Rationalizability and Minimal Complexity in Dynamic Games

Andrés Perea*
Maastricht University

May 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 two key conditions upon the players' beliefs: (a) throughout the game, every move by the opponent should be interpreted as a rational move, and (b) the belief about the opponents' relative utilities between two terminal nodes should only be revised if you are sure that the opponent has decided to avoid one of these nodes. Common belief about these events leads to the concept of *persistent rationalizability*. It is shown that persistent rationalizability implies the backward induction procedure in generic games with perfect information. We next focus on persistently rationalizable types having beliefs with "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.

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 the opponents' utility functions. 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.

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

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 (1992, 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 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 preference relations.

Our first result is a representation theorem similar to Armbruster and Böge (1979), Böge and Eisele (1979) and Mertens and Zamir (1985) which shows that the infinite preference hierarchies within our epistemic model can be handled elegantly by means of *types.* 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 *belief revision consistency,* 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. Belief revision consistency states that, whenever a player i type decides to revise his belief about player j's utility function, he should not change the relative utilities between two terminal nodes unless player i is sure that

player j has decided to avoid one of these terminal nodes.

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.

The last condition, *utility consistency,* states that the utility function of a type at a certain stage of the game should always be in accordance with the utility function he held at the beginning of the game. Types that, throughout the game, respect common belief (1) about the event that types are updating consistent, belief revision consistent and utility consistent, and (2) about the event that types believe in sequential rationality, are called persistently rationalizable. Strategies that may be chosen rationally by persistently rationalizable types are called persistently rationalizable strategies.

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. 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.

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 game. 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 throughout the game. 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 outline of this paper is as follows. Section 2 presents some preliminary definitions in

extensive form games. In Section 3 we develop 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, and discusses the relationship between proper rationalizability (Schuhmacher (1999) and Asheim (2001)) and persistent rationalizability. It is also shown that for every profile of utility functions there exists at least one persistently rationalizable strategy for each player. Section 5 focusses on the relationship between persistent and minimal rationalizability on the one hand, and backward induction, Nash equilibrium and forward induction in outside option games on the other hand.

2. Extensive Form Structures

In this section we present the notation and some basic definitions in extensive form games that will be employed throughout this paper. The rules of the game are represented by an *extensive form structure* $\mathcal{S}$ consisting of a finite game tree, a finite set of players I, a finite collection H_i of information sets for each player i and at each information set $h_i \in H_i$ a finite collection $A(h_i)$ of actions for the player. The set of terminal nodes in $\mathcal{S}$ is denoted by Z, whereas $H = \cup_{i \in I} H_i$ denotes the collection of all information sets. We assume throughout that the extensive form structure satisfies perfect recall and that no chance moves occur. The latter assumption is not crucial for our analysis, but simplifies the presentation. Let S_i denote the set of player i (pure) strategies, and let $S = \times_{i \in I} S_i$ be the set of all strategy profiles.

Throughout the paper, we shall make the assumption that the extensive form structure is with *observable deviators* (see Battigalli (1996), among others). In order to formalize this condition, we need the following definitions. For a given information set h, let $S(h)$ be the set of strategy profiles that reach h. For a given player i, not necessarily the player who moves at h, let $S_i(h)$ be the set of strategies s_i that do not avoid h. We say that $\mathcal{S}$ is with observable deviators if $S(h) = \times_{i \in I} S_i(h)$ for every information set h. That is, an information set h can only be avoided if there is at least one player who chooses a strategy that already avoids h by itself.

3. Epistemic Framework

In this section we formally model the players in an extensive form structure as decision makers under uncertainty. In order to do so, we first introduce some preliminary decision theoretic and epistemic concepts upon which this model shall be built.

