Robust Knowledge and Rationality

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

November 22, 2010

Abstract

In 1995, Aumann proved that in games of perfect information, common knowledge of rationality yields backward induction. In 1998, Stalnaker provided an example of a game in which common knowledge of rationality, once belief revision is taken into account, does not yield backward induction. However, in some pertinent situations in this example, players are allowed to forfeit the rationality condition. We introduce the notion of *robust knowledge* which extends common knowledge to all relevant situations, including counterfactual ones. *Robust knowledge of rationality*, in a general belief revision setting, represents the "no irrationality in the system" condition which is at the heart of the backward induction argument. We show that in games of perfect information, robust knowledge of rationality yields backward induction. This may be regarded as a natural form of Aumann's theorem which accommodates belief revision.

1 Introduction

Stalnaker's approach to games of perfect information (PI games) introduces belief revision into players' reasoning [Stalnaker, 1998]. The paradigmatic example is provided by the common interest game in Figure 1. In Aumann's setting [Aumann, 1995], 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, once belief revision is taken into account.

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

Stalnaker's reasoning proceeds as follows. Consider a 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.

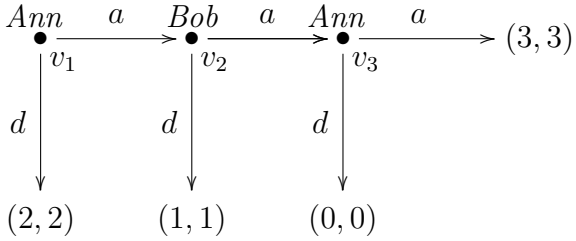

Figure 1: Stalnaker's game

It suffices to check that both players are rational in (dda); this would yield 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 at v_1 (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 his belief in Ann's rationality for the remainder of the game;
2. Bob does not revise his belief in Ann's rationality for the remainder of the game.

Stalnaker describes what happens when the first possibility is allowed. This case was cast in a formal logical framework in [Halpern, 2001].

We offer a general logical treatment of the second case and show that 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 does our approach correspond to Aumann's? Aumann obtains *BI*, but not via this route. In his treatment, there is no explicit belief revision and the condition "there is no irrationality in the system" is represented by the common knowledge of rationality

[1] [Stalnaker, 1998] claims that, in case 2, the game in Figure 1 will end in the backward induction solution, but he does not offer a formal argument for this example or more generally.

assumption. In contrast, we allow belief revision, but preserve knowledge of rationality for the remainder of the game; the condition "there is no irrationality in the system" is represented by a stronger assumption which we call *robust knowledge of rationality*. So our approach may be regarded as an extension of Aumann's to a belief revision setting. In particular, when the belief revision is trivial, i.e., epistemic states do not change, our approach coincides with Aumann's.

2 Models of rationality and belief revision

Let us recap basic terminology. An extensive game consists of the following components.

1. A finite set $I = \{1, 2, \ldots, n\}$ of players.
2. A finite rooted tree H with set of nodes (also called 'vertices') N. 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 (making a move) to each nonterminal node.
4. For each player, a payoff function defined on Z.

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

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

Each s_i is a strategy of player i, i.e., an assignment of a move at each node v such that $P(v) = i$. We write $\mathbf{s}_i(\omega)$ for i's component of the strategy profile $\mathbf{s}(\omega)$, i.e., s_i. Also, for a strategy profile s, 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 following *measurability* property is usually assumed: players know their own strategies, i.e., $\mathbf{s}(\omega) = \mathbf{s}(\omega')$ whenever $\omega' \in \mathcal{K}_i(\omega)$.

The definition of rationality is formalized as follows.

Definition 1 *Player i is* **rational** *at vertex v in state ω if, for each strategy s^i, there exists $\omega' \in \mathcal{K}_i(\omega)$ such that*

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

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 *ω' is the closest state to ω in which vertex v 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 *Player i is* **Stalnaker-rational** *in state ω at vertex v if i is rational at v in $f(\omega, v)$.*

Substantive rationality is rationality at all vertices of the game tree. This definition also extends to Stalnaker rationality.

