Rationalizability and Minimal Complexity in Dynamic Games

Andrés Perea*
Maastricht University

June 2003

Abstract

This paper presents a formal epistemic framework for dynamic games in which players, during the course of the game, may revise their beliefs about the opponents' utility functions. We impose three key conditions upon the players' beliefs: (a) throughout the game, every move by the opponent should be interpreted as being part of a rational strategy, (b) the belief about the opponents' relative ranking of two strategies should not be revised unless one is certain that the opponent has decided not to choose one of these strategies, and (c) the players' *initial* beliefs about the opponents' utility functions should agree on a given profile u of utility functions. Types that, throughout the game, respect common belief about these three events, are called *persistently rationalizable* for the profile u of utility functions. It is shown that persistent rationalizability implies the backward induction procedure in generic games with perfect information. We next focus on persistently rationalizable types for u that hold a theory about the opponents of "minimal complexity", resulting in the concept of *minimal rationalizability*. For two-player simultaneous move games, minimal rationalizability is equivalent to the concept of Nash equilibrium strategy. In every outside option game, as defined by van Damme (1989), minimal rationalizability uniquely selects the forward induction outcome.

Keywords: Rationalizability, belief revision, dynamic games, backward induction, forward induction.

JEL Classification: C72

1. Introduction

In the epistemic approach to noncooperative games every player is modeled as a decision maker under uncertainty, endowed with a preference ordering on the possible strategy choices. Under the assumption that each player is of the expected utility type, such preference orderings may be represented by a utility function over the possible consequences and a subjective probability distribution, or belief, over the uncertain parameters in the game. Most epistemic models that have been proposed in the literature assume that the players face no uncertainty about

*Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: a.perea@ke.unimaas.nl

the opponents' utility functions (some papers that explicitly allow for uncertainty about the opponents' utility functions will be discussed below). This property is usually modeled by the presence of an exogenously given profile of utility functions and the implicit requirement that, whatever happens in the game, these utility functions are never to be questioned. The uncertainty faced by a player at a given instance of the game will then consist of the opponents' strategy choices, the opponents' beliefs about the other players' strategy choices, the opponents' beliefs about the other players' beliefs about the other players' strategy choices, and so forth.

Within a given epistemic model for games, the problem of how to model rational behavior cannot be reduced to one-person decision theory since a player should not only choose rationally given his beliefs, but these beliefs should also be based upon the conjecture that his opponents choose rationally as well. Also should a player realize that each of his opponents will hold beliefs that are based upon the conjecture that the other players act rationally, and so on. This intuitive argument may be formalized by the notion of *common belief of rationality,* a concept that plays a central role in theories of rationality such as rationalizability (Bernheim (1984) and Pearce (1984)), Nash equilibrium and all refinements thereof. Indeed, Tan and Werlang (1988) have shown that, within a formal epistemic model, the strategies that may be chosen rationally when there is common belief of rationality coincide exactly with the set of rationalizable strategies.

A fundamental problem arises, however, if the notion of common belief of rationality is to be applied to *dynamic* games, and no uncertainty about the utility functions is allowed. The difficulty is that there may be information sets in the game that cannot be reached if players were to act in accordance with common belief of rationality. Reny (1992a, 1993) has shown that for the class of perfect information games, this phenomenon occus on a rather structural basis. A natural question which then arises is: how should a player revise his beliefs about the opponents' strategy choices and the opponents' beliefs if an information set is reached that contradicts common belief of rationality? At this stage, the player should conclude that there is at least one opponent who (a) did not act rationally given his beliefs, or (b) bases his beliefs upon the conjecture that some other player does not act rationally given his belief, or (c) believes that some other player believes that some other player acts irrationally, and so on. A concept of rationality should specify which of the above scenarios is to be viewed as "most plausible", thus imposing a restriction on how beliefs are to be revised at such "problematic" information sets.

In the literature, several rationalizability concepts for dynamic games have been proposed that hold different views on how to revise beliefs when common belief of rationality has been contradicted by the play of the game. The concept of *common certainty of rationality at the beginning of the game* (Ben-Porath (1997)) and its extension to general dynamic games, to which we shall refer as *weak sequential rationalizability,* require common belief *at the beginning of the game* about the event that players choose rationally at each of their information sets, but impose no restriction upon the players' belief revisions at information sets where the player's initial belief about the opponents has been contradicted. In particular, if common belief of rationality has been contradicted at a given information set, the corresponding player may from now on believe that one or more opponents chooses suboptimally.

Backward induction, and backward induction based rationalizability concepts such as *sequential* and *quasi-perfect rationalizability* (Asheim and Perea (2002)), state that a player, at an

information set where common belief of rationality has been contradicted, should conclude that the event of reaching this information set is due to a suboptimal move by one of his opponents, but should maintain his belief in common belief of rationality for the remainder of the game.

The concept of *extensive form rationalizability* (Pearce (1984) and Battigalli (1997)) holds yet another viewpoint by requiring that a player, at an information set contradicting common belief of rationality, should not conclude immediately that an opponent has chosen suboptimally, but should rather seek for the "highest possible degree of interactive belief of rationality[1]" that is compatible with the event of reaching this information set. From then on, the player should base his beliefs about the opponents upon this degree of interactive belief of rationality until some further information set is reached that contradicts this degree. At this occasion, the player should again search for the highest possible degree of interactive belief of rationality that explains the event of reaching this information set, and so on.

In this paper we choose an alternative approach by allowing the players to revise their beliefs about the opponents' utility functions during the game, while insisting on common belief of rationality at every possible instance in the game (see Perea (2002) for a similar approach within an equilibrium framework). Accordingly, we develop an epistemic model in which every player, at each of his information sets, has uncertainty about the opponents' strategy choices, the opponents' utility functions, the opponents' first-order beliefs about the other players' strategy choices, the opponents' first-order beliefs about the other players' utility functions, the opponents' second-order beliefs about the other players' first-order beliefs, etcetera. This leads, for every player at each of his information sets, to an infinite hierarchy of succesively richer uncertainty spaces, to which we refer as the first-order uncertainty space, second-order uncertainty space, and so on, and to an infinite hierarchy of preference orderings over his own strategies. In this hierarchy, the k-th order preference ordering at a given information set is induced by a subjective probability distribution (belief) over the k-th order uncertainty space and a utility function at reachable terminal nodes. In turn, the k-th order uncertainty space contains the opponents' possible $(k-1)$-th order preference relations, and hence a player, at each of his information sets, should hold a belief about the opponents' first-order, second-order, and higher order preference relations. In Perea (2003) it has been shown that the infinite preference hierarchies within our epistemic model can be handled efficiently by means of *types.* More precisely, it can be shown that each preference hierarchy of a player can be identified with a type, which specifies at each of the information sets a subjective probability distribution over the opponents' strategy choices and opponents' types, and which determines a utility function at the terminal nodes. This representation result thus justifies the use of a relatively simple, implicit type-model that makes the analysis easier.

We then proceed by imposing some restrictions upon the types, eventually leading to the concept of *persistent rationalizability.* The first two requirements, *updating consistency* and *proper belief revision,* are concerned with the belief updating and belief revision policies carried out by the types. Updating consistency simply states that Bayesian updating should be used whenever the observed behavior is still in accordance with the previously held beliefs. Proper belief revision states that, whenever a player i type decides to revise his belief about player j's ranking of his own strategy choices, then he should not change his belief about j's relative

[1]Or, "highest possible degree of strategic sophistication", as Battigalli and Siniscalchi (2002) put it.

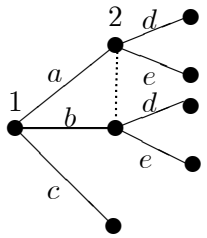

Figure 1

ranking of two strategies unless player i is absolutely certain that j has decided not to choose one of these strategies. The underlying principle is that a player should base his belief revision policy solely on the actual observed behavior, and not on conjectures concerning possible past and future behavior about which he is not absolutely certain. In order to illustrate this principle, consider Figure 1.

Suppose that player 2 initially believes that player 1 strictly prefers c over a, and strictly prefers a over b. If player 2 finds himself at his information set, he is certain that player 1 has not chosen c. If player 2 believes at this stage that player 1 has chosen rationally, player 2 is led to revise his belief about player 1's preference relation. Proper belief revision requires player 2 to maintain his belief that player 1 strictly prefers a over b, while allowing player 1 to change his belief about the ranking of c relative to a and b. The reason is that the observed behavior, namely that player 1 has not chosen c, does not reveal new evidence about player j's ranking of a relative to b, and therefore player 2 should not change his belief about this relative ranking.

The reason to call it *proper* belief revision is that this belief revision principle is implicitly present in the concepts of proper equilibrium (Myerson (1978)) and proper rationalizability (Schuhmacher (1999), Asheim (2001)). The key restriction in both concepts is that a player should never exclude any strategy choice by an opponent, yet should deem one opponent strategy "infinitely more likely" than another if he believes that the opponent strictly prefers the former above the latter. The notion of "infinitely more likely" can either be formalized by taking sequences of full-support probability distributions over the opponents' strategy choices and considering the relative likelihood of two strategy choices in the limit, as is done in Myerson (1978) and Schuhmacher (1999), or can be established by considering lexicographic probability distributions over strategy choices, as used in Blume, Brandenburger and Dekel (1991a, 1991b) and Asheim (2001) in their characterizations of proper equilibrium and proper rationalizability, respectively. If we would apply proper equilibrium or proper rationalizability to the situation in Figure 1, and assume that player 2 initially believes that player 1 strictly prefers c over a, and strictly prefers a over b, then player 2 should deem c infinitely more likely than a, and deem a infinitely more likely than b. Hence, if player 2 observes that player 1 has chosen a or b, he should still deem a infinitely more likely than b. Consequently, if player 2 is faced with the fact that player 1 has chosen a or b, he should still believe that player 1 strictly prefers a over b, and hence proper equilibrium and proper rationalizability imply the proper belief revision principle in this example. Since this argument can be applied to any given dynamic game, proper equilibrium

and proper rationalizability generally support the proper belief revision principle.