3.1. Preference Hierarchies

The decision theoretic framework to be presented here is based on the models by Savage (1954) and Anscombe and Aumann (1963) for decision making under uncertainty. Let X be a compact metric space provided with some topology, and Y some finite set. Let $\Delta(Y)$ denote the space of probability distributions on Y. By $\mathcal{F}(X, Y)$ we denote the set of all measurable functions $f : X \to \Delta(Y)$ to which we shall refer as *acts*[1]. The set X is to be interpreted as the space

[1]The definition of an act as we use it coincides with the notion of *compound horse lottery* in Anscombe and Aumann (1963).

of relevant variables about which the decision maker has uncertainty, whereas Y represents the set of possible consequences. As such, $\Delta(Y)$ contains all objective lotteries on possible consequences. For a given act f in $\mathcal{F}(X,Y)$ and $x \in X$, let $f(x) \in \Delta(Y)$ be the objective lottery induced by x on Y, and let $f(x)(y)$ be the objective probability that $f(x)$ assigns to consequence y. By $\mathcal{P}^{eu}(X,Y)$ we denote the set of all nontrivial preference relations on $\mathcal{F}(X,Y)$ that are of the expected utility type, that is, for which is there is some probability distribution μ on X and some nonconstant utility function $u : Y \to \mathbb{R}$ such that act f is weakly preferred over act g if and only if

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

Here,

$$u(f(x)) = \sum_{y \in Y} f(x)(y)\ u(y)$$

denotes the expected utility induced by the objective lottery $f(x) \in \Delta(Y)$ and the utility function u.

Since for a given preference relation $p \in \mathcal{P}^{eu}(X,Y)$, the probability distribution μ is unique and the utility function u is unique up to some positive affine transformation, we may uniquely identify every $p \in \mathcal{P}^{eu}(X,Y)$ with a pair (μ, u) where μ is a subjective probability distribution on X and $u : Y \to \mathbb{R}$ with $\min_{y \in Y} u(y) = 0$ and $\max_{y \in Y} u(y) = 1$. Let $U(Y)$ be the set of all utility functions $u : Y \to \mathbb{R}$ with the latter property, and let $\Delta(X)$ be the set of probability distributions on X. Hence, we may identify $\mathcal{P}^{eu}(X,Y)$ with the set $\Delta(X) \times U(Y)$. Let τ_1 be the weak topology on $\Delta(X)$, let τ_2 be the natural topology on $U(Y)$ and τ the product topology on $\mathcal{P}^{eu}(X,Y)$ induced by τ_1 and τ_2. Then, the topological space $(\mathcal{P}^{eu}(X,Y), \tau)$ is a compact metric space.

Having established the model for individual decision making under uncertainty, we may now formalize an epistemic model for extensive form games in which players, at each of their information sets, have uncertainty about the opponents' strategy choices, uncertainty about the opponents' first-order preference relations (including their utility functions), uncertainty about the opponents' second-order preference relations, and so forth. This will eventually lead to the concept of *preference hierarchies* for players. The epistemic model combines elements from Epstein and Wang (1996) and Battigalli and Siniscalchi (1999). Epstein and Wang (1996) propose a model for *static* games in which players have uncertainty about the opponents' preference relations (possibly including the opponents' utility functions) and players may hold preference relations that do not conform to expected utility. Battigalli and Siniscalchi (1999), in turn, propose a model for *dynamic* games in which players hold expected utility preferences, players have no doubts about the opponents' utility functions but have uncertainty about the opponents' subjective probability distributions.

Consider some player i in an extensive form structure. Let h_0 be the information set that coincides with the beginning of the game, and let $H_i^* = H_i \cup \{h_0\}$. The primary source of uncertainty faced by player i at information set $h_i \in H_i^*$ concerns the strategy choices by his opponents. We may thus define the first-order space of uncertainty $X_i^1(h_i)$ by