In epistemology, 'knowledge' is usually understood as non-defeasible (cf. [Steup, 2005]), and not subject to revision. Stalnaker allows revision of 'common knowledge of rationality' in some hypothetical situations and hence treats 'common knowledge of rationality,' rather, as 'rationality and common belief of rationality.' The latter has been formalized in the belief-based literature in various ways (cf. [Battigalli and Friedenberg, 2009, Brandenburger and Friedenberg, 2010]); these also allow solution (*dda*) for the game in Figure 1 under assumptions of rationality and common belief in rationality.

3 Common knowledge is too weak for belief revision

We argue that the common knowledge of rationality assumption is too weak to guarantee that there is no irrationality in the system. Consider the game in Figure 1. Following [Halpern, 2001], we introduce[2] 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;
- 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 [Halpern, 2001], consider the extended model $\mathcal{A} = (\Omega, \mathcal{K}_{Ann}, \mathcal{K}_{Bob}, \mathbf{s}, f)$ where

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$;
- $\mathcal{K}_{Ann} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;
- $\mathcal{K}_{Bob} = \{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4\}, \{\omega_5\}\}$;

[2] in slightly different notation

- $\mathbf{s}(\omega_j) = s^j$ 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 true epistemic state is assumed to be ω_1. The Stalnaker-Halpern argument claims that

$$\textit{Stalnaker rationality is common knowledge in } \omega_1. \tag{1}$$

Since ω_1 is not a backward induction solution, (1) implies that in model $\mathcal{A}$, common knowledge of rationality does not yield backward induction. Let us prove (1). 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 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 will stay Stalnaker-rational**. Indeed, Ann is not Stalnaker-rational in ω_3 (since $f(\omega_3, v_3) = \omega_5$ and Ann is 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 at v_2 revises his belief from ω_1 to ω_2 in which Ann plays *across* at v_1. Accidentally, Bob also forfeits his knowledge of Ann's rationality at v_3. Technically speaking, this is not a violation of the common knowledge of rationality assumption since ω_2 is not epistemically reachable from the original state ω_1; this is the essence of the Stalnaker example.

As we can see, the common knowledge of substantive Stalnaker rationality assumption does not reach its intended goal to secure, in Aumann's words ([Aumann, 2010]), that "there is no irrationality in the system," and hence stronger notions of shared rationality should be considered.

4 Robust knowledge

Robust knowledge is common knowledge that is maintained during belief revision.

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 nonterminal vertex v of the game tree. We define a notion of *relevant situation* that reflects our goal to maintain common knowledge for the remainder of the game at any depth of the belief revision process.

The set RS of situations **relevant** to a situation (ω, v) is the minimal set of situations containing (ω, v) and closed under

a) reachability of states: $(\omega', v') \in RS$ *and* ω'' *is reachable from* ω' *yields* $(\omega'', v') \in RS$ and

b) belief revision: $(\omega', v') \in RS$ *and* $v' \preceq v''$[3] *yields* $(f(\omega', v''), v'') \in RS$.

Definition 3 *Event F is* **robust knowledge** *in situation (ω, v) if F holds in all situations relevant to (ω, v).*

We can say that F is robust knowledge in state ω if F is robust knowledge in situations (ω, v) for all nonterminal vertices v. However, in the belief revision setting, it seems more appropriate to consider knowledge/belief **in situations** rather than states since moving from a node v to a later node v' can be accompanied by revision of epistemic states.

Definition 4 *Event F is* **universal knowledge** *if F holds in all situations.*

From the definitions, if follows that in any situation,

Universal Knowledge $\Rightarrow$ *Robust Knowledge* $\Rightarrow$ *Common Knowledge* $\Rightarrow$ *Knowledge.*

If belief revision in a model is trivial, i.e., does not change epistemic states, $f(\omega, v) = \omega$ for each state ω and vertex v, then robust knowledge in the initial situation (ω, r), where r is the root node of the game tree, is equivalent to common knowledge in state ω[4]

5 Robust knowledge of rationality

Robust knowledge of rationality means that rationality holds in all relevant situations. This notion represents the belief revision strategy under which knowledge of rationality for the remainder of the game is maintained.

Definition 5 *Robust knowledge of rationality in a given situation means that, in each relevant situation (ω, v), player $P(v)$ is rational at vertex v in state ω.*

Robust knowledge of rationality represents the condition that neither iterated reasoning nor belief revision at any node can reach an 'irrational' situation.

