Mathematical Social Sciences 75 (2015) 102–114

Contents lists available at ScienceDirect

Mathematical Social Sciences

journal homepage: www.elsevier.com/locate/econbase

Rethinking common belief, revision, and backward induction

Patricia Rich

Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA

HIGHLIGHTS

- I consider backward induction in finite, extensive form, perfect information games.
- I argue that the structure of common beliefs warrants a constraint on their revision.
- I augment the AGM theory of belief revision with this constraint, yielding "AGM+".
- AGM+ prevents un-forced revision from common belief in rationality to irrationality.
- Rationality and common belief in rationality entail backward induction.

ARTICLE INFO

Article history:
Received 27 August 2013
Received in revised form
2 March 2015
Accepted 3 March 2015
Available online 11 March 2015

ABSTRACT

Whether rationality and common belief in rationality jointly entail the backward inductive outcome in centipede games has long been debated. Stalnaker's compelling negative argument appeals to the AGM belief revision postulates to argue that off-path moves may be rational, given the revisions they may prompt. I counter that the structure of common belief and the principles of AGM justify an additional assumption about revision. I then prove that, given my proposed constraint, for all finite, n-player, extensive form, perfect information games with a unique backward inductive solution, if there is initial common belief in rationality, then backward induction is guaranteed.

1. Introduction

Backward induction as a solution concept and a pattern of reasoning for perfect information games is an old idea in game theory, and a seemingly simple one. This reasoning process has an agent who begins by considering the last possible node in a game (or in a branch of a game) and determines what the player to act would do at that node, assuming their rationality, i.e. given that the player acts so as to maximize their payoff. Taking that as given, the agent determines what the player to act at the node before the last would do; this process continues, backwards, through all nodes of the game. The backward inductive solution is the result of all players taking the actions recommended by the backward inductive reasoning process (von Neumann and Morgenstern, 1953).

Unfortunately, in many situations the backward inductive solution is counterintuitive or leads to disappointingly low payoffs all around, as will be shown later. As a result, and despite the intuitive appeal of backward induction, there is much disagreement about when it is rational to follow its recommendations and when doing so is either irrational or not the only rational option. Improved understanding of this disagreement and its possible solutions is of particular importance because of the ties between backward induction and the prisoner's dilemma.

The prisoner's dilemma is one of the better-known games studied by game theorists because of the social situations it represents. The structure of the prisoner's dilemma is that two agents have to decide independently whether to cooperate with the other or to defect. The payoffs to the agents are better if they both cooperate than if they both defect, while if one agent cooperates but the other defects, then the defector receives the highest possible payoff and the cooperator the lowest. It is therefore strictly dominant for each agent to defect (Flood, 1958). This is of special interest because the prisoner's dilemma seems to reflect the common situation faced by people of choosing whether to cooperate with others or to renege on their agreements. Political philosophers see the prisoner's dilemma as one of the fundamental problems for understanding how people could have first chosen to come together in a society and begin cooperating, and therefore of understanding what kind of social contract society might be based on or what other principles could justify the authority of governments. The solution to the prisoner's dilemma seems to suggest that rational people could never initiate cooperation in a state of nature, which would make it impossible for them to create a social contract and cooperate by

E-mail address: ptlrich@gmail.com.

http://dx.doi.org/10.1016/j.mathsocsci.2015.03.001
0165-4896/© 2015 Elsevier B.V. All rights reserved.

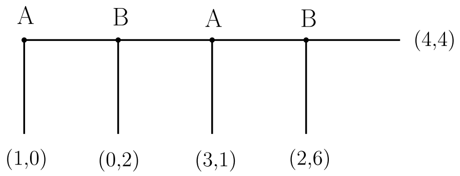

Fig. 1. A typical centipede game.

following its rules. Jean Hampton's introductory book *Political Philosophy* provides a thorough discussion of this problem (Hampton, 1997).

Backward induction enters the dialog because it blocks the simple would-be solution to the problem of initiating cooperation. It is often suggested that people could rationally cooperate in one circumstance if they expected the benefits of reciprocal cooperation in future interactions as a result; the idea is that if the prisoner's dilemma is to be encountered over and over by the same people, the long-term benefits of repeated cooperation might outweigh the short-term benefits of defecting (Kreps et al., 1982). Backward induction, however, dictates that since in the final instance of the prisoner's dilemma defecting is dominant, a rational player will not cooperate even in a repeated prisoner's dilemma if there is a known end point (Hampton, 1997).

