2010

TR-2010007: Robust Knowledge of Rationality

Sergei Artemov

Robust Knowledge of Rationality

Sergei Artemov*
The CUNY Graduate Center
365 Fifth Avenue, 4319
New York City, NY 10016, USA
sartemov@gc.cuny.edu

July 26, 2010

Abstract

Stalnaker provided an example of a perfect information game in which common knowledge of rationality does not yield backward induction. However, in his example, knowledge is treated as defeasible: players forfeit their knowledge of rationality at some vertices. This is not how 'knowledge' is understood in epistemology where, unlike belief, it is not subject to revision. In this respect, the Stalnaker example is a fit for 'rationality and common *belief* of rationality' rather than 'common knowledge of rationality.' In order to represent *knowledge* in the belief revision setting we introduce the notion of 'robust knowledge' which is maintained whenever possible during belief revision. We show that robust knowledge of Stalnaker rationality in games of perfect information yields backward induction.

1 Introduction

Stalnaker's approach to games of perfect information (PI games) introduces belief revision into players' reasoning ([8]). The paradigmatic example is given by the common interest game in Figure 1. In Aumann's setting ([1]), given common knowledge of rationality, players play the backward induction solution (*aaa*), i.e., *across* at all three nodes. Stalnaker's approach claims that the solution (*dda*), i.e., *down* at v_1, *down* at v_2, and *across* at v_3 can be regarded as rational under 'the same' assumption of common knowledge of rationality.

Stalnaker's reasoning proceeds as follows. Consider the variant of the game in which (*dda*) is common knowledge. Then it is common knowledge that both players are rational, but the only solution, (*dda*), is not the backward induction solution.

*This work is supported by the National Science Foundation under Grant No. 0830450.

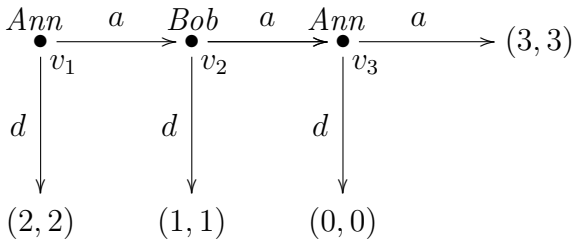

Figure 1: Stalnaker's game

It suffices to check that both players are rational in (*dda*); this would yield the common knowledge of rationality since (*dda*) is common knowledge.

- Ann is rational at v_3 according to the game tree.
- Bob is rational at v_2 since if Ann were to play across (an obviously irrational move by Ann given her knowledge that Bob is playing *down*), then Bob revises his initial belief of Ann's rationality and no longer assumes that Ann will play *across* at v_3. Under these circumstances, playing *down* at v_2 is not irrational for Bob.
- Ann is rational at v_1 since she knows that Bob is playing *down*.

In this proof, the heart of the matter is how Bob would react to being surprised by Ann's (irrational) move *across* at v_1. There are various possibilities:

1. Bob revises some of his beliefs, including his belief in Ann's rationality for the remainder of the game;
2. Bob revises some of his beliefs, but not his belief in Ann's rationality for the remainder of the game.

Stalnaker describes what happens when the first possibility is allowed, which makes good sense. This case was cast in a formal logical framework in [4].

We offer a general logical treatment of the second case. In the context of the 'knowledge of rationality,' it leads to the backward induction solution, *BI*, in all PI games.[1] Our goal is not to defend or attack *BI*, but to formulate the underlying issues fully and formally.

How is our approach different from Aumann's? Aumann obtains *BI*, but not via this route. In his treatment, there is no explicit belief revision. In contrast, we allow belief revision, but knowledge of (Stalnaker) rationality for the remainder of the game is maintained.

[1] Stalnaker in [8] also indicates that in case 2, the game in Figure 1 will end in the backward induction solution, *BI*. While discussing general notions of "robust belief in rationality" and "rationalization principle" which correspond to case 2, Stalknaker points out that they lead to a potentially infinite tower of belief revision priorities which quickly loses intuitive plausibility.