The third condition we impose on types, *belief in sequential rationality,* reflects the principle that, whatever happens in the game, a player should always interpret observed moves as rational ones. In particular, if a player i observes a move that would not have been optimal for an opponent j, were player i to keep his previously held belief about j's utility function, then player i should actually revise his belief about j's utilities in order to rationalize this move. Types that, throughout the game, respect common belief about the events that (1) types are updating consistent, (2) types satisfy proper belief revision, and (3) types believe in sequential rationality, are called *persistently rationalizable*.

The literature usually assumes some exogenously given restrictions upon the players' utility functions, and the beliefs they have about the opponents' utilities, modeled by the specification of a fixed profile of utility functions. The implicit interpretation is that players are assumed to hold these utility functions, and are to believe throughout the game that the opponents hold the utility functions as specified by the profile. As to link the concept of persistent rationalizability to existing rationality concepts for given utility functions, we subsequently impose some exogenous restrictions upon the players' utility functions and beliefs about the opponents' utilities. In order to do so, we proceed as above by taking as given a profile u of utility functions, but a different interpretation shall now be attached to it. Players are required to hold the utility functions as specified by u, and to respect common belief about the event that players *initially* believe that opponents hold utility functions as given by u. Persistently rationalizable types that satisfy these additional requirements are said to be *persistently rationalizable for* u, and strategies that are optimal for such types at each of their information sets are called *persistently rationalizable strategies for* u. We thus leave open the possibility that players may change their belief about the opponents' utilities as the game is under way, while requiring that the players' beliefs agree on the same profile of utility functions at the beginning of the game.

In light of the latter property, our approach is related to the model of *games with randomly disturbed payoffs*, as used in Harsanyi (1973), Fudenberg, Kreps and Levine (1988), Dekel and Fudenberg (1990), Zauner (2002) and Stinchcombe and Zauner (2002), among others. In all of these papers, players are assumed to have "infinitesemal uncertainty" about the opponents' utility functions at the beginning of the game, modeled by a sequence of games with randomly perturbed utility functions in which the perturbation vanishes in the limit. The analysis then focusses on the behavior of players as the perturbation tends to zero. When applied to dynamic games, the choice of a sequence of utility perturbations may be seen as a way to model a particular belief revision policy for each player about the possible opponents' utilities. By letting the perturbation vanish in the limit, one imposes that the players' beliefs about the opponents' utilities should (approximately) agree on a particular profile of utility functions at the beginning of the game. A key factor that distinguishes our model from the ones above is that our *proper belief revision* condition imposes an explicit restriction upon the way players should revise their beliefs about the opponents' utilities, while the above mentioned papers, with the exception of Stinchcombe and Zauner (2002), put no constraints on the sequences of utility perturbations that may be chosen.[2] Battigalli (2003), in his analysis of rationalizability in games

[2]Stinchcombe and Zauner (2002) *do* impose a restriction upon the sequence of utility perturbations, however it is different from proper belief revision.

with "genuine incomplete information", takes another approach by assuming that players may have uncertainty about the other players' utility functions, without requiring that the players' beliefs at the beginning of the game agree on a particular profile of utility functions.

Having established the concept of persistent rationalizability for a given profile u of utility functions, our next step is to present a refinement that focusses on types holding beliefs that are "as simple as possible". As to formalize the latter, we introduce the notion of the *complexity* of a type t_i, which, loosely speaking, represents the total number of types that t_i considers directly or indirectly in his theory about the opponents. More precisely, the complexity of a type t_i first counts the number of types t_j that t_i attaches positive probability to in his beliefs. For each of these types t_j, one counts the number of types that t_j attaches positive probability to and that have not been counted already, and so on. By summing up all these types, one gets the total number of types that t_i directly or indirectly refers to in his beliefs throughout the game, and this number is called the complexity of t_i. For a given profile of utility functions u, we say that a type is *minimally rationalizable* for u if (1) it is persistently rationalizable for u, and (2) it has minimal complexity among all types that are persistently rationalizable for u. Accordingly, a strategy is called minimally rationalizable for u if it can be chosen rationally by a type that is minimally rationalizable for u.

The second part of this paper is devoted to relating persistent and minimal rationalizability to existing rationality concepts in the literature. First of all, in Perea (2003) it has been shown that for every given profile u of utility functions, every properly rationalizable strategy for u for types with "non-increasing type supports" is persistently rationalizable for u. Here, the latter concept is a non-empty refinement of proper rationalizability, thus establishing the existence of persistently rationalizable types and strategies for all game trees and all utility functions u. Moreover, the proof of this result shows that persistently rationalizable types with finite complexity can always be found for all u, and hence minimally rationalizable types and strategies always exist.

We next find that persistent rationalizability may be viewed as a possible epistemic foundation for backward induction, since for generic games with perfect information the only persistently rationalizable strategy for each player is his backward induction strategy. Moreover, it establishes an interpretation of backward induction where moves off the backward induction path are not viewed as mistakes by one of the players, but rather as moves in accordance with common belief of rationality, while allowing players to adjust their beliefs about the opponents' utilities when observing such deviations from the backward induction path.

The concept of minimal rationalizability, on the other hand, turns out to have a strong *forward induction* flavour, at least in some classes of games. It is shown that in the class of outside option games, as introduced by van Damme (1989), the unique minimally rationalizable strategy for each player is his forward induction strategy. In particular, minimal rationalizability uniquely selects the forward induction outcome, which is the only Nash equilibrium outcome in the game that dominates the outside option from player 1's point of view. What is remarkable about this result is that no explicit *equilibrium* condition is needed in minimal rationalizability to filter this forward induction Nash equilibrium outcome.

The relation between minimal rationalizability and Nash equilibrium is even more transparent in the class of two-player simultaneous move games, since for such games the set of minimally

rationalizable strategies for a player coincides exactly with the set of Nash equilibrium strategies. As such, minimal rationalizability provides an epistemic characterization of Nash equilibrium strategies for two-player static games. The major difference with the epistemic characterization provided by Aumann and Brandenburger (1995) is that minimal rationalizability does not explicitly impose mutual belief (or knowledge) concerning the players' beliefs about the opponent's strategy choice.

The outline of this paper is as follows. Section 2 presents some preliminary definitions in extensive form games. In Section 3 we present the epistemic framework that will be used as a basis for the rationalizability concepts. Section 4 lays out the concepts of persistent and minimal rationalizability. Sections 5, 6 and 7 prove the above mentioned results on backward induction, Nash equilibrium in two-player simultaneous move games and outside option games, respectively. Section 8, finally, discusses the relationship with other rationalizability concepts such a weak sequential rationalizability and extensive form rationalizability.

2. Extensive Form Structures

In this section we introduce the necessary notation and some preliminary definitions for dynamic games. It is assumed that the reader is familiar with the precise definition of an extensive form game, which we therefore do not present in order to save space. Let I be some finite set of players that faces a dynamic game. The rules of the game are given by an *extensive form structure* $\mathcal{S}$ consisting of (1) a finite game tree, (2) for every player i some finite collection H_i of information sets, (3) for every information set h some finite set $A(h)$ of available actions, and (4) a finite set Z of terminal nodes. We assume that the extensive form structure $\mathcal{S}$ satisfies perfect recall and contains no chance moves. The latter requirement is not really needed for our analysis, but rather simplifies the exposition. Let $H = \cup_{i \in I} H_i$ be the collection of all information sets in the game. By h_0 we denote the information set that marks the beginning of the game. Let $H_i^* = H_i \cup \{h_0\}$ for every player i.

The definition of a strategy we use in this paper follows Rubinstein's (1991) notion of a *plan of action,* specifying an action only at those information sets which are not avoided by the strategy itself. Let $\tilde{H}_i \subseteq H_i$ be some collection of player i information sets, not necessarily including all, and let s_i be a function that assigns an available action $s_i(h_i) \in A(h_i)$ to every information set h_i in $\tilde{H}_i$. We say that an information set $h \in H$ (possibly controlled by some player other than i) is *avoided* by the function s_i if every profile of available actions in $\times_{h \in H} A(h)$ that coincides with s_i on $\tilde{H}_i$, avoids the information set h. The function s_i defined on $\tilde{H}_i$ is called a *strategy* if $\tilde{H}_i$ is exactly the collection of player i information sets not avoided by s_i. Obviously, every player i strategy may be obtained by first prescribing an action at *all* player i information sets, and then discarding those player i information sets that are avoided by it. Let S_i denote the set of player i strategies.

An additional restriction we impose upon the extensive form structure is that it should be *with observable deviators* (see e.g. Battigalli (1996)). For its definition, we need some additional notation. For a given information set h, let $S(h)$ be the set of strategy profiles $(s_i)_{i \in I}$ that reach h, and let $S_i(h)$ be the set of player i strategies that do not avoid h. Here, player i needs not be the player who controls h. We say that the extensive form structure $\mathcal{S}$ is with observable

deviators if $S(h) = \times_{i \in I} S_i(h)$ for every information set h. Hence, if every player i chooses a strategy that cannot avoid h by itself, then the resulting strategy profile will reach h. The condition is implied by perfect recall if there are only two players involved. For more than two players, however, this is no longer true.

3. The Epistemic Model

Our next step will be to lay out an epistemic model for dynamic games in which players, during the course of the game, may revise their beliefs about the opponents' preference relations, including the opponents' utilities at the terminal nodes. The model is identical to the one used in Perea (2003) and the reader is referred to that paper for the proofs of the results to be presented in this section. Before discussing the epistemic model, we shall briefly introduce the notions of *acts* and *expected utility* preference relations over acts, as they play a crucial role in the model.