The general structure of the above-described problem is that both players gain higher payoffs the longer they cooperate with each other, but since there is a known end to their interactions, since defecting at the last stage of interaction dominates cooperating at that stage, and given that it is always better for a player to defect during the interaction prior to that in which their opponent defects, the players end up in a kind of race to defect first. The result is that each player defects as soon as possible with low payoffs all around. This general structure is instantiated in centipede games (introduced in Rosenthal (1981)), a class of games in which the backward inductive solution typically yields the lowest possible combined payoffs to the players. Fig. 1 is a typical centipede game.

This game is characteristic of centipedes in that the payoffs to each player tend to increase during the course of the game, but the payoff to a player for moving down at any given node is always greater than the payoff if the player instead moves across, only for the next player to move down. Because of this feature, as in the finitely repeated prisoner's dilemma discussed above, the backward inductive solution to centipedes generally requires the first player to move down and end the game immediately. This is counterintuitive to those who expect both players to prefer reaching later stages of the game with higher payoffs, and disappointing to those who note that by doing what is ostensibly rational all parties are worse off than had they behaved irrationally or cooperated. There is something unsettling about a solution which results in the absolute lowest possible total utility for the players involved. This leads philosophers and social scientists to puzzle over backward induction, in hopes of better understanding its recommendations and when they apply.

Unsurprisingly, it is rare for players of the centipede game in economics experiments to follow the recommendation of backward induction. (See section 5.3 of Colin F. Camerer's *Behavioral Game Theory* for a summary of experimental results with centipede games (Camerer, 2003).) This suggests that backward induction involves some assumptions about players that fail to hold in an experimental setting; one proposed explanation is that the backward inductive solution requires, or at least follows from, common knowledge of (or belief in) rationality between the players.

For a player to act rationally in a game is for them to act so as to maximize their payoffs, given what they believe others will do. A rational player would act in this way at any decision node reached during play. Common knowledge of rationality is rationality of all players, mutual knowledge of the rationality of all players (i.e. first level mutual knowledge), mutual knowledge of first level mutual knowledge (i.e. second level mutual knowledge), and indeed infinite levels of mutual knowledge of rationality; common belief, of interest here, is defined analogously. Naturally this is not a condition that one would expect to hold in an anonymous experiment, but the focus of this paper will be the claim that *when* common belief in rationality is present between players in a perfect information game, *then* the backward inductive solution results.

This claim has been put forth by many people over the years in various incarnations, and much energy has been spent denying it as well. For a detailed picture of the history of the debate and an analysis of some of the attempts to justify or refute the claim in question, see Graciela Kuechle's *What Happened to the Three-Legged Centipede?* (Kuechle, 2009).

Despite the rarity of common belief in rationality among individuals, especially in anonymous experimental settings, it is an important question whether it is a sufficient condition for backward induction. For one, although it may not be the norm, it surely obtains in many situations of interest. It should not be surprising that a group of strangers would be uncertain of each other's rationality, but many games are played by friends, colleagues, and classmates who have had ample time to gather evidence of one another's rationality. On a theoretical level, if it is important to understand backward induction itself (as is argued above), then it is also important to pinpoint what assumptions it makes about game players, not only whether rationality is sufficient for backward inductive play but whether (if rationality is insufficient) then common belief in rationality is sufficient instead. It is not possible to fully understand backward induction without knowing what assumptions it makes or what conditions it depends on, and as this paper endeavors to show, common belief in rationality may be the right condition.

Robert Stalnaker provides a compelling argument that purports to demonstrate that common belief in rationality is not sufficient for backward induction without additional, unjustified assumptions about players' belief revision processes (Stalnaker, 1998). I argue that one particular assumption about how players revise their beliefs when there is common belief in rationality is in fact justified, due not to some special feature of rationality but due to the structure of common beliefs and the AGM axioms themselves. This assumption is that when there is common belief in a proposition, or levels of mutual belief, then upon learning that a particular level of mutual belief in the proposition cannot obtain, rational agents revise their beliefs so as to retain the highest level of mutual belief consistent with what was learned. I use amodeling due to Grove (1988) in advocating for this additional assumption. I then prove that if it is accepted as a requirement of rational belief revision (and therefore that it is also common belief among the players), then common belief in rationality is sufficient for backward induction.

2. Stalnaker's belief-based framework

In a typical game situation where players are reasoning about what their opponents are thinking and planning to do, it is clear that players may think that they know something when in fact they only believe it, because it is false. Furthermore, players may come to discover that some of their beliefs, particularly those about other players, are indeed false when an action is taken that is inconsistent with the beliefs initially held. A player's rational actions, then, depend on which new beliefs replace the original ones, or how players' beliefs are revised. Accordingly, Stalnaker argues that this revision process must be explicitly included in our models of games, so that unjustified assumptions about belief revision are not hidden within the framework.

Stalnaker's framework combines a representation of players' beliefs with their policies for how they would revise those beliefs if they were to learn some surprising information during the