There is a principal point at which our approach differs from that of Stalnaker. Stalnaker's basic epistemic assumptions concern belief rather than knowledge: while assuming 'knowledge of rationality,' Stalnaker's players treat knowledge as defeasible whereas since Descartes, epistemology has usually attributed to knowledge a certain degree of non-defeasibility (infallibility, reliability, truth-tracking, necessity, etc., cf. [5, 6, 7, 9]). What is known is true in a robust way and is not subject to revision. Defeasibility, openness to revision, is a property of belief (hence 'belief revision') rather than of knowledge. The standard game-theoretical assumption

$$\textit{common knowledge of players' rationality} \tag{1}$$

does not suggest the possibility of revising the rules of the game, payoffs, or rationality assumptions. Belief revision's approach to PI games is a fit for the assumption

$$\textit{players' rationality and common belief of players' rationality,} \tag{2}$$

which is itself a fascinating subject that is not, however, identical to studying games with assumption (1). Note that (2) has been formalized in the belief-based literature in various ways (cf. [2, 3]) and these also allow solution (*dda*) for the game in Figure 1.

2 Models of rationality and belief revision

Let us recap basic terminology ([1, 4]). An extensive game consists of the following components.

1. A finite set $N = \{1, 2, \ldots, n\}$ of players.

2. A finite rooted tree H. Each node has a unique path from the root called the history of this node. The leaves of the game tree are called terminal nodes, or outcomes. The set of all terminal nodes is called Z.

3. A player function P that assigns a player (who makes a move) to each nonterminal node.

4. For each player, a payoff function defined on Z.

The root node is the starting point of the game. At any node $v \in (N \setminus Z)$, player $P(v)$ chooses one of the successor nodes (move).

An **Aumann model** is a tuple $\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$, where Ω is a set of "epistemic states" of the world, $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are knowledge partitions of Ω corresponding to players $1, 2, \ldots, n$, and $\mathbf{s}$ is a mapping from Ω to the set of all strategy profiles: for a state ω,

$$\mathbf{s}(\omega) = (s_1, \ldots, s_n).$$

We write $\mathbf{s}_i(\omega)$ for i's component of the strategy profile $\mathbf{s}(\omega)$, i.e., s_i. Also, let (s_{-i}, s^i) be the strategy profile obtained from s by replacing s_i by s^i, $h_i^v(s)$ be i's conditional payoff if strategy profile s is followed starting at v, and $\mathcal{K}_i(\omega)$ be the cell in $\mathcal{K}_i$ that includes ω.

The definition of rationality is formalized as follows.

Definition 1 *Player i is* **rational** *at vertex v in state ω if, for all strategies s^i,*

$$h_i^v(\mathbf{s}(\omega')) \geq h_i^v(\mathbf{s}_{-i}(\omega'), s^i)$$

for some $\omega' \in \mathcal{K}_i(\omega)$. Player i is rational in state ω if i is rational at any node in ω.

Extended models formalize Stalnaker's representation of counterfactuals via the selection function "the closest world where a given vertex is reached." In a formal setting, the extended model is a tuple

$$\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$$

where $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$ is an Aumann model and a selection function f maps pairs of states and vertices to states. The intended reading of $f(\omega, v) = \omega'$ is

$$\omega' \textit{ is the closest state to } \omega \textit{ in which vertex } v \textit{ is reached.}$$

It is assumed that f satisfies the following conditions:

F1. Vertex v is reached in $f(\omega, v)$.

F2. If v is reached in ω, then $f(\omega, v) = \omega$.

F3. $\mathbf{s}(f(\omega, v))$ and $\mathbf{s}(\omega)$ agree on the subtree of the game tree at and below v.

Definition 2 ([4]) *Player i is* **Stalnaker-rational** *in state ω at vertex v if i is rational at vertex v in $f(\omega, v)$. Player i is Stalnaker-rational in state ω if i is Stalnaker-rational at any of its vertices in ω.*

3 Common knowledge is too weak for belief revision

The principal reservation[2] concerning extended models is that common knowledge of Stalnaker rationality is defeasible. It can be immediately observed that Stalnaker rationality spills over epistemic reachability – state $f(\omega, v)$ can be unreachable from ω – which is an indication that reachability-based common knowledge may be not adequate.

Consider, for example, the game in Figure 1. Following [4], we introduce[3] the following strategy profiles:

- s^1 is the strategy profile (dda), i.e., Ann plays *down* at v_1, Bob plays *down* at v_2, and Ann plays *across* at v_3;

[2]There is also the technical issue that the principle 'players are aware of their own rationality,' which is usually adopted as a property of rationality, can be violated in extended models.