Consider a compact metric space X endowed with some topology. The space X is to be interpreted as a collection of relevant parameters about which the decision maker is uncertain. Let Y be some finite set of possible consequences, and $\Delta(Y)$ the set of probability distributions on Y, endowed with the natural topology. Following Anscombe and Aumann (1963), every decision may be identified with a mapping from X to $\Delta(Y)$, to which we refer as *acts.* [3] Let $\mathcal{F}(X,Y)$ be the set of measurable acts from X to $\Delta(Y)$, and assume that the decision maker holds a preference relation over all acts in $\mathcal{F}(X,Y)$. We say that this preference relation is of the *expected utility* type if there is some probability distribution μ on X and some von Neumann-Morgenstern utility function u from Y to the real numbers such that act f is weakly preferred to act g if and only if

$$\int_X u(f(x))\ d\mu \geq \int_X u(g(x))\ d\mu.$$

Here, $u(f(x))$ is the expected utility induced by the probability distribution $f(x) \in \Delta(Y)$ and the utility function u. Similarly for $u(g(x))$. Let $\mathcal{P}^{eu}(X,Y)$ be the set of non-trivial expected utility preference relations on $\mathcal{F}(X,Y)$. For a given preference relation in $\mathcal{P}^{eu}(X,Y)$ the probability distribution μ is unique and the utility function is unique up to a positive affine transformation. Hence, every member of $\mathcal{P}^{eu}(X,Y)$ may be uniquely identified with a pair (μ, u) where μ is a probability distribution on X and u is a utility function on Y with maximum value 1 and minimum value 0. Let $U(Y)$ be the set of utility functions on Y with the latter property. We thus may identify $\mathcal{P}^{eu}(X,Y)$ with the space $\Delta(X) \times U(Y)$. If we endow this space with the product toplogy induced by the weak topology on $\Delta(X)$ and the natural toplogy on $U(Y)$, the set $\mathcal{P}^{eu}(X,Y)$ becomes a compact metric space.

Let us now return to dynamic games. The basic assumption is that every player i, at each of his information sets $h_i \in H_i^*$, holds a preference relation on the set $S_i(h_i)$ of strategies that are compatible with h_i, while facing uncertainty about the opponents' strategy choices. Recall that $H_i^* = H_i \cup \{h_0\}$. At information set h_i, the set of feasible opponents' strategies is given by

[3] In Anscombe and Aumann (1963), such acts are called *compound horse lotteries.*

$S_{-i}(h_i) = \times_{j \neq i} S_j(h_i)$. Hence, we may define

$$X_i^1(h_i) = S_{-i}(h_i)$$

as player i's first-order space of uncertainty at information set h_i. Let $Z(h_i)$ be the set of terminal nodes following h_i. Every strategy s_i in $S_i(h_i)$ may now be identified with an act f_{s_i} in $\mathcal{F}(S_{-i}(h_i), Z(h_i))$ which assigns to every opponents' strategy profile $s_{-i} \in S_{-i}(h_i)$ the probability distribution on $Z(h_i)$ that attaches probability one to the terminal node reached by (s_i, s_{-i}). Hence, $S_i(h_i)$ may be embedded in the set of acts $\mathcal{F}(X_i^1(h_i), Z(h_i))$. It is assumed that player i at h_i holds a non-trivial expected utility preference relation $p_i^1(h_i)$ on all acts in $\mathcal{F}(X_i^1(h_i), Z(h_i))$, that is, $p_i^1(h_i) \in \mathcal{P}^{eu}(X_i^1(h_i), Z(h_i))$. We refer to $p_i^1(h_i)$ as player i's first-order preference relation at h_i, while $\mathcal{P}^{eu}(X_i^1(h_i), Z(h_i))$ is the set of player i's first-order preference relations at h_i. By choosing the appropriate topology as described above, the latter set becomes a compact metric space.

The set $X_i^1(h_i)$, however, does not contain *all* relevant parameters about which player i is uncertain at h_i, since player i also faces uncertainty about the first-order preference relations that his opponents hold at each of their information sets. We may thus define player i's second-order space of uncertainty at h_i by

$$X_i^2(h_i) = X_i^1(h_i) \times (\times_{j \neq i} \times_{h_j \in H_j^*} \mathcal{P}^{eu}(X_j^1(h_j), Z(h_j))),$$

containing both the opponents' possible strategy choices that may have led to h_i and the opponents' possible first-order preference relations at each of their information sets. Similarly as above, we assume that player i at h_i holds an expected utility preference relation $p_i^2(h_i)$ on the set of all acts in $\mathcal{F}(X_i^2(h_i), Z(h_i))$. We refer to p_i^2 as player i's second-order preference relation at h_i, and to $\mathcal{P}^{eu}(X_i^2(h_i), Z(h_i))$ as the set of possible second-order preference relations at h_i. The players' k-th order uncertainty spaces at their respective information sets may then be defined inductively by

$$X_i^k(h_i) = X_i^{k-1}(h_i) \times (\times_{j \neq i} \times_{h_j \in H_j^*} \mathcal{P}^{eu}(X_j^{k-1}(h_j), Z(h_j)))$$

for all players i, information sets $h_i \in H_i^*$ and $k \geq 2$. At each information set h_i, player i is assumed to hold a hierarchy of preference relations $p_i(h_i) = (p_i^k(h_i))_{k \in \mathbb{N}}$, where $p_i^k(h_i)$ represents the k-th order preference relation in $\mathcal{P}^{eu}(X_i^k(h_i), Z(h_i))$. A vector $p_i = (p_i(h_i))_{h_i \in H_i^*}$ of such hierarchies of preference relations, one for each information set, is simply called a preference hierarchy for player i. Let P_i be the set of all preference hierarchies for player i.

Similar to Epstein and Wang (1996), a preference hierarchy p_i is called *coherent* if for every information set h_i and every order $k \geq 2$, the marginal of the k-th order preference relation $p_i^k(h_i)$ on the $(k-1)$-th order space of acts $\mathcal{F}(X_i^{k-1}(h_i), Z(h_i))$ coincides with the $(k-1)$-th order preference relation $p_i^{k-1}(h_i)$.[4] In words, coherence means that the different preference relations in the hierarchy should coincide on overlapping subspaces. We obtain the following representation result for coherent preference hierarchies, for which a proof can be found in Perea (2003).

[4]Every act in $\mathcal{F}(X_i^{k-1}(h_i), Z(h_i))$ may be identified with some act in $\mathcal{F}(X_i^k(h_i), Z(h_i))$ that only depends on the argument in $X_i^{k-1}(h_i)$. As such, $\mathcal{F}(X_i^{k-1}(h_i), Z(h_i))$ may be viewed as a subspace of $\mathcal{F}(X_i^k(h_i), Z(h_i))$, and hence the marginal of $p_i^k(h_i)$ on $\mathcal{F}(X_i^{k-1}(h_i), Z(h_i))$ is well-defined.

Lemma 3.1. *For every player i, the set of coherent preference hierarchies is homeomorphic to the space $\times_{h_i \in H_i^*} \mathcal{P}^{eu}(S_{-i}(h_i) \times P_{-i}, Z(h_i))$.*

Here, $P_{-i} = \times_{j \neq i} P_j$ is the space of all opponents' preference hierarchies. As such, every coherent preference hierarchy p_i for player i induces at every information set $h_i \in H_i^*$ some expected utility preference relation on acts in $\mathcal{F}(S_{-i}(h_i) \times P_{-i}, Z(h_i))$, representable by a probability distribution $\mu_i(p_i, h_i)$ on $S_{-i}(h_i) \times P_{-i}$, and some utility function $u_i(p_i, h_i)$ on $Z(h_i)$. For an opponent j, let $\mu_i(p_i, h_i|\ P_j)$ be the marginal of the probability distribution $\mu_i(p_i, h_i)$ on the set of player j's preference hierarchies. Let $P_j(p_i, h_i) = \text{supp}\ \mu_i(p_i, h_i|\ P_j)$ be the set of player j's preference hierarchies to which p_i assigns positive probability at h_i. For $j = i$, we define $P_i(p_i, h_i) = \{p_i\}$. By $P(p_i, h_i) = \cup_{j \in I} P_j(p_i, h_i)$ we denote the set of all preference hierarchies to which p_i assigns positive probability at h_i. Let $P(p_i) = \cup_{h_i \in H_i^*} P(p_i, h_i)$ be the set of all preference hierarchies to which p_i assigns positive probability somewhere in the game.

Let $\tilde{P} \subseteq \cup_{j \in I} P_j$ be a set of profiles of preference hierarchies, or simply an *event.* We say that the coherent preference hierarchy p_i *believes* the event $\tilde{P}$ if $P(p_i) \subseteq \tilde{P}$, that is, if p_i only assigns positive probability to preference hierarchies that belong to $\tilde{P}$. We now define *common belief about coherence* by means of the following recursive definition of sets: let $P_i^{c,1}$ be the set of coherent preference hierarchies for player i, and for every $k \geq 2$ let $P_i^{c,k}$ be the set of preference hierarchies in $P_i^{c,k-1}$ that believe $\cup_{j \in I} P_j^{c,k-1}$. By $P_i^{c,\infty} = \cap_{k \in \mathbb{N}} P_i^{c,k}$ we denote the set of preference hierarchies that respect common belief about coherence. Hence, $p_i \in P_i^{c,\infty}$ if and only if p_i is coherent, believes that all opponents' preference hierarchies are coherent, believes that all opponents' preference hierarchies believe that all other players' preference hierarchies are coherent, and so on. In the spirit of Armbruster and Böge (1979), Böge and Eisele (1979) and Mertens and Zamir (1985) we may now derive the following representation result for preference hierarchies that respect common belief about coherence. A proof may be found in Perea (2003).