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

where $S_{-i}(h_i) = \times_{j \neq i} S_j(h_i)$. If $h_i = h_0$, we set $S_j(h_i) = S_j$ for all players j. Let $Z(h_i)$ be the set of terminal nodes that follow h_i. Every player i strategy $s_i \in S_i(h_i)$ may now be identified with an act $f_{s_i} : X_i^1(h_i) \to Z(h_i)$ assigning to every $s_{-i} \in X_i^1(h_i)$ the terminal node $z \in Z(h_i)$ reached by the strategy profile (s_i, s_{-i}). Hence, every strategy $s_i \in S_i(h_i)$ corresponds to an act in $\mathcal{F}(X_i^1(h_i), Z(h_i))$. We assume that player i holds a nontrivial preference relation of the expected utility type $p_i^1(h_i) \in \mathcal{P}^{eu}(X_i^1(h_i), Z(h_i))$. We refer to $\mathcal{P}^{eu}(X_i^1(h_i), Z(h_i))$ as the set of *first-order* preference relations for player i at h_i.

At information set h_i, player i does not only have uncertainty about the strategies chosen by the opponents, but also about the first-order preference relations held by his opponents at each of their information sets. The second-order space of uncertainty for player i at h_i is therefore given by

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

which, together with the product topology induced by the topologies on $X_i^1(h_i)$ and $\mathcal{P}^{eu}(X_j^1(h_j), Z(h_j))$, is a compact metric space.

By the same argument as above, player i at h_i is assumed to hold a *second-order* preference relation $p_i^2(h_i) \in \mathcal{P}^{eu}(X_i^2(h_i), Z(h_i))$. Since player i has uncertainty about the second-order preference relations held by the other players at each of their information sets, the third-order space of uncertainty at h_i becomes

$$X_i^3(h_i) = X_i^2(h_i) \times (\times_{j \neq i} \times_{h_j \in H_j^*} \mathcal{P}^{eu}(X_j^2(h_j), Z(h_j))),$$

which, together with the induced product topology, is again a compact metric space. By repeating this construction, we obtain an infinite sequence of "succesively richer" spaces of uncertainty, defined 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 $k \geq 2$, which are all compact metric spaces.

A *preference hierachy* for player i at h_i is a sequence $p_i(h_i) = (p_i^k(h_i))_{k \in \mathbb{N}}$ where $p_i^k(h_i) \in \mathcal{P}^{eu}(X_i^k(h_i), Z(h_i))$ for all k. Hence, it specifies an infinite hierarchy of expected utility preference relations over succesively richer spaces of uncertainty. A vector $p_i = (p_i(h_i))_{h_i \in H_i^*}$, specifying a preference hierarchy at each of player i information sets, is simply called a *preference hierarchy* for player i. Let P_i be the set of all preference hierarchies for player i.

3.2. Coherence of Preference Hierarchies

A preference hierarchy p_i is called *coherent* if it holds a sequence of preference relations that do not contradict one another at overlapping layers. Let P_i^c be the set of coherent preference hierarchies for player i, and let $P_{-i} = \times_{j \neq i} P_j$ be the set of all opponents' preference hierarchies.

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

Hence, there is a homeomorphism f_i from P_i^c to $\times_{h_i \in H_i^*} \mathcal{P}^{eu}(S_{-i}(h_i) \times P_{-i}, Z(h_i))$ for every player i. Hence, every preference hierarchy $p_i \in P_i^c$ can be identified with the vector

$$f_i(p_i) = (\mu_i(p_i, h_i), u_i(p_i, h_i))_{h_i \in H_i^*}$$

where $\mu_i(p_i, h_i) \in \Delta(S_{-i}(h_i) \times P_{-i})$ and $u_i(p_i, h_i) \in U(Z(h_i))$. A subset $E \subseteq S_{-i}(h_i) \times P_{-i}$ is called an *event* at information set h_i. We say that preference hierarchy $p_i \in P_i^c$ *believes* the event E at information set h_i if

$$\text{supp } \mu_i(p_i, h_i) \subseteq E.$$

We do not only require that every preference hierarchy is coherent, but also that there be *common belief* among the players that all preference hierarchies are coherent. This may be formalized as follows. Let $P_{-i}^c = \times_{j \neq i} P_j^c$. Define the sets $P_i^{c,1}, P_i^{c,2}, ...$ by