[3]in slightly different notation

- s^2 is the strategy profile (ada);
- s^3 is the strategy profile (add);
- s^4 is the strategy profile (aaa) (which is the backward induction solution);
- s^5 is the strategy profile (aad).

As in [4], consider extended model $\mathcal{A} = (\Omega, \mathcal{K}_{Ann}, \mathcal{K}_{Bob}, \mathbf{s}, f)$ where

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$;
- $\mathcal{K}_{Ann} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;
- $\mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;
- $\mathbf{s}(\omega_j) = s^j$ for j =1–5;
- $f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = \omega_4$, $f(\omega_2, v_3) = \omega_4$, $f(\omega_3, v_3) = \omega_5$, and $f(\omega, v) = \omega$ in all other situations.

The real epistemic state is assumed to be ω_1. The Stalnaker-Halpern argument claims that

$$\textit{Stalnaker rationality is common knowledge in } \omega_1. \tag{3}$$

Since ω_1 is not a backward induction solution, (3) implies that in model $\mathcal{A}$, common knowledge of rationality does not yield backward induction. Let us prove (3). Since

$$\mathcal{K}_{Ann}(\omega_1) = \mathcal{K}_{Bob}(\omega_1) = \{\omega_1\},$$

everything that is true in ω_1 is common knowledge in ω_1. Let us check that Stalnaker rationality of both players holds in ω_1, in particular that Bob is Stalnaker-rational in ω_1 at v_2. Selection function f reduces this question to the claim that Bob is (Aumann-) rational in epistemic state ω_2 at vertex v_2 which is established by direct application of Definition 1.

The problem is that in state ω_2 at vertex v_2 **Bob cannot know that Ann is Stalnaker-rational**. Indeed, Ann is not Stalnaker-rational in ω_3 (since $f(\omega_3, v_3) = \omega_5$ and Ann in not rational in ω_5 at v_3), and $\omega_3 \in \mathcal{K}_{Bob}(\omega_2)$. Speaking informally, following selection function $f(\omega_1, v_2) = \omega_2$, Bob in ω_1 revises his belief that Ann plays *down* at v_1 and considers the case ω_2 in which Ann plays *across* at v_1. Accidentally, Bob also forfeits his knowledge of Ann's rationality at v_3, thus treating this knowledge as a mere belief.

The interpretation of the aforementioned Stalnaker-Halpern result as an example of a PI game with condition (1) of common knowledge of rationality is not entirely convincing: rather, this game corresponds to condition (2) of rationality and common *belief* of rationality.

4 Robust knowledge of Stalnaker rationality

Common knowledge of rationality (in a given state ω) requires that rationality holds in all reachable states, which turned out to be too weak to represent some nuances of belief revision. In this section, we introduce a notion of *robust knowledge of Stalnaker rationality* in which Stalnaker rationality holds in all relevant situations. This notion captures the essence of belief revision under which knowledge of rationality at a vertex is maintained whenever possible. This can also be considered as a case study which sketches a general framework for different sorts of rationality: (Aumann) rationality is required for some sets X of situations, i.e., pairs *(state,vertex)*, and the choice of X is used to specify the corresponding notion of rationality. In particular,

1. knowledge of rationality in state ω: $X = \{(\omega, v) \mid v \textit{ is a vertex}\}$;
2. common knowledge of rationality in state ω:
   $X = \{(\omega', v) \mid \omega' \textit{ is reachable from } \omega, v \textit{ is a vertex}\}$;
3. Stalnaker rationality in state ω: $X = \{(f(\omega, v), v) \mid v \textit{ is a vertex}\}$.

Given an extended model $\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$, a **situation** is a pair (ω, v) of a state ω and vertex v of the game tree. We define a notion of **relevant situation** which reflects our goal to maintain common knowledge of Stalnaker rationality for the remainder of the game at any depth of the belief revision process. The set of situations relevant in (ω, v) is closed under belief revision, epistemic reachability, and advancing to a later moment in the game. **Robust knowledge of Stalnaker rationality** (Definition 3) is defined then via rationality in any relevant situation.