Lemma 3.2. *For every player i, the set $P_i^{c,\infty}$ of preference hierarchies that respect common belief of coherence is homeomorphic to the space $\times_{h_i \in H_i^*} \mathcal{P}^{eu}(S_{-i}(h_i) \times P_{-i}^{c,\infty}, Z(h_i))$.*

By $T_i = P_i^{c,\infty}$ we denote the set of player i *types.* Then, by the lemma above, the type-space T_i for every player i is homeomorphic to $\times_{h_i \in H_i^*} \mathcal{P}^{eu}(S_{-i}(h_i) \times T_{-i}, Z(h_i))$. That is, every type t_i may be identified with a vector that induces at every information set $h_i \in H_i^*$ a probability distribution, or belief, $\mu_i(t_i, h_i)$ on $S_{-i}(h_i) \times T_{-i}$, and a utility function $u_i(t_i, h_i)$ on $Z(h_i)$. Similarly as we have done above for coherent preference hierarchies p_i, we may define for every type t_i the set $T(t_i) \subseteq \cup_{j \in I} T_j$ as the set of types to which t_i assigns positive probability somewhere in the game. We recursively define the sets of types $T^1(t_i), T^2(t_i), ...$ by $T^1(t_i) = T(t_i)$, and

$$T^k(t_i) = \bigcup_{t \in T^{k-1}(t_i)} T(t)$$

for all $k \geq 2$. By construction, $T^k(t_i)$ for $k \geq 2$ is the set of types t for which one can build a sequence $(t_i = t^1, t^2, ..., t^k = t)$ of length k such that t_i assigns positive probability to t^2, t^2 assigns positive probability to $t^3, ..., t^{k-1}$ assigns positive probability to t somewhere in the

game. By $T^\infty(t_i) = \cup_{k\in\mathbb{N}} T^k(t_i)$ we denote the set of types that may be reached from t_i by building such sequences of arbitrary length.

Let $\tilde{T} \subseteq \cup_{j\in I} T_j$ be a collection of types, or simply an event. We say that type t_i respects *common belief about* $\tilde{T}$ if $T^\infty(t_i) \subseteq \tilde{T}$. In words, t_i believes that all types belong to $\tilde{T}$, believes that all types believe that all types belong to $\tilde{T}$, and so on.

4. Persistent and Minimal Rationalizability

4.1. Persistent Rationalizability

We shall now impose three restrictions upon types in our epistemic model: updating consistency, proper belief revision and belief in sequential rationality. The concept of *persistent rationalizability* then selects those types that respect common belief about these three events throughout the game.

Updating consistency requires a type to update his beliefs using Bayes' rule, whenever possible.

Definition 4.1. *A type t_i is called updating consistent if for every two information sets $h_i^1, h_i^1 \in H_i^*$ such that h_i^2 follows h_i^1, and every event $E \subseteq S_{-i}(h_i^2) \times T_{-i}$, it holds that*

$$\mu_i(t_i, h_i^2)(E) = \frac{\mu_i(t_i, h_i^1)(E)}{\mu_i(t_i, h_i^1)(S_{-i}(h_i^2) \times T_{-i})}$$

whenever $\mu_i(t_i, h_i^1)(S_{-i}(h_i^2) \times T_{-i}) > 0$.

Proper belief revision states that a type, when revising his belief about an opponent's preference relation, should not change his belief about the opponent's relative ranking of two strategies unless he is certain that the opponent has not chosen one of these strategies. Formally, let t_i be a type and $h_i \in H_i^*$ an information set for player i. By definition, t_i knows at h_i that opponent j has chosen some strategy in $S_j(h_i)$, without being able to exclude any of these strategies. Proper belief revision then argues that t_i may revise his belief at h_i about player j's preference relation over strategies in S_j, but should not change his belief about player j's ranking of strategies in $S_j(h_i)$. In the formal definition below, let $T_j(t_i, h_i)$ be the set of player j types to which t_i assigns positive probability at information set h_i.

Definition 4.2. *A type t_i is said to satisfy proper belief revision if for every two information sets $h_i^1, h_i^2 \in H_i^*$ with h_i^2 following h_i^1, and every type $t_j^2 \in T_j(t_i, h_i^2)$, there is some type $t_j^1 \in T_j(t_i, h_i^1)$ with the property that t_j^1 and t_j^2 hold the same preference relation over strategies in $S_j(h_i^2) \cap S_j(h_j)$ at every information set $h_j \in H_j^*$.*

Finally, belief in sequential rationality reflects the principle that a player, at each of his information sets, should believe that his opponents are carrying out strategies that are optimal for them at each of *their* information sets. Formally, we say that a strategy-type pairs (s_i, t_i) is *sequentially rational* if at every information set $h_i \in H_i^*(s_i)$ it holds that

$$u_i(t_i, t_i|\ h_i) = \max_{s_i' \in S_i(h_i)} u_i(s_i', t_i|\ h_i).$$

Here, $H_i^*(s_i)$ denotes the collection of information sets in H_i^* that are not avoided by s_i. By $u_i(t_i, t_i|\ h_i)$ we denote the expected utility at h_i induced by the utility function $u_i(t_i, h_i)$, the strategy s_i and the marginal of the belief $\mu_i(t_i, h_i)$ on the set $S_{-i}(h_i)$ of opponents' strategies. For every player j, let $(S_j \times T_j)^{sr}$ be the set of sequentially rational strategy-type pairs. For a given player i, let $(S_{-i} \times T_{-i})^{sr} = \times_{j \neq i}(S_j \times T_j)^{sr}$ be the set of opponents' sequentially rational strategy-type pairs.

Definition 4.3. *A type t_i is said to believe in sequential rationality if for every information set $h_i \in H_i^*$ it holds that supp $\mu_i(t_i, h_i) \subseteq (S_{-i} \times T_{-i})^{sr}$.*

We are now ready to define the concept of persistent rationalizability.

Definition 4.4. *A type t_i is called persistently rationalizable if it respects common belief about the events that (1) types are updating consistent, (2) types satisfy proper belief revision, and (3) types believe in sequential rationality.*

4.2. Restrictions on Utility Functions and Initial Beliefs

We now proceed by imposing some exogenous restrictions upon the players' actual utility functions, and the initial beliefs players have about the utility functions of others. Consider a profile $u = (u_i)_{i \in I}$ of utility functions over the terminal nodes. Together with the extensive form structure $\mathcal{S}$, this induces a pair $(\mathcal{S}, u)$ which is usually called an *extensive form game.* The most common interpretation of $(\mathcal{S}, u)$ is that players hold utility functions as specified by u, and that they should believe *throughout the game* that the opponents have utility functions as given by u. Our interpretation of $(\mathcal{S}, u)$ will be different since we require players to *initially* believe that opponents have utility functions in u, while allowing the players to change their beliefs about the opponents' utility functions later on in the game. Formally, we say that a type t_i *initially believes* u if supp $\mu_i(t_i, h_0)$ only contains types t_j with $u_j(t_j, h_j) = u_j|_{Z(h_j)}$ for all $h_j \in H_j^*$. Here, $u_j|_{Z(h_j)}$ denotes the restriction of u_j on the set $Z(h_j)$ of terminal nodes following h_j.

Definition 4.5. *Let the profile $u = (u_i)_{i \in I}$ of utility functions be given. We say that a type t_i is persistently rationalizable for $(\mathcal{S}, u)$ if (1) t_i is persistently rationalizable, (2) $u_i(t_i, h_i) = u_i|_{Z(h_i)}$ for all $h_i \in H_i^*$, and (3) t_i respects common belief about the event that types initially believe u. We say that a strategy s_i is persistently rationalizable for $(\mathcal{S}, u)$ if there is a persistently rationalizable type t_i for $(\mathcal{S}, u)$ such that (s_i, t_i) is sequentially rational.*

4.3. Minimal Rationalizability

We shall next focus on types that are persistently rationalizable for a given extensive form game $(\mathcal{S}, u)$, and, moreover, hold a theory about the opponents that in some sense is "as simple as possible". In order to formalize the latter, we introduce the notion of the *complexity* of a type. Recall from Section 3 that for a given type t_i, the set $T^\infty(t_i) \subseteq \cup_{j \in I} T_j$ denotes the collection of types t_j such that either (1) t_i assigns positive probability to t_j at one of his information sets, (2) t_i assigns positive probability to some type t_k which assigns positive probability to t_j at some of his information sets, and so on. Hence, one could say that $T^\infty(t_i)$ represents the set

of types that t_i uses, directly or indirectly, in his theory about the opponents' strategies and beliefs throughout the game. Let $c(t_i)$ be the total number of types in $T^\infty(t_i)$, which could in principle be infinite. We refer to $c(t_i)$ as the *complexity* of type t_i.

Definition 4.6. *Let $(\mathcal{S}, u)$ be an extensive form game. Then, a type t_i is called minimally rationalizable for $(\mathcal{S}, u)$ if t_i is persistently rationalizable for $(\mathcal{S}, u)$ and has minimal complexity among all player i types that are persistently rationalizable for $(\mathcal{S}, u)$. A strategy s_i is said to be minimally rationalizable for $(\mathcal{S}, u)$ if there is some minimally rationalizable type t_i for such that (s_i, t_i) is sequentially rational.*

4.4. Existence

In Perea (2003) it has been shown that for every extensive form structure $\mathcal{S}$ with observable deviators, and every profile u of utility functions, every player has at least one persistently rationalizable type and strategy for $(\mathcal{S}, u)$. The existence follows from a theorem establishing a general relationship between the concept of proper rationalizability (Schuhmacher (1999) and Asheim (2001)) on the one hand, and the concept of persistent rationalizability on the other hand. More precisely, the theorem states that for a given extensive form game $(\mathcal{S}, u)$, every properly rationalizable strategy for $(\mathcal{S}, u)$ for types with "non-inceasing type-supports" is a persistently rationalizable strategy for $(\mathcal{S}, u)$. The former notion constitutes a refinement of proper rationalizability, and following Asheim (2001) it can easily be shown that properly rationalizable strategies for types with non-increasing type supports always exist for every $(\mathcal{S}, u)$.

The proof of the theorem not only establishes the existence of persistently rationalizable types for all $(\mathcal{S}, u)$, it also shows that we can always find a persistently rationalizability type with *finite* complexity for each player . The reason is that the proof chooses for every $(\mathcal{S}, u)$ a properly rationalizable type with "non-increasing type supports" (this can always be found) and then explicitly transforms this type into a persistently rationalizable type for $(\mathcal{S}, u)$ with finite complexity. As such, the notion of minimal rationalizability for $(\mathcal{S}, u)$ is always well-defined, and every minimally rationalizable type will always have a finite complexity. In the remainder of this paper, we shall apply the concepts of persistent and minimal rationalizability to several special classes of games, and investigate their relationships to existing rationality concepts.