$$\begin{aligned} P_i^{c,1} &= \{p_i \in P_i^c |\ p_i \text{ believes } S_{-i}(h_i) \times P_{-i}^c \text{ at every } h_i \in H_i^*\}, \\ P_i^{c,k} &= \{p_i \in P_i^{c,k-1} |\ p_i \text{ believes } S_{-i}(h_i) \times P_{-i}^{c,k-1} \text{ at every } h_i \in H_i^*\} \end{aligned}$$

for $k \geq 2$. Define $P_i^{c,\infty} = \cap_{k \in \mathbb{N}} P_i^{c,k}$ for all players i. We say that $P_i^{c,\infty}$ is the set of preference hierarchies for player i which *respect common belief of coherence.* We now obtain the following representation result for infinite preference hierarchies respecting common belief of coherence. The result is similar in spirit to results in Armbruster and Böge (1979), Böge and Eisele (1979), Mertens and Zamir (1985) and Epstein and Wang (1996).

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

3.3. Types, Common Belief and Complexity

In view of Lemma 3.2, we may identify each preference hierarchy $p_i \in P_i^{c,\infty}$ with a vector specifying at each information set $h_i \in H_i^*$ an expected utility preference relation $(\mu_i(p_i, h_i), u_i(p_i, h_i))$ where $\mu_i(p_i, h_i)$ is a probability measure on $S_{-i}(h_i) \times P_{-i}^{c,\infty}$ and $u_i(p_i, h_i)$ is a utility function from $Z(h_i)$ to the real numbers. A preference hierarchy $p_i \in P_i^{c,\infty}$ is called a *type* for player i, and by $T_i = P_i^{c,\infty}$ we denote the set of all player i types. Hence, every type $t_i \in T_i$ corresponds to a vector $(\mu_i(t_i, h_i), u_i(t_i, h_i))_{h_i \in H_i^*}$ where $\mu_i(t_i, h_i)$ is a probability distribution on $S_{-i}(h_i) \times T_{-i}$ and $u_i(t_i, h_i)$ is a utility function on $Z(h_i)$ for every information set $h_i \in H_i^*$. Using Lemma 3.2, we thus obtain the following representation result for types.

Corollary 3.3. *For every player i, the space T_i of player i types is homeomorphic to the space $\times_{h_i \in H_i^*} \mathcal{P}^{eu}(S_{-i}(h_i) \times T_{-i}, Z(h_i))$.*

We now formalize what it means that a type respects *common belief* about the event that types have certain properties. In order to do so, we use the following definitions. For a given type t_i, information set $h_i \in H_i^*$, and opponent j, let $\mu_i(t_i, h_i |\ T_j)$ be the marginal of the probability distribution $\mu_i(t_i, h_i)$ on the set of player j types. By

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

we denote the set of player j types that t_i attaches positive probability to at h_i, whereas

$$T_j^1(t_i) = \cup_{h_i \in H_i^*} T_j^1(t_i, h_i)$$

is the set of player j types that t_i attaches positive probability to somewhere in the game. For $j = i$, we define $T_i^1(t_i) = \{t_i\}$. Let

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

Hence, in some sense, $T^1(t_i)$ is the set of types that t_i uses *at the first level* in his theory about the opponents' behavior and beliefs. In turn, the behavior of each of the types t in $T^1(t_i)$ is driven by the beliefs that t has about the other players' types throughout the game. More exactly, every $t \in T^1(t_i)$ uses the set $T^1(t)$ of types at a first level for his theory about the other players' behavior and beliefs. By

$$T^2(t_i) = \bigcup_{t \in T^1(t_i)} T^1(t)$$

we denote the set of types that t_i uses, *at a first or second level*, for his theory about the game. By repeating this argument recursively, we obtain that

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

for $k \geq 2$ represents the set of types that t_i uses, up to level k, in his theory about the game. By $T^\infty(t_i) = \cup_{k \in \mathbb{N}} T^k(t_i)$ we denote the set of types that t_i uses, directly or indirectly, for his theory about the opponents' strategy choices and beliefs, and upon which t_i shall base his final decision. By

$$c(t_i) = |T^\infty(t_i)|$$

we denote the *complexity* of the type t_i. Hence, it specifies how many different types are used by t_i in his theory about the opponents' decisions and the opponents' beliefs. For every player j, let $T_j^\infty(t_i)$ be the set of player j types in $T^\infty(t_i)$. Note that $T_i^\infty(t_i)$ may contain more types than t_i, since t_i may believe that his opponents believe that player i has some other type than t_i.