A situation (ω', v') is *relevant in* (ω, v), if there is a finite sequence of situations

$$(\omega, v) = (\omega_0, v_0),\ (\omega_1, v_1),\ (\omega_2, v_2),\ \ldots, (\omega_m, v_m) = (\omega', v')$$

such that for each $k = 0, \ldots, m-1$,

1. $v_k \preceq v_{k+1}$, i.e., v_{k+1} is a future node with respect to v_k;
2. $\omega_{k+1} = f(\widetilde{\omega_k}, v_{k+1})$ for some $\widetilde{\omega_k}$ reachable from ω_k.

It is easy to see that to get from (ω_k, v_k) to (ω_{k+1}, v_{k+1}), one has to pick a state $\widetilde{\omega_k}$ reachable from ω_k (e.g., $\widetilde{\omega_k} = \omega_k$) and a future vertex v_{k+1} (e.g., $v_{k+1} = v_k$), and advance to the revised state $f(\widetilde{\omega_k}, v_k)$. Iteration of this procedure generates all relevant situations.

Example 1 In model $\mathcal{A}$, the set U of situations relevant in (ω_1, v_3) is $U = \{(\omega_4, v_3)\}$. The set V of situations relevant in (ω_1, v_2) is $V = U \cup \{(\omega_2, v_2),\ (\omega_3, v_2),\ (\omega_5, v_3)\}$. The set W of situations relevant in (ω_1, v_1) is $W = V \cup \{(\omega_1, v_1)\}$. Intuitively, Stalnaker rationality in state ω_1 is determined by (Aumann) rationality in five situations from W.

Definition 3 *Robust knowledge of Stalnaker rationality in state ω at vertex v means that in any situation (ω', v') relevant in (ω, v), player $P(v')$ is rational in ω' at v'. Robust knowledge of Stalnaker rationality in state ω means robust knowledge of Stalnaker rationality in state ω at v for each vertex v.*

This definition justifies the notion of a 'relevant situation': robust knowledge of Stalnaker rationality guarantees common knowledge of 'Stalnaker rationality is maintained for the remainder of the game' in any relevant situation.

Example 2 In model $\mathcal{A}$, Stalnaker rationality is common knowledge in ω_1. However, robust knowledge of Stalnaker rationality does not hold in (ω_1, v_1). Indeed, situation (ω_5, v_3) is relevant in (ω_1, v_1), but Ann is not rational in ω_5 at v_3.

Example 3 Consider the length-three Centipede game in Figure 2

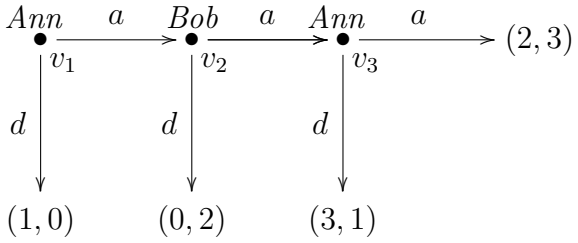

Figure 2: Length-three Centipede game

and an extended model $\mathcal{A}'' = (\Omega, \mathcal{K}_{Ann}, \mathcal{K}_{Bob}, \mathbf{s}, f)$ where

- $\Omega = \{\omega_1, \omega_2, \omega_3\}$;
- $\mathcal{K}_{Ann} = \mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$;
- $\mathbf{s}(\omega_1) = (ddd)$, $\mathbf{s}(\omega_2) = (add)$, $\mathbf{s}(\omega_3) = (aad)$;
- $f(\omega_1, v_2) = \omega_2$, $f(\omega_1, v_3) = f(\omega_2, v_3) = \omega_3$.

Stalnaker rationality does not hold in some situations, e.g., (ω_2, v_1) and (ω_3, v_2). However, such 'bad' situations are irrelevant in 'real' state ω_1 and robust Stalnaker rationality holds in ω_1. Indeed, relevant situations in ω_1 for all possible v's are

$$\{(\omega_1, v_1),\ (\omega_2, v_2),\ (\omega_3, v_3)\},$$

and the corresponding players are rational in all of them.

Example 3 shows how to model belief revision while maintaining knowledge of Stalnaker rationality for the remainder of the game in a meaningful way. Counterfactual strategy profiles are represented in the model (states ω_2 and ω_3), but they don't spoil rationality at relevant vertices, which is exactly what we wanted.

