Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 159 (2015) 443–464

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

The logic of backward induction ☆

Itai Arieli [a], Robert J. Aumann [b,∗,1]

[a] *Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Israel*
[b] *Department of Mathematics and Federmann Center for the Study of Rationality, the Hebrew University of Jerusalem, Israel*

Received 24 October 2013; final version received 30 June 2015; accepted 1 July 2015

Available online 13 July 2015

Dedicated to Moshe Arieli on the occasion of his 70th birthday

Abstract

Call a perfect information (PI) game *simple* if each player moves just once. Call a player *rational* if he never takes an action while believing, with probability 1, that a different action would yield him a higher payoff. Using syntactic logic, we show that an outcome of a simple PI game is consistent with common strong belief of rationality iff it is a backward induction outcome. The result also applies to general PI games in which a player's agents act independently, rendering forward inferences invalid.

JEL classification: C72; D83

Keywords: Backward induction; Common strong belief; Perfect information; Syntactic interactive epistemology; Strong belief

☆ The authors wish to express their deep gratitude to Adam Brandenburger; in the nineties of the last century, he was instrumental in formulating the underlying concepts, and he also contributed importantly to shaping the current presentation. We also thank Joe Halpern, who referred us to the work cited in Footnote 23. Finally, we are much indebted to the editor and the referees, whose comments brought about a very significant improvement of the paper. Needless to say, none of the above are responsible for errors, misstatements, or expressions of opinion herein.

∗ Corresponding author.
E-mail addresses: iarieli@tx.technion.ac.il (I. Arieli), raumann@math.huji.ac.il (R.J. Aumann).

1 Aumann thanks the Israel Science Foundation (Grant number 1216/10) for research support.

http://dx.doi.org/10.1016/j.jet.2015.07.004

1. Introduction

The backward induction (BI) algorithm for perfect information (PI) games is based on the following reasoning: The last player, who must choose between outcomes of the game, chooses an action that maximizes his payoff; taking this as given, the previous player maximizes *his* payoff; and so on, until the beginning of the game is reached.

Though on its face convincing enough, this reasoning has for the past quarter century been discussed, scrutinized, analyzed intensely, and even rejected. Our goal here is to clarify the assumptions on the players' knowledge and rationality on which it rests.

Two preliminary observations: First, the above reasoning, and indeed the BI algorithm, is unchanged if each player i is split into several independent "agents," one for each of i's decision nodes, each with the same payoff as i. So[2] we restrict attention to *simple* PI games—those in which each player moves at just one node. Call a player *rational* if he has no action that he believes (with probability 1) yields him a higher payoff than the action he took.

Second, for the reasoning to work, more than just rationality is needed; roughly, the players must also ascribe rationality to each other. More precisely, we may take the players' rationality (r) to be *common knowledge* (ck); i.e., all players know that all are rational, all know that all know it, all know *that,* and so on. This implies that players at preterminal nodes[3] choose rationally; knowing that, players at prepreterminal[4] nodes choose rationally—i.e., the BI action; and so on. Formally, that *ckr* entails BI follows from a theorem of Aumann (1995).

Aumann's work was criticized because *ckr* involves a conceptual conundrum. A player i on the BI path must either continue on the BI path or go off. If he goes off, the player j who is reached is unreachable under *ckr*; and j knows this, as commonly known propositions are a fortiori known by all players. Since j bases his choice on what he thinks subsequent players will do, and his "knowledge" has been contradicted, it is not clear what he (j) will do. But i must base *his* choice on what *he* thinks j will do, so it is not clear what i should do, either. Specifically, it is not clear that the BI action is indeed rational for i—that he could not do better by leaving the BI path.

To avoid grappling with this conundrum,[5] we replace "knowledge" by "strong belief", where a player *strongly believes* a proposition p if he believes it unless it is logically inconsistent with his node being reached. *Common* strong belief of p signifies that p holds, all players strongly believe p, all strongly believe the foregoing, all strongly believe the foregoing, and so on. We then have our

Main Theorem. *An outcome of a simple PI game is consistent with common strong belief of rationality* (*csbr*) *iff it is a BI outcome.*

With *csbr*, the conundrum disappears. As before, suppose a player i on the BI path considers going off, say to j. When *ckr* obtains, j knows that he cannot be reached, as he knows *ckr*; so if he *is* reached, then his knowledge is contradicted, and we get the conundrum. Also when *csbr* obtains, j cannot be reached, by our theorem. But j does *not know* that he cannot be reached,

[2] Please see Section 9.1 for an explanation of this point.

[3] Those all of whose sons are outcomes.

[4] Those all of whose sons are terminal or preterminal.

[5] As have Binmore (1996) and Aumann (1996), inter alia.