Now, let $\tilde{T} \subseteq \times_{j \in I} T_j$ be some set of profiles of types, or, simply, and event. We say that type t_i *respects common belief about* $\tilde{T}$ if $T^\infty(t_i) \subseteq \tilde{T}$. That is, in his theory about the opponents' behavior and the opponents' beliefs, type t_i only uses, directly or indirectly, types that belong to $\tilde{T}$. Or, in other words, t_i believes that all opponents' types belong to $\tilde{T}$, believes that all opponents' types believe that all the other players' types belong to $\tilde{T}$, and so forth.

4. Persistent and Minimal Rationalizability

4.1. Persistent Rationalizability

In the concept of persistent rationalizability we impose four conditions on types, to which we refer as common belief about *updating consistency, utility consistency, belief revision consistency*

and *belief in sequential rationality.* Types that satisfy these requirements are called *persistently rationalizable,* and strategies that are sequentially optimal for a persistently rationalizable type are called *persistently rationalizable strategies.*

In the previous section, we have seen that every type $t_i \in T_i$ corresponds to a vector $(\mu_i(t_i, h_i), u_i(t_i, h_i))_{h_i \in H_i^*}$, where $\mu_i(t_i, h_i)$ is a probability measure on $S_{-i}(h_i) \times T_{-i}$ and $u_i(t_i, h_i)$ is a utility function on $Z(h_i)$ for every information set $h_i \in H_i^*$. Updating consistency states that, whenever the game moves from a player i information set h_i^1 to another player i information set h_i^2, player i should derive his new belief $\mu_i(t_i, h_i^2)$ from his old belief $\mu_i(t_i, h_i^1)$ by Bayesian updating, if possible.

Definition 4.1. *A type t_i is said to be updating consistent if for all information sets $h_i^1, h_i^2 \in H_i^*$, where h_i^2 follows h_i^1, it holds that*

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

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

Utility consistency simply states that the utility function of a type should be time-consistent in the sense that his utility function at some later stage in the game should not contradict his utility function earlier on in the game.

Definition 4.2. *A type t_i is called utility consistent if $u_i(t_i, h_i) = u_i(t_i, h_0)|_{Z(h_i)}$ for all $h_i \in H_i^*$.*

Here, $u_i(t_i, h_0)|_{Z(h_i)}$ denotes the restriction of $u_i(t_i, h_0)$ on the terminal nodes following h_i.

While updating consistency states how to change the belief when the observed behavior is still in accordance with the previously held beliefs, *belief revision consistency* imposes a condition upon the players' belief revision policies when the observed behavior contradicts the previous beliefs. In words, the condition states that, whenever player i at some information set h_i is led to revise his beliefs about some opponent j's utility function, then he should not change his belief about j's relative utilities between two terminal nodes unless i is certain that j has decided to avoid one of these nodes. More precisely, if player i finds himself at h_i, then, by the observable deviators property, player i knows that opponent j has chosen some strategy in $S_j(h_i)$, without knowing which one. Let $Z_j(h_i)$ be the set of terminal nodes that may be reached by strategies in $S_j(h_i)$. Hence, the event of reaching information set h_i only tells player i that player j has decided to avoid terminal nodes that are not in $Z_j(h_i)$. Belief revision consistency then states that player i may only revise his belief about j's relative utilities between two nodes if at least one of these nodes is not in $Z_j(h_i)$.

