Games and Economic Behavior 83 (2014) 231–254

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Belief in the opponents' future rationality ☆

Andrés Perea

Maastricht University, EpiCenter & Department of Quantitative Economics, P.O. Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 5 August 2011
Available online 27 November 2013

JEL classification:
C72

Keywords:
Epistemic game theory
Dynamic games
Belief in future rationality
Backwards rationalizability procedure
Backward dominance procedure
Backward induction

ABSTRACT

For dynamic games we consider the idea that a player, at every stage of the game, will always believe that his opponents will choose rationally in the future. This is the basis for the concept of *common belief in future rationality*, which we formalize within an epistemic model. We present an iterative procedure, *backward dominance*, that proceeds by eliminating strategies from the game, based on strict dominance arguments. We show that the backward dominance procedure selects precisely those strategies that can rationally be chosen under common belief in future rationality if we would not impose (common belief in) Bayesian updating.

1. Introduction

The goal of *epistemic game theory* is to describe plausible ways in which a player may reason about his opponents before he makes a decision in a game. In static games, the epistemic program is largely based upon the idea of *common belief in rationality* (Tan and Werlang, 1988), which states that a player believes that his opponents choose rationally, believes that every opponent believes that each of his opponents chooses rationally, and so on.

Extending this idea to dynamic games, however, does not come without problems. One obstacle is that in dynamic games it may be impossible to require that a player always believes that his opponents have chosen rationally *in the past*. Consider, for instance, a two-player game where player 1, at the beginning of the game, can choose between stopping the game and entering a subgame with player 2. If player 1 stops the game he would receive a utility of 10, whereas entering the subgame would always give him a lower utility. If player 2 observes that player 1 has entered the subgame, he cannot believe that player 1 has chosen rationally in the past.[1] In particular, it will not be possible in this game to require that player 2 always believes that player 1 chooses rationally *at all points in time*. In dynamic games, we are therefore forced to weaken the notion of *belief in the opponent's rationality*. But how?

In this paper we present one such way. We require that a player, under all circumstances, believes that his opponents will *choose rationally now and in the future*. So, even if a player observes that an opponent has chosen irrationally in the past, this should not be a reason to drop his belief in this opponent's present and future rationality. In order to keep our terminology short, we refer to this condition as *belief in the opponent's future rationality*, so we omit belief in *present* rationality in this

☆ I would like to thank Christian Bach, Pierpaolo Battigalli, Aviad Heifetz, an associate editor, and some anonymous referees for useful comments and suggestions.

E-mail address: a.perea@maastrichtuniversity.nl.
URL: http://www.personeel.unimaas.nl/a.perea/.

[1] At least, if we stick to a framework with complete information in which the players' utility functions are transparent to everyone, as we assume in this paper.

0899-8256/$ – see front matter
http://dx.doi.org/10.1016/j.geb.2013.11.008

phrase. The reader should bear in mind, however, that we always assume belief in the opponent's *present* rationality as well. A first observation is that belief in the opponents' future rationality is always possible: Even if an opponent has behaved irrationally in the past, it is always possible to believe that he will choose optimally now and at all future instances.

Belief in the opponents' future rationality is certainly not the only reasonable condition one can impose on a player's beliefs in a dynamic game, but we think it provides a natural and plausible way of reasoning about the opponents. In a sense, it assumes that the player is *completely forward looking* – he only reasons about the opponents' behavior in the future of the game, and takes the opponents' past choices for granted without drawing any conclusions from these. A possible explanation the player could give for unexpected past choices is that his opponents were making mistakes, or misjudging the situation at hand, but this should, according to the concept of belief in the opponents' future rationality, not be a reason to give up the belief that these opponents will choose rationally in the future. This condition can thus be viewed as a typical *backward induction* condition, as opposed to *forward induction* reasoning which assumes that the player, at every stage of the game, tries to interpret the opponents' past choices as being part of some rational plan. There is something to say for both lines of reasoning, but in this paper we focus on the first one.