5. Games with Perfect Information

In this section we show that in generic games with perfect information, every player has a unique persistently rationalizable strategy, namely his backward induction strategy. A game with perfect information $(\mathcal{S}, u)$ is said to be *in generic position* if for every player i and every pair z_1, z_2 of different terminal nodes, we have that $u_i(z_1) \neq u_i(z_2)$. For such a game, let $a^*(h_i) \in A(h_i)$ denote the unique backward induction action at information set h_i. For every player i, there is a unique strategy s_i^* with $s_i^*(h_i) = a^*(h_i)$ for all $h_i \in H_i(s_i^*)$, to which we shall refer as the *backward induction strategy.*

Theorem 5.1. *Let $(\mathcal{S}, u)$ be a game with perfect information in generic position. Then, a strategy is persistently rationalizable for $(\mathcal{S}, u)$ if and only if it is a backward induction strategy for $(\mathcal{S}, u)$.*

Proof. Let $(\mathcal{S}, u)$ be a game with perfect information in generic position. For every player i and every information set h_i, let $S_i^*(h_i)$ denote the set of strategies s_i such that (1) at every information set $\tilde{h}_i \in H_i$ preceding h_i the strategy s_i prescribes the unique action at $\tilde{h}_i$ which leads to h_i, and (2) at every information set $\tilde{h}_i$ following h_i which is not avoided by s_i, the strategy s_i prescribes the unique backward induction action $a^*(h_i)$. We refer to $S_i^*(h_i)$ as the set of player i *backward induction strategies conditional on* h_i in the game $(\mathcal{S}, u)$. Let $T_i^*(h_i)$ be the set of player i types t_i such that t_i's most preferred strategies at h_i all belong to $S_i^*(h_i)$. For a given type t_i, let $\mu_i(t_i, h_i|\ T_j)$ be the marginal probability distribution of $\mu_i(t_i, h_i)$ on player j's types. Let $T_i^\infty(u)$ be the set of player i types that respect common belief about the event that types initially believe u. We prove the following claim.

Claim. Let t_i be persistently rationalizable and $t_i \in T_i^\infty(u)$. Then, for every information set $h_i \in H_i^*$, every opponent j and information set h_j following h_i, it holds that $\text{supp}\mu_i(t_i, h_i|\ T_j) \subseteq T_j^*(h_j)$.

Proof of claim. By induction on the number of decision nodes following h_j. Suppose first that h_j is not followed by any decision node. Let $t_j \in \text{supp}\mu_i(t_i, h_i|\ T_j)$. Then, since t_i satisfies proper belief revision, there is some $t_j^0 \in \text{supp}\mu_i(t_i, h_0|\ T_j)$ such that t_j^0 and t_j have at h_j the same preference relation over strategies in $S_j(h_i)$. Since h_i precedes h_j, we have that $S_j(h_j) \subseteq S_j(h_i)$, and hence t_j^0 and t_j have the same preference relation over strategies in $S_j(h_j)$. Since $t_i \in T_i^\infty(u)$ we have that t_i initially believes u, and hence $u_j(t_j^0, h_j) = u_j|_{Z(h_j)}$. Since h_j is not followed by any decision node, it must hold that t_j^0 at h_j strictly prefers the backward induction action. But then, since t_j at h_j has the same preference relation over actions at h_j as t_j^0, it must hold that also t_j at h_j strictly prefers the backward induction action, and hence $t_j \in T_j^*(h_j)$. We thus have shown that $\text{supp}\mu_i(t_i, h_i|\ T_j) \subseteq T_j^*(h_j)$.

Now, suppose that the claim holds for every t_i, h_i and h_j where h_j is followed by at most $K-1$ decision nodes. We prove the claim for information sets h_j followed by K decision nodes. Choose h_i and h_j such that h_j follows h_i and h_j is followed by K decision nodes. Let $t_j \in \text{supp}\mu_i(t_i, h_i|\ T_j)$. We show that $t_j \in T_j^*(h_j)$. Since t_i is persistently rationalizable and $t_i \in T_i^\infty(u)$, we have that t_j is persistently rationalizable and $t_j \in T_j^\infty(u)$. Let player $k \neq j$ and information set $h_k \in H_k$ be such that h_k follows h_j. Since h_k is followed by at most $K-1$ decision nodes, we may apply the induction assumption to t_j, h_j and h_k and conclude that $\text{supp}\mu_j(t_j, h_j|\ T_k) \subseteq T_k^*(h_k)$. Hence, for all players $k \neq j$ and all information sets h_k following h_j we have that $\text{supp}\mu_j(t_j, h_j|\ T_k) \subseteq T_k^*(h_k)$. Since t_j is persistently rationalizable, we have in particular that t_j believes in sequential rationality. By the above, we may thus conclude that t_j believes at h_j that at all future information sets the corresponding player chooses the backward induction action.

Since $t_j \in \text{supp}\mu_i(t_i, h_i|\ T_j)$ and t_i satisfies proper belief revision, there is some $t_j^0 \in \text{supp}\mu_i(t_i, h_0|\ T_j)$ such that t_j^0 and t_j hold the same preference relation over strategies in $S_j(h_i) \supseteq S_j(h_j)$. As t_i initially believes u, it must hold that $u_j(t_j^0, h_j) = u_j|_{Z(h_j)}$. Moreover, since $t_j^0 \in \text{supp}\mu_i(t_i, h_0|\ T_j)$ we may apply the same reasoning as above to conclude that t_j^0 is persistently rationalizable and $t_j^0 \in T_j^\infty(u)$. But then, by copying the argument above for t_j, this leads to the conclusion that t_j^0 believes at h_j that at all future information sets the corresponding player chooses the backward induction action. Together with the fact that $u_j(t_j^0, h_j) = u_j|_{Z(h_j)}$, this implies that t_j^0's most preferred strategies at h_j are backward induction strategies conditional

on h_j.

Above we have seen that t_j^0 and t_j hold the same preference relation over strategies in $S_j(h_j)$, and hence t_j's most preferred strategies at h_j are backward induction strategies conditional on h_j, that is, $t_j \in T_j^*(h_j)$. This completes the proof of the claim.

Now, choose an arbitrary strategy s_i that is persistently rationalizable for $(\mathcal{S}, u)$. Then, there must be a type t_i which is persistently rationalizable for $(\mathcal{S}, u)$ such that s_i is sequentially rational for t_i. By definition, $t_i \in T_i^\infty(u)$, and hence, by the claim, it follows that for every information set $h_i \in H_i^*$, every opponent j and information set h_j following h_i, we have $\text{supp}\mu_i(t_i, h_i | T_j) \subseteq T_j^*(h_j)$. Since t_i believes in sequential rationality, we may conclude that t_i believes at every information set h_i that at all future information sets the corresponding player chooses his backward induction action. By assumption, $u_i(t_i, h_i) = u_i|_{Z(h_i)}$ for all information sets h_i, and hence t_i prefers at every information set h_i a backward induction strategy conditional on h_i. Since s_i is sequentially rational for t_i, it must be that s_i is player i's unique backward induction strategy in $(\mathcal{S}, u)$. This completes the proof of this theorem. ■

In view of Theorem 5.1, the concept of persistent rationalizability may be employed as an alternative epistemic foundation for backward induction in games with perfect information. There is an important difference with other foundations proposed in the literature, such as Aumann (1995), Samet (1996), Balkenborg and Winter (1997), Stalnaker (1998) and Asheim (2000), as persistent rationalizability allows players to revise their conjectures about the opponents' utility functions during the game, whereas the latter foundations do not. In turn, persistent rationalizability requires players to interpret "unexpected moves" (in this case, moves that deviate from the backward induction play) always as being in accordance with common belief of rationality.

6. Simultaneous Move Games

In Section 4, we have defined *minimally rationalizable* types for $(\mathcal{S}, u)$ as those persistently rationalizable types for $(\mathcal{S}, u)$ that have minimal complexity. Recall that the complexity of a type t_i denotes the total number of types that t_i, directly or indirectly, uses in his theory about the opponents' beliefs. In this section, we show that the minimal complexity criterion has non-trivial implications even for the class of simultaneous move games in which belief revision plays no role. In these games, persistent rationalizability is equivalent to rationalizability, as defined in Bernheim (1984) and Pearce (1984). Together with the restriction that types hold utility functions as specified by u and that there be common belief about the event that types initially believe u, the epistemic model of Section 3, when applied to simultaneous move games, is equivalent to the one used by Tan and Werlang (1988). Minimal rationalizability thus restricts attention to those rationalizable strategies that can be justified by an epistemic rationalizability theory (cf. Tan and Werlang (1988)) which involves as few types as possible. We shall prove that for the case of two-player simultaneous move games, this concept is equivalent to the notion of *Nash equilibrium strategies.*

In order to formalize this result, we first need the definition of a Nash equilibrium *strategy.* For a given two-person simultaneous move game, a first-order belief about player i is a probability distribution $\mu_i \in \Delta(S_i)$, reflecting player j's belief about player i's strategy choice. A profile

(μ_1, μ_2) of first-order beliefs is a Nash equilibrium if $\mu_i(s_i) > 0$ implies that s_i is a best response against μ_j. A strategy s_i is a *Nash equilibrium strategy* if there is some Nash equilibrium (μ_1, μ_2) such that s_i is a best response against μ_j. Since not every rationalizable strategy in a two-player game is a Nash equilibrium strategy, the following result implies that minimal rationalizability is indeed stronger than rationalizability in two-player simultaneous move games.

