [17] Stalnaker, R. (1994): "On the Evaluation of Solution Concepts", *Theory and Decision* 37, 49-73.
- [18] Stalnaker, R. (1996): "Knowledge, Belief and Counterfactual Reasoning in Games", *Economics and Philosophy* 12, 133-163.
- [19] Tan, T, and S. R. C. Werlang (1988): "The Bayesian Foundations of Solution Concepts of Games", *Journal of Economic Theory* 45, 370-391.