Definition 4.3. *A type t_i is said to be belief revision consistent if for every two information sets $h_i^1, h_i^2 \in H_i^*$ such that h_i^2 follows h_i^1 the following holds: if $t_j^2 \in \text{supp}\mu_i(t_i, h_i^2 |\ T_j)$ then there exists some $t_j^1 \in \text{supp}\mu_i(t_i, h_i^1 |\ T_j)$ such that $u_j(t_j^2, h_j)|_{Z_j(h_i^2)} = u_j(t_j^1, h_j)|_{Z_j(h_i^2)}$ for every $h_j \in H_j^*$.*

Here, $\mu_i(t_i, h_i^1|\ T_j)$ and $\mu_i(t_i, h_i^2|\ T_j)$ denote the marginals of the probability distributions $\mu_i(t_i, h_i^1)$ and $\mu_i(t_i, h_i^2)$ on T_j, and $u_j(t_j^1, h_j)|_{Z_j(h_i^2)}$ and $u_j(t_j^2, h_j)|_{Z_j(h_i^2)}$ denote the restrictions of the utility functions $u_j(t_j^1, h_j)$ and $u_j(t_j^2, h_j)$ on the terminal nodes in $Z(h_j) \cap Z_j(h_i^2)$.

We finally define belief in sequential rationality. For a given strategy s_i, let $H_i^*(s_i)$ be the set of information sets in H_i^* that are not avoided by s_i. A strategy-type pair $(s_i, t_i) \in S_i \times T_i$ is called *sequentially rational* if at every information set $h_i \in H_i^*(s_i)$, we have that

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

Here, $u_i(s_i, t_i|\ h_i)$ denotes the expected utility induced by strategy s_i with respect to the probability distribution $mrg(\mu_i(t_i, h_i)|\ S_{-i}(h_i)) \in \Delta(S_{-i}(h_i))$ and the utility function $u_i(t_i, h_i)$. Let $(S_i \times T_i)^{sr}$ be the set of sequentially rational strategy-type pairs, and let $(S_{-i} \times T_{-i})^{sr} = \times_{j \neq i}(S_j \times T_j)^{sr}$. By

$$T_i^{sr} = \{t_i \in T_i|\ \text{supp}\mu_i(t_i, h_i) \subseteq\ (S_{-i} \times T_{-i})^{sr} \text{ for every } h_i \in H_i^*\}$$

we denote the set of those player i types that *believe in sequential rationality.*

Definition 4.4. *A type $t_i \in T_i$ is called persistently rationalizable if it respects common belief about the events that (1) types are updating consistent, (2) types are utility consistent, (3) types are belief revision consistent, and (4) types believe in sequential rationality. A strategy $s_i \in S_i$ is called persistently rationalizable if there is some persistently rationalizable type t_i such that (s_i, t_i) is sequentially rational.*

4.2. Exogenous Restrictions on Utility Functions and Beliefs

In the literature, it is usually assumed that there be common belief about the players' actual utility functions throughout the game, and these utility functions are usually modeled as being part of the extensive form game itself. Indeed, an *extensive form game* is normally defined as a pair $(\mathcal{S}, u)$, where $\mathcal{S}$ is an extensive form structure and $u = (u_i)_{i \in I}$ is a profile of utility functions, the interpretation being that at any stage of the game, there is common belief about u. Therefore, if we wish to compare our concept of persistent rationalizability to other rationalizability concepts proposed in the literature, we should formalize what it means in our model that players "face an exogenously given profile of utility functions", while allowing these players to revise their belief about the opponents' utility functions as the game proceeds.

Let $\mathcal{S}$ be an extensive form structure and $u = (u_i)_{i \in I}$ an exogenously given profile of utility functions. We say that a type t_i *initially believes* u if $\mu_i(t_i, h_0)$ assigns probability one to the event that every opponent j has some type t_j with $u_j(t_j, h_j) = u_j|_{Z(h_j)}$ for all $h_j \in H_j^*$.

Definition 4.5. *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.*

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, have a complexity that is as small as possible. Recall from Section 3.3 that the complexity of a type t_i denotes the total number of types that t_i uses in his theory about the opponents.