Theorem 6.1. *Let $(\mathcal{S}, u)$ be a two-player simultaneous move game. Then, s_i is minimally rationalizable for $(\mathcal{S}, u)$ if and only if s_i is a Nash equilibrium strategy for $(\mathcal{S}, u)$.*

Proof. Let s_i be a Nash equilibrium strategy. Then, there is some Nash equilibrium (μ_i, μ_j) in first-order beliefs such that s_i is optimal against μ_j. We may construct two types t_i and t_j such that $u_i(t_i, h_0) = u_i$, $\mu_i(t_i, h_0)$ assigns probability one to t_j, $\mu_i(t_i, h_0)(s_j, t_j) = \mu_j(s_j)$ for all $s_j \in S_j$, and similarly for type t_j. Then, by the properties of Nash equilibrium, t_i is persistently rationalizable. Since $T^\infty(t_i) = \{t_i, t_j\}$, the complexity of t_i is 2, which is clearly minimal. Hence, t_i is minimally rationalizable. Since s_i is optimal against μ_j, it follows that (s_i, t_i) is sequentially rational, which implies that s_i is minimally rationalizable for $(\mathcal{S}, u)$.

Now, let s_i be minimally rationalizable. Then, there exists some minimally rationalizable type t_i for this game such that (s_i, t_i) is sequentially rational. We have seen above that every Nash equilibrium induces a type that is persistently rationalizable for $(\mathcal{S}, u)$ with complexity 2. Since Nash equilibria always exist, we may thus conclude that t_i must have complexity 2, that is, $T^\infty(t_i) = \{t_i, t_j\}$ for some t_j. Hence, $\mu_i(t_i, h_0)$ assigns probability one to t_j, and $\mu_j(t_j, h_0)$ assigns probability one to t_i. Let μ_i be the marginal of $\mu_j(t_j, h_0)$ on S_i, and μ_j the marginal of $\mu_i(t_i, h_0)$ on S_j. Since t_i is persistently rationalizable, it follows that (μ_i, μ_j) is a Nash equilibrium. Since (s_i, t_i) is sequentially rational, it follows that s_i is optimal against μ_j, and hence s_i is a Nash equilibrium strategy. This completes the proof. ■

The characterization result no longer holds for more than two players. In order to see this, consider the following three-player simultaneous move game, represented by its normal form.

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

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

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

Here, player 1 chooses between a, b and c, player 2 chooses between d, e and f, whereas player 3 chooses between g, h and i. We show that g is a minimally rationalizable strategy for player 3, but not a Nash equilibrium strategy. Consider types t_1, t_2, t_3 such that $\mu_1(t_1, h_0)$ puts probability one on $((e, t_2), (g, t_3))$, $\mu_2(t_2, h_0)$ puts probability one on $((a, t_1), (h, t_3))$ and $\mu_3(t_3, h_0)$ puts probability one on $((a, t_1), (e, t_2))$. Then, it may be verified that t_3 is persistently rationalizable for $(\mathcal{S}, u)$. Type t_3 has complexity 3, which is the minimum possible complexity, and hence t_3 is minimally rationalizable. Since g is sequentially rational for t_3, it follows that g is minimally rationalizable.

Suppose that g would be a Nash equilibrium strategy. Then, there would be a Nash equilibrium (μ_1, μ_2, μ_3) in first-order beliefs such that g would be optimal against (μ_1, μ_2). Here,

$\mu_i \in \Delta(S_i)$ represents player i's opponents' common belief about player i's strategy choice. However, g can only be optimal against (μ_1, μ_2) if $\mu_1(a) > 0$ and $\mu_2(e) > 0$. Since (μ_1, μ_2, μ_3) is a Nash equilibrium, this implies that a is optimal against (μ_2, μ_3) and e is optimal against (μ_1, μ_3). This, in turn, implies that $\mu_3(g) \geq \frac{2}{3}$ and $\mu_3(h) \geq \frac{2}{3}$, which is clearly impossible. Hence, g is not a Nash equilibrium strategy. The difference with minimal rationalizability is that the concept of Nash equilibrium requires player 3 to believe that players 1 and 2 hold the same belief about player 3's strategy choice, whereas minimal rationalizability does not impose such restriction. Indeed, types t_1 and t_2 above hold different beliefs about t_3's strategy choice, and this is what makes g minimally rationalizable.

One direction of Theorem 6.1 remains true, however, if more than two players are allowed. Namely, in every n-player simultaneous move game, every Nash equilibrium strategy is minimally rationalizable. The proof is similar to the proof above and is left to the reader.

7. Outside Option Games

In this section, we shall prove that the concept of minimal rationalizability singles out the unique forward induction outcome in so-called *outside option games* as defined in van Damme (1989). An outside option game is a two-player game $(\mathcal{S}, u)$ with the following properties:
(1) At the beginning, player 1 may choose an outside option and leave the game or not choose the outside option and stay in the game; actions that will be denoted by Out and In, respectively.
(2) When taking the outside option, player 1 receives utility $u_1(Out)$.
(3) If player 1 does not take the outside option, players 1 and 2 enter a simultaneous move game with action sets A_1 and A_2. In this subgame, there is a strict Nash equilibrium (a_1^*, a_2^*) which yields player 1 utility $u_1(a_1^*, a_2^*) > u_1(Out)$. All other Nash equilibria (μ_1, μ_2) in first-order beliefs yield player 1 an expected utility strictly lower than $u_1(Out)$.

In van Damme (1989) it is argued that (In, a_1^*) and a_2^* are the unique "forward induction strategies" in this game. The argument runs as follows. If player 2 observes that player 1 has not chosen the outside option, he should conclude that player 1 is heading for the only Nash equilibrium that dominates the outside option for him, that is, (a_1^*, a_2^*). As such, he should believe that player 1 will play a_1^*, and hence player 2 should respond with a_2^*. Player 1, anticipating on player 2 reasoning in this way, should therefore choose (In, a_1^*). In the following theorem, we prove that this argument is supported by the concept of minimal rationalizability.

Theorem 7.1. *Let $(\mathcal{S}, u)$ be an outside option game in the sense of van Damme (1989). Then, the unique minimally rationalizable strategies for $(\mathcal{S}, u)$ are the forward induction strategies (In, a_1^*) and a_2^*.*

Proof. Let h_1 and h_2 denote the information sets at which players 1 and 2 move in the simultaneous move game, respectively. We can construct types t_1^* and t_2^* with the following properties: (1) t_1^* and t_2^* have utility functions u_1 and u_2 throughout the game, where u_1 and u_2 are as specified by $(\mathcal{S}, u)$, (2) $\mu_1(t_1^*, h_0)$ and $\mu_1(t_1^*, h_1)$ assign probability one to strategy a_2^* and type t_2^*, and (3) $\mu_2(t_2^*, h_0)$ and $\mu_2(t_2^*, h_2)$ assign probability one to strategy (In, a_1^*) and type t_1^*. Then, it may be verified that t_1^* and t_2^* are persistently rationalizable for $(\mathcal{S}, u)$ with complexity

2, hence minimally rationalizable. Since $((In, a_1^*), t_1^*)$ and (a_2^*, t_2^*) are sequentially rational, it follows that (In, a_1^*) and a_2^* are minimally rationalizable.

Now, let t_1 be minimally rationalizable for $(\mathcal{S}, u)$. From the above, we must conclude that t_1 should have complexity 2, that is, $T^\infty(t_1) = \{t_1, t_2\}$ for some t_2. Hence, $\mu_1(t_1, h_0)$ and $\mu_1(t_1, h_1)$ assign probability one to t_2, whereas $\mu_2(t_2, h_0)$ and $\mu_2(t_2, h_2)$ assign probability one to t_1. Since t_1 has utility function u_1 throughout the game, it follows, in particular, that t_2 believes at h_2 that player 1 has utility function u_1 at h_0 and h_1. Moreover, since t_2 has utility function u_2, it follows that t_1 believes at h_1 that player 2 has utility function u_2 at h_2. Let μ_1 be the marginal of $\mu_2(t_2, h_2)$ on player 1's action set A_1 in the simultaneous move game, and let μ_2 be the marginal of $\mu_1(t_1, h_1)$ on A_2. Since both t_1 and t_2 believe in sequential rationality, we may conclude that (μ_1, μ_2) constitutes a Nash equilibrium in first-order beliefs in the subgame.

Since t_2 believes in sequential rationality, we know that $\mu_2(t_2, h_2)$ only assigns positive probability to player 1 strategies (In, a_1) that are optimal for t_1 at h_0. Hence, $\mu_2(t_2, h_2)(In, a_1) > 0$ only if (In, a_1) is optimal at h_0 given $\mu_1(t_1, h_0)$. Since t_1 satisfies updating consistency, we know that $\mu_1(t_1, h_0) = \mu_1(t_1, h_1)$. Recall that we have denoted the marginal of $\mu_1(t_1, h_1)$ on A_2 by μ_2, and the marginal of $\mu_2(t_2, h_2)$ on A_1 by μ_1. It thus follows that $\mu_1(a_1) > 0$ only if (In, a_1) is optimal at h_0 given μ_2. In particular, we have that $\mu_1(a_1) > 0$ only if $u_1(a_1, \mu_2) > u_1(Out)$, where $u_1(a_1, \mu_2)$ denotes the expected utility of playing a_1 when having belief μ_2 about player 2's strategy choice.

In summary, we thus have that (μ_1, μ_2) must constitute a Nash equilibrium, and $\mu_1(a_1) > 0$ only if $u_1(a_1, \mu_2) > u_1(Out)$. However, by the definition of an outside option game, there is only one Nash equilibrium (μ_1, μ_2) with this property, namely (a_1^*, a_2^*). Hence, we may conclude that $\mu_1(t_1, h_1)$ puts probability one on a_2^*, and $\mu_2(t_2, h_2)$ puts probability one on (In, a_1^*). By updating consistency, we then have that $\mu_1(t_1, h_0)$ puts probability one on a_2^*. The unique strategy that is sequentially rational for t_1 is (In, a_1^*), and hence, (In, a_1^*) is the unique minimally rationalizable strategy for player 1 in $(\mathcal{S}, u)$.