In this paper, we do not only impose that a player always believes in his opponents' future rationality, we also require that a player always believes that every opponent always believes in his opponents' future rationality, and that he always believes that every opponent always believes that every other player always believes in his opponents' future rationality, and so on. This leads to the concept of *common belief in future rationality*, which is the central idea in this paper.

As a first step, we lay out a formal epistemic model for finite dynamic games with complete information, and formalize the notion of common belief in future rationality within this model. This enables us to define precisely which strategies can be chosen by every player under common belief in future rationality. It turns out that the strategies that can rationally be chosen under common belief in future rationality are precisely the strategies that survive Penta's (2009) *backwards rationalizability procedure* – a procedure that iteratively eliminates strategies and conditional belief vectors from the game.

We go on by presenting a *new* iterative procedure – which we call the *backward dominance procedure* – that solely eliminates strategies (not conditional belief vectors) from the game, based on strict dominance arguments. The procedure works as follows: In the first round we eliminate, at every information set, those strategies for player i that are strictly dominated at a present or future information set for player i. In every further round k we eliminate, at every information set, those strategies for player i that are strictly dominated at a present or future information set h_i for player i, given the opponents' strategies that have survived until round k at that information set h_i. We continue until we cannot eliminate anything more. The strategies that eventually survive at the beginning of the game are those that survive the backward dominance procedure.

We show that the strategies that survive the backward dominance procedure are exactly the strategies that can rationally be chosen under common belief in future rationality if we would *not impose* (*common belief in*) *Bayesian updating*. In fact, imposing (common belief in) Bayesian updating can matter for the strategies that can eventually be chosen under common belief in future rationality. Consequently, the backward dominance procedure will in general select a set of strategies that *contains* the strategies that can rationally be chosen under common belief in future rationality, but it can actually contain more. In that sense, the backward dominance procedure can be used as a first selection towards the precise set of strategies that corresponds to common belief in future rationality. This is an important result, since the backward dominance procedure is particularly easy to use.

If we apply the backwards rationalizability procedure – or even the backward dominance procedure – to games with perfect information, then we obtain precisely the well-known backward induction procedure. As a consequence, applying common belief in future rationality to games with perfect information leads to backward induction.

The idea of (common) belief in the opponents' future rationality is not entirely new. For games with perfect information, some variants of it have served as an epistemic foundation for backward induction. See, for instance, Asheim (2002), Baltag et al. (2009), Feinberg (2005) and Samet (1996). In fact, the condition of "stable belief in dynamic rationality" in Baltag et al. (2009) matches exactly our definition of belief in the opponents' future rationality, although they restrict to a non-probabilistic framework. The reader may consult Perea (2007) for a detailed overview of the various epistemic foundations that have been offered for backward induction in the literature.

For general dynamic games, belief in the opponents' future rationality is *implicitly* present in "backward induction concepts" such as sequential equilibrium (Kreps and Wilson, 1982) and sequential rationalizability (Dekel et al., 1999, 2002 and Asheim and Perea, 2005). In fact, we show in Section 7 that sequential equilibrium and sequential rationalizability are both more restrictive than common belief in future rationality.

Now, why should we be interested in common belief in future rationality as a concept if it is already implied by sequential equilibrium and sequential rationalizability? We believe there are several important reasons.

First, the concept of common belief in future rationality is based upon very elementary decision theoretic and epistemic conditions, namely that players should always believe that their opponents will choose rationally in the remainder of the game, and that there is common belief throughout the game in this event. No other conditions, besides (common belief in) Bayesian updating, are imposed. In particular, we impose no equilibrium conditions as in sequential equilibrium. So, in this sense, common belief in future rationality constitutes a very basic concept. Compared to sequential rationalizability, the concept of common belief in future rationality is very explicit about the epistemic assumptions being made. In the formulation of sequential rationalizability, the epistemic conditions imposed are somewhat more hidden in the various ingredients of its definition.