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 $(\mathcal{S}, u)$ such that (s_i, t_i) is sequentially rational.*

Reducing the complexity of a persistently rationalizable type t_i to a minimum typically implies that t_i should involve as few belief revisions as possible during the course of the game. The reason is that belief revisions typically increase the complexity of a type. In order to see this, note that a belief revision is necessary whenever t_i, at some information set h_i^1, holds some belief about the possible types of opponent j, and discovers at a later information set h_i^2 that none of these types could have chosen a strategy leading to h_i^2 that is in accordance with common belief of sequential rationality. In such a case, t_i should include new player j types to his theory, at least at a first level, in order to explain the event of reaching information set h_i^2, and this will typically increase the complexity of t_i.

Moreover, minimizing the complexity of a type t_i also typically implies that t_i should restrict his attention, in his theory about the opponents, as much as possible to types that use few belief revisions. The reason should be clear: if t_i, directly or indirectly, uses a type t_j with many belief revisions in his theory, then t_j's complexity will typically be large, which in turn would contribute to a larger complexity of type t_i.

4.4. Existence and Relation to Proper Rationalizability

Schuhmacher (1999) introduced the concept of *proper rationalizabily* as some non-equilibrium analogue to proper equilibrium, and showed that it uniquely selects the backward induction strategies in generic games with perfect information. Subsequently, Asheim (2001) provided a characterization of proper rationalizability in terms of lexicographic beliefs for the case of two players, which can be extended to games with more than two players. Asheim's characterization states, in words, that a properly rationalizable type should respect common belief about the event that (1) types take all opponents' strategies into account, and (2) types deem one opponent strategy infinitely more likely than some other strategy whenever the opponent prefers the former over the latter. We show that for a given extensive form structure $\mathcal{S}$ and profile u of utility functions, every properly rationalizable strategy for $(\mathcal{S}, u)$ is persistently rationalizable for $(\mathcal{S}, u)$. Since properly rationalizable strategies always exist for every $(\mathcal{S}, u)$, this result implies the existence of persistently rationalizable strategies for every $(\mathcal{S}, u)$.

Theorem 4.7. *Let $\mathcal{S}$ be an extensive form structure with observable deviators and $u = (u_i)_{i \in I}$ a profile of utility functions. Then, every properly rationalizable strategy for $(\mathcal{S}, u)$ is persistently rationalizable for $(\mathcal{S}, u)$.*

5. Relation to Other Concepts

5.1. Backward Induction

In this section we will see 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 $(\mathcal{S}, u)$.*

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.

5.2. Nash Equilibrium Strategies

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' strategy choices and opponents' beliefs. It turns out 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). Minimal rationalizability thus restricts attention to those strategies by player i that can be justified by an epistemic rationalizability theory (cf. Tan and Werlang (1988)) which involves as few types as possible. We 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 5.2. *Let (S, u) be a two-player simultaneous move game. Then, s_i is minimally rationalizable for (S, u) if and only if s_i is a Nash equilibrium strategy for (S, u).*

The characterization result no longer holds for more than two players, since in this case a minimally rationalizable strategy need no longer be a Nash equilibrium strategy.

5.3. Forward Induction in Outside Option Games

In the class of so-called *outside option games,* the concept of minimal rationalizability singles out the unique forward induction outcome, as defined in van Damme (1989). An outside option game is a two-player game (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 games 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^*). The following theorem shows that this argument is supported by the concept of minimal rationalizability.

Theorem 5.3. *Let (S, u) be an outside option game in the sense of van Damme (1989). Then, the unique minimally rationalizable strategies for (S, u) are the forward induction strategies (In, a_1^*) and a_2^*.*

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

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

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

[8] Battigalli, P. and M. Siniscalchi (1999), Hierarchies of conditional beliefs, and interactive epistemology in dynamic games, *Journal of Economic Theory* **88,** 188-230.

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

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

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

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

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

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

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

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

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

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

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

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

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

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