Now, suppose that t_2 is minimally rationalizable for $(\mathcal{S}, u)$. Then, t_2 must have complexity 2, which implies that $T^\infty(t_2) = \{t_1, t_2\}$ for some t_1. In particular, $\mu_2(t_2, h_0)$ and $\mu_2(t_2, h_2)$ must assign probability one to t_1. Since t_2 initially believes u, it follows that $u_1(t_1, h_0) = u_1$ and $u_1(t_1, h_1) = u_1|_{Z(h_1)}$. Together with the observation that t_1 must be persistently rationalizable and must respect common belief about the event that types initially believe u, this implies that t_1 is persistently rationalizable for $(\mathcal{S}, u)$. Since $T^\infty(t_2) = \{t_1, t_2\}$, we have that $T^\infty(t_1) = \{t_1, t_2\}$ as well, and hence we may conclude that t_1 is minimally rationalizable for $(\mathcal{S}, u)$. Above we have already seen that in such a case, (In, a_1^*) is the unique strategy which is sequentially rational for t_1. Since t_2 believes in sequential rationality, and $\mu_2(t_2, h_2)$ assigns probability one to t_1, it follows that $\mu_2(t_2, h_2)$ must assign probability one to the strategy (In, a_1^*). But then, a_2^* is the only strategy which is sequentially rational for t_2. Hence, a_2^* is the only minimally rationalizable strategy for player 2 in $(\mathcal{S}, u)$. This completes the proof. ■

8. Relation to Other Concepts

The purpose of this section is to compare persistent and minimal rationalizability with the concepts of weak sequential rationalizability (Ben-Porath (1997)) and extensive form rational-

izability (Pearce (1984) and Battigalli (1997)).

8.1. Weak Sequential Rationalizability

Formally speaking, the concept of weak sequential rationalizability as we use it, is an extension of the notion of *common certainty of rationality at the beginning of the game,* defined in Ben-Porath (1997) for the class of games with perfect information, to the general class of extensive form games. We first formally define weak sequential rationalizability within our epistemic framework. Let $\mathcal{S}$ be an extensive form structure and $u = (u_i)_{i \in I}$ a profile of utility functions. In the concept of weak sequential rationalizability, it is assumed that there be common belief about u *at every information set,* and not only at the beginning of the game. By the latter, we formally mean that every type respects common belief about the event that all types t_i, at every information set $h_i \in H_i^*$, hold the utility function $u_i(t_i, h_i) = u_i|_{Z(h_i)}$. Moreover, there is *initial* common belief about the event that types *initially* believe that players act sequentially rationally. In particular, a type is allowed to believe that an opponent is no longer acting sequentially rationally whenever this type finds out that the opponent has made a move that contradicts his initial beliefs. In order to state this property formally, we need the following definitions.

For a given type t_i and opponent j, let

$$T_j^1(t_i, h_0) = \text{supp}\mu_i(t_i, h_0 |\ T_j)$$

be the set of player j types that t_i assigns positive probability to at the beginning of the game. Let $T_i^1(t_i, h_0) = \{t_i\}$, and let

$$T^1(t_i, h_0) = \cup_{j \in I} T_j^1(t_i, h_0).$$

For all $k \geq 2$, let

$$T^k(t_i, h_0) = \cup_{t \in T^{k-1}(t_i, h_0)} T^1(t, h_0).$$

Define $T^\infty(t_i, h_0) = \cup_{k \in \mathbb{N}} T^k(t_i, h_0)$. For a given event $\tilde{T} \subseteq \times_{j \in I} T_j$, we say that t_i respects *initial* common belief about $\tilde{T}$ if $T^\infty(t_i, h_0) \subseteq \tilde{T}$.

For a given player i, recall that $(S_{-i} \times T_{-i})^{sr}$ denotes the set of sequentially rational opponents' strategy-type pairs. Let $T_i^{sr}(h_0)$ be the set of player i types t_i that *initially believe in sequential rationality*, that is, with $\text{supp}\mu_i(t_i, h_0) \subseteq (S_{-i} \times T_{-i})^{sr}$.

Definition 8.1. *A type t_i is said to be weakly sequentially rationalizable for the game $(\mathcal{S}, u)$ if (1) t_i holds utility function u_i, (2) respects common belief about u at every information set, and (3) respects initial common belief about the event that types initially believe in sequential rationality. A strategy s_i is called weakly sequentially rationalizable for $(\mathcal{S}, u)$ if there is a weakly sequentially rationalizable type t_i for $(\mathcal{S}, u)$ such that (s_i, t_i) is sequentially rational.*

We shall now prove that for every extensive form game, persistently rationalizable strategies are always weakly sequentially rationalizable. The other direction is not true, since it is well-known that weakly sequentially rationalizable strategies in a game with perfect information need not be backward induction strategies, while we have seen that persistent rationalizability always

yields backward induction strategies in such games. One may even find examples of games with perfect information in which a profile of weakly sequentially rationalizable strategies need not lead to the backward induction *outcome* (see for instance Figure 1 in Ben-Porath (1997)).

Theorem 8.2. *Let $\mathcal{S}$ be an extensive form structure and u a profile of utility functions. Then, every persistently rationalizable strategy for $(\mathcal{S}, u)$ is weakly sequentially rationalizable for $(\mathcal{S}, u)$.*

Proof. Let s_i be persistently rationalizable for $(\mathcal{S}, u)$. Then, there is some type t_i that is persistently rationalizable for $(\mathcal{S}, u)$ such that (s_i, t_i) is sequentially rational. By definition, t_i respects initial common belief about the event that types initially believe in sequential rationality. However, t_i need not respect common belief about u at every information set since t_i may revise his belief about the opponents' utility functions as the game proceeds. However, the type t_i may be transformed into a type $\tilde{t}_i$ such that (1) $\tilde{t}_i$ respects common belief about u at all information sets, and (2) at every information set h_i, the marginal probability distribution of $\mu_i(\tilde{t}_i, h_i)$ on the opponents' strategies coincides with the corresponding marginal probability distribution of $\mu_i(t_i, h_i)$. Then, by construction, $\tilde{t}_i$ has utility function u_i, respects initial common belief about the event that types initially believe in sequential rationality, and respects common belief about u at all information sets. Hence, $\tilde{t}_i$ is weakly sequentially rationalizable for $(\mathcal{S}, u)$. Moreover, since the beliefs of $\tilde{t}_i$ about the opponents' strategies coincide with t_i's beliefs at every information set h_i, and since (s_i, t_i) is sequentially rational, it follows that $(s_i, \tilde{t}_i)$ is sequentially rational. We may thus conclude that s_i is weakly sequentially rationalizable for $(\mathcal{S}, u)$, which completes the proof. ■

8.2. Extensive Form Rationalizability

The concept of extensive form rationalizability has been introduced in Pearce (1984) by means of an iterated elimination procedure. Later, Battigalli (1997) provided an alternative procedure that always leads to the same sets of strategies, whereas Battigalli and Siniscalchi (2002) give an epistemic foundation for the concept of extensive form rationalizability. Instead of delivering a precise definition of extensive form rationalizability, we shall restrict ourselves to a verbal expression of Battigalli and Siniscalchi's epistemic characterization in order to save space. This informal desciption will then be sufficient to prove that there is no general logical relationship between persistent rationalizability and extensive form rationalizability, at least in terms of strategy choices.

In the concept of extensive form rationalizability, the players' utility functions are never to be questioned during the game. Battigalli and Siniscalchi (2002) show that extensive form rationalizability can be characterized by the requirement that a player, at each of his information sets, should seek for the "highest possible degree of interactive belief in sequential rationality" that is still compatible with the event of reaching this information set, and should base his current and future beliefs upon this degree until it is contradicted at some later information set. As to illustrate the concept, consider the perfect information game in Figure 2, which is taken from Reny (1992b).

Suppose that player 2 observes that player 1 has chosen r_1 at his first move. Among the feasible strategies (r_1, r_3) and (r_1, d_3), only (r_1, r_3) can possibly be a sequentially rational strategy for player 1, given the restriction that player 2 should still believe at this point that player

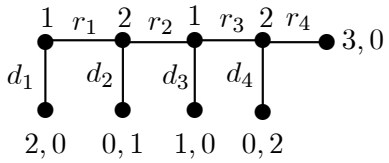

Figure 2

1's utility function is as depicted at the terminal nodes. As such, extensive form rationalizability requires player 2 to believe that player 1 has chosen (r_1, r_3) after observing r_1. Given these beliefs, player 2 should choose (r_2, d_4). If player 1 believes, at the beginning, that player 2 chooses (r_2, d_4), player 1 should choose d_1. Hence, d_1 and (r_2, d_4) are the unique extensive form rationalizable strategies in this game. Since the unique backward induction strategies are d_1 and d_2, we know by Theorem 5.1 that d_1 and d_2 are the unique persistently rationalizable (and hence unique minimally rationalizable) strategies in this game. We may thus conclude that a minimally rationalizable strategy need not be extensive form rationalizable, and an extensive form rationalizable strategy need not be persistently rationalizable.

In the perfect information game above, we see however that the concepts of extensive form rationalizability and persistent (minimal) rationalizability lead to the same unique outcome: the backward induction outcome $(2, 0)$. This is a structural phenomenon for perfect information games, since Battigalli (1997) has shown that in a generic game with perfect information, every profile of extensive form rationalizable strategies leads to the backward induction outcome.

There are other games, however, where minimal rationalizability leads to outcomes that cannot be reached by extensive form rationalizability. Consider, for instance, the Burning-Money game in Figure 3, which is due to van Damme (1989) and Ben-Porath and Dekel (1992). In this game, player 1 may choose between burning a dollar (*burn*) or not burning a dollar (*not*) at round 1, after which players 1 and 2 face a simultaneous move game at round 2. We show that (not, e) and (c, g) are the unique extensive from rationalizable strategies, leading to player 1's most preferred outcome $(3, 1)$. Suppose that player 2 observes *burn*. Among the feasible strategies $(burn, a)$ and $(burn, b)$, only $(burn, a)$ can possibly be sequentially rational, and hence extensive form rationalizability imposes that player 2 should believe that player 1 chooses $(burn, a)$ after observing *burn*. As such, player 2 should choose c after observing *burn*. Player 1, anticipating on player 2 choosing c after *burn*, can thus guarantee 2 by choosing $(burn, a)$. Player 2, realizing this, should thus believe that player 1 is choosing (not, e) after observing *not*, since (not, f) cannot give player 1 more than 2. Hence, player 2 should choose g after *not*, which makes player 1 choosing (not, e) at the beginning of the game. The only strategies that remain are thus (not, e) for player 1 and (c, g) for player 2.[5]

We will show, however, that minimal (and hence persistent) rationalizability may lead to

[5]Shimoji (2002) shows that in the more general class of Burning-Money games discussed in Ben-Porath and Dekel (1992), extensive form rationalizability always leads to the "forward induction outcome", that is, player 1's most preferred outcome in the game.