The following theorem states that robust knowledge of Stalnaker rationality yields backward induction in all PI games.

Theorem 1 *In extended models over generic game trees, robust knowledge of Stalnaker rationality yields backward induction.*

Proof. Let

$$\mathcal{M} = (\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$$

be an extended model such that robust knowledge of Stalnaker rationality holds in state ω of $\mathcal{M}$. This yields that a corresponding player is rational in ω' at v' for each situation (ω', v') relevant in (ω, v_0) where v_0 is the root vertex. We claim that for every relevant situation (ω', v'), restriction of profile $\mathbf{s}(\omega')$ on the subtree Γ below v' coincides with *BI*. Theorem 1 follows from this claim since (ω, v_0) is relevant in itself and the subtree Γ below v_0 is the entire game tree.

To prove the claim, assume the opposite, i.e., that $\mathbf{s}(\omega') \neq BI$ on the subtree Γ below v' for some relevant situation (ω', v'). Let (ω', v') be such a situation with the lowest non-terminal vertex v'. Let also i be the player making a choice at v'.

Note that $\mathbf{s}(\omega')$ coincides with *BI* at any vertex v'' strictly below v'. Indeed, situation $(f(\omega', v''), v'')$ for any vertex v'' strictly below v' is relevant, by the definition. By choice of (ω', v'), $\mathbf{s}(f(\omega', v''))$ coincides with *BI* on v''. By condition F3 on the selection function, $\mathbf{s}(f(\omega', v''))$ agrees with $\mathbf{s}(\omega')$ on v'', hence $\mathbf{s}(\omega')$ coincides with *BI* on v''.

Then i is not Aumann-rational at v' in ω'. Indeed, the backward induction at v' chooses the best move for i given *BI*-moves at all other nodes of the subtree Γ below v'. Since the choice of $\mathbf{s}(\omega')$ at v' is different from those of *BI* and the game tree is generic, it can only be strictly worse. By Definition 1, i is not rational in ω' at v'. □

5 Discussion

Extended models treat knowledge as defeasible: players revise not only their beliefs in other players' moves but also their 'knowledge' of rationality for the remainder of the game. However, in epistemology, 'knowledge' is usually understood as non-defeasible, and not subject to revision. In this respect, the Stalnaker example reflects the assumption 'rationality and common belief of rationality' rather than 'common knowledge of rationality.'

The notion of robust knowledge of Stalnaker rationality reflects the idea of common knowledge of Stalnaker rationality for the remainder of the game at any depth of the belief revision process; it necessarily goes beyond reachability-based common knowledge.

For games with a 'small' number of irrational moves, robust knowledge of Stalnaker rationality can be justified by strong *a priori* rationality reputation of players, their history of rational behavior, etc. An isolated irrational move can be viewed as a technical error. However, trust in rationality fades with each irrational move and given a 'large' number of such moves, robust knowledge of Stalnaker rationality becomes unfeasible. More realistic models of robust rationality should include a bound on the number of errors (e.g., one) allowed for each player.

6 Acknowledgments

The author is grateful to Adam Brandenburger for inspiring discussions and valuable advice. The author is indebted to Vladimir Krupski, Elena Nogina, and Cagil Tasdemir for many useful suggestions. Special thanks to Karen Kletter for editing this paper.

References

[1] R. Aumann. Backward Induction and Common Knowledge of Rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[2] P. Battigalli, A. Friedenberg. Context-Dependent Forward Induction Reasoning. *Working Paper n. 351*, IGIER Università Bocconi, Milano Italy August 2009.

[3] A. Brandenburger, A. Friedenberg. Self-admissible sets. *Journal of Economic Theory*, 145(2):785-811, 2010.

[4] J. Halpern. Substantive Rationality and Backward Induction. *Games and Economic Behavior*, 37:425–435, 2001.

[5] L. Newman. Descartes' Epistemology. *Stanford Encyclopedia of Philosophy*, 2005.

[6] L. Newman, A. Nelson. Circumventing Cartesian Circles. *Noûs*, 33:370–404, 1999.

[7] R. Nozick. Philosophical Explanations. *Philosophical Explanations*. Cambridge: Harvard University Press, 1981.

[8] R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.

[9] M. Steup. Epistemology. *Stanford Encyclopedia of Philosophy*, 2005.