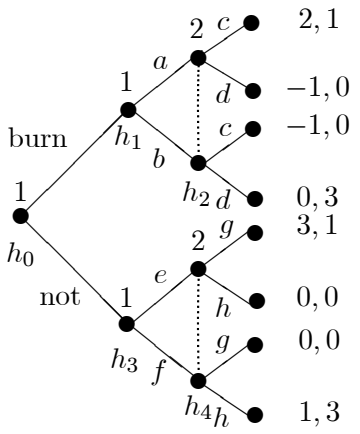

Figure 3

outcomes other than $(3, 1)$. As a preparatory step, we first prove that any persistently rationalizable type in $(\mathcal{S}, u)$ should have a complexity strictly greater than 2. Assume, on the contrary, that t_1 would be a persistently rationalizable type in $(\mathcal{S}, u)$ with complexity 2. Then, there is some type t_2 such that $T^\infty(t_1) = \{t_1, t_2\}$, that is, t_1 believes at h_0 that player 2 has type t_2, and that t_2 believes at h_2 and h_4 that player 1 has type t_1. In particular, t_2 believes at h_2 and h_4 that player 1 has utility function u_1 as specified at the terminal nodes in Figure 3. Since t_2 should believe at h_2 that t_1 chooses sequentially rationally, t_2 should believe at h_2 that t_1 chooses $(burn, a)$. Accordingly, t_1 should believe at h_0 that t_1 chooses c after $burn$, and hence t_1's expected utility by choosing $(burn, a)$ is 2. Since t_2 should believe at h_4 that t_1 chooses sequentially rationally, t_2 should believe at h_4 that t_1 chooses (not, e). Hence, t_1 should believe that t_2 chooses g after not, and hence t_1's expected utility by choosing (not, a) is 3, which is greater than his expected utility by choosing $(burn, a)$ or $(burn, b)$. Therefore, t_2 cannot believe at h_2 that t_1 chooses sequentially rationally, and hence t_1 cannot be persistently rationalizable in $(\mathcal{S}, u)$ while having complexity 2. Similarly, one may show that no persistently rationalizable type t_2 in $(\mathcal{S}, u)$ can have complexity 2.

Now, let the utility functions u_1 and u_2 be as specified in Figure 3, and let $\tilde{u}_1$ be the utility function that coincides with u_1 at $Z(h_3)$, and for which $\tilde{u}_1(z) = u_1(z) - 2$ for all $z \in Z(h_1)$. Let the types $t_1, \tilde{t}_1$ and t_2 be such that:
(1) t_1 has utility function u_1 and $\mu_1(t_1, h_0), \mu_1(t_1, h_1)$ and $\mu_1(t_1, h_3)$ assign probability one to $((c, h), t_2)$;
(2) $\tilde{t}_1$ has utility function $\tilde{u}_1$ and $\mu_1(t_1, h_0), \mu_1(t_1, h_1)$ and $\mu_1(t_1, h_3)$ assign probability one to $((c, h), t_2)$;
(3) t_2 has utility function u_2, $\mu_2(t_2, h_0)$ and $\mu_2(t_2, h_2)$ assign probability one to $((burn, a), t_1)$, and $\mu_2(t_2, h_4)$ assigns probability one to $((not, f), \tilde{t}_1)$.

It may be verified that t_1 and t_2 are persistently rationalizable for $(\mathcal{S}, u)$. Since both have complexity 3, it follows that both t_1 and t_2 are minimally rationalizable for $(\mathcal{S}, u)$. The strategies

$(burn, a)$ and (c, h) are sequentially rational for t_1 and t_2, and hence $(burn, a)$ and (c, h) are minimally rationalizable for $(\mathcal{S}, u)$, leading to the outcome $(2, 1)$. We may thus conclude that not every minimally rationalizable outcome is extensive form rationalizable in this game.

There are also games in which not every extensive form rationalizable outcome is minimally rationalizable. Consider the following two-player simultaneous move game (see Perea (2001), p.204) represented by its normal form.

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

In this game $(\mathcal{S}, u)$, every strategy is rationalizable. Since extensive form rationalizability coincides with rationalizability in simultaneous move games, it follows that all strategies are extensive form rationalizable. However, strategy c is not a Nash equilibrium strategy. Suppose, on the contrary, that c were a Nash equilibrium strategy. Then, there should be some Nash equilibrium $(\mu_1, \mu_2) \in \Delta(S_1) \times \Delta(S_2)$ in first-order beliefs such that c is a best response to μ_2. This implies that $\mu_2(d) > 0$ and $\mu_2(e) > 0$. Hence, both d and e should be a best response to μ_1, which is impossible. Consequently, c cannot be a Nash equilibrium strategy. Since we know from Theorem 6.1 that the set of minimally rationalizable strategies coincides with the set of Nash equilibrium strategies in every two-player simultaneous move game, it follows that c is not minimally rationalizable. In particular, the outcomes $(c, d), (c, e)$ and (c, f) are extensive form rationalizable but not minimally rationalizable.

At this stage, it remains an open question whether in any given game, every extensive form rationalizable *outcome* is a persistently rationalizable outcome. Up to this point, I have not been able to provide a counterexample, nor to produce a general proof.

References

[1] Anscombe, F.J. and R. Aumann (1963), A definition of subjective probability, *Annals of Mathematical Statistics* **34,** 199-205.

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

[3] Asheim, G.B. (2000), On the epistemic foundation for backward induction, Memorandum No. 30, Department of Economics, University of Oslo.

[4] Asheim, G.B. (2001), Proper rationalizability in lexicographic beliefs, *International Journal of Game Theory* ***30,*** 453-478.

[5] Asheim, G.B. and A. Perea (2002), Sequential and quasi-perfect rationalizability in extensive games, University of Oslo and Maastricht University.

[6] Aumann, R. (1995), Backward induction and common knowledge of rationality, *Games and Economic Behavior* **8,** 6-19.

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

[8] Balkenborg, D. and E. Winter (1997), A necessary and sufficient epistemic condition for playing backward induction, *Journal of Mathematical Economics* **27,** 325-345.

[9] Battigalli, P. (1996), Strategic independence and perfect Bayesian equilibria, *Journal of Economic Theory* **70,** 201-234.

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

[11] Battigalli, P. (2003), Rationalizability in infinite, dynamic games with incomplete information, *Research in Economics* **57,** 1-38.

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

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

[14] Ben-Porath, E. (1997), Rationality, Nash equilibrium and backwards induction in perfect-information games, *Review of Economic Studies* **64,** 23-46.

[15] Ben-Porath, E. and E. Dekel (1992), Signaling future actions and the potential for sacrifice, *Journal of Economic Theory* **57,** 36-51.

[16] Blume, L.E., Brandenburger, A. and E. Dekel (1991a), Lexicographic probabilities and choice under uncertainty, *Econometrica* **59,** 61-79.

[17] Blume, L.E., Brandenburger, A. and E. Dekel (1991b), Lexicographic probabilities and equilibrium refinements, *Econometrica* **59,** 81-98.

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

[19] Dekel, E. and D. Fudenberg (1990), Rational behavior and payoff uncertainty, *Journal of Economic Theory* **52,** 243-267.

[20] Epstein, L. and T. Wang (1996), "Beliefs about beliefs" without probabilities, *Econometrica* **64,** 1343-1373.

[21] Fudenberg, D., Kreps, D. and D. Levine (1988), On the robustness of equilibrium refinements, *Journal of Economic Theory* **44,** 354-380.

[22] Harsanyi, J. (1973), Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points, *International Journal of Game Theory* **2,** 1-23.

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

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

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

[26] Perea, A. (2001), *Rationality in Extensive Form Games,* Kluwer Academic Publishers, Boston, Dordrecht.

[27] Perea, A. (2002), Forward induction and the minimum revision principle, Meteor Research Memorandum RM/02/010, Maastricht University.

[28] Perea, A. (2003), Proper rationalizability and belief revision in dynamic games, Maastricht University.

[29] Reny, P.J. (1992a), Rationality in extensive-form games, *Journal of Economic Perspectives* **6,** 103-118.

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

[31] Reny, P.J. (1993), Common belief and the theory of games with perfect information, *Journal of Economic Theory* **59,** 257-274.

[32] Rubinstein, A. (1991), Comments on the interpretation of game theory, *Econometrica* **59,** 909-924.

[33] Samet, D. (1996), Hypothetical knowledge and games with perfect information, *Games and Economic Behavior* **17,** 230-251.

[34] Savage, L.J. (1954), *The Foundations of Statistics,* Wiley, New York.

[35] Schuhmacher, F. (1999), Proper rationalizability and backward induction, *International Journal of Game Theory* **28,** 599-615.

[36] Shimoji, M. (2002), On forward induction in money-burning games, *Economic Theory* **19,** 637-648.

[37] Stalnaker, R. (1998), Belief revision in games: forward and backward induction, *Mathematical Social Sciences* **36,** 31-56.

[38] Stinchcombe, M.B. and K.G. Zauner (2002), Inference preserving perturbations of extensive-form games: the agent normal form approach, Mimeo.

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

[40] van Damme, E. (1989), Stable equilibria and forward induction, *Journal of Economic Theory* **48,** 476-496.

[41] Zauner, K.G. (2002), The existence of equilibrium in games with randomly perturbed payoffs and applications to experimental economics, *Mathematical Social Sciences* **44,** 115-120.