How is this related to other methods of expressing "no irrationality in the system" in a belief revision setting? It is immediately apparent that the common knowledge of substantive Stalnaker rationality is subsumed by robust knowledge of rationality. Indeed, suppose common knowledge of substantive Stalnaker rationality does not hold in a true situation (ω, v). Then there is a state ω' reachable from ω and node $v' \succeq v$ such that rationality of the corresponding player fails in situation $(f(\omega', v'), v')$. The latter is relevant in (ω, v), hence robust knowledge of rationality fails in (ω, v). The fact that robust knowledge of rationality is strictly stronger than common knowledge of substantive Stalnaker rationality is demonstrated in Example 1.

[3] i.e., v'' is v' or a future node with respect to v'.

[4] [Aumann, 1987] notes that in strategic games, one can restrict the universe to the smallest common knowledge event, hence universal, robust, and common knowledge coincide.

Example 1 In model $\mathcal{A}$, there are five states and three non-terminal nodes, hence $5 \times 3 = 15$ situations total. Assuming common knowledge of substantive Stalnaker rationality in true state ω_1 means rationality in the following three situations:

$$S = \{(\omega_1, v_1),\ (\omega_2, v_2),\ (\omega_4, v_3)\}.$$

However, as we have seen earlier, this assumption is not sufficient to ensure that there is no irrationality in the system. In particular, in (ω_2, v_2), Bob considers ω_3 possible, and to guarantee that Bob in (ω_2, v_2) knows Ann's rationality we have to add the condition that Bob knows Ann's rationality in (ω_3, v_2) as well. Further closure with respect to belief revision produces one more relevant situation (ω_5, v_3) in which rationality should be maintained to guarantee that "there is no irrationality in the system."

We end up with the set W of all realizable situations relevant to true situation (ω_1, v_1):

$$W = \{(\omega_1, v_1),\ (\omega_2, v_2),\ (\omega_4, v_3),\ (\omega_3, v_2),\ (\omega_5, v_3)\}.$$

Robust knowledge of 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.

This example illustrates the difference between Halpern-Stalnaker's common knowledge of substantive Stalnaker rationality and robust knowledge of rationality: the former is not necessarily closed under reachability in relevant situations which allows for irrationality in the system.

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

Theorem 1 *In extended models over generic game trees, robust knowledge of 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 rationality holds in true situation (ω_0, r) where r is the root vertex. This means that a corresponding player is rational in each relevant situation (ω, v). We claim that for every relevant situation (ω, v), restriction of profile $\mathbf{s}(\omega)$ on the subtree Γ below v (which includes v itself) coincides with *BI*. Theorem 1 follows from this claim since (ω_0, r) is relevant in itself and the subtree Γ below r 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 situations (ω, v). Let $(\widetilde{\omega}, \widetilde{v})$ be such a situation with the lowest non-terminal vertex $\widetilde{v}$. Also, let i be the player making a move at $\widetilde{v}$.

Lemma 1 *Let $\omega' \in \mathcal{K}_i(\widetilde{\omega})$. Then $\mathbf{s}(\omega')$ coincides with BI at any vertex v' strictly below $\widetilde{v}$.*

Proof. Let v' be a vertex strictly below $\widetilde{v}$. Situation $(f(\omega', v'), v')$ is relevant, by the definition. By the choice of $\widetilde{v}$, $\mathbf{s}(f(\omega', v'))$ 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'. □

In particular, $\mathbf{s}(\widetilde{\omega})$ coincides with BI at any vertex v' strictly below $\widetilde{v}$. By the choice of $(\widetilde{\omega}, \widetilde{v})$, strategy profile $\mathbf{s}(\widetilde{\omega})$ suggests a non-BI move at or below $\widetilde{v}$. By Lemma 1, such a move can only occur at $\widetilde{v}$. So, $\mathbf{s}(\widetilde{\omega})$ is different from BI at $\widetilde{v}$.

We now show that i is not rational in $(\widetilde{\omega}, \widetilde{v})$. Let $\omega' \in \mathcal{K}_i(\widetilde{\omega})$, then situation $(\omega', \widetilde{v})$ is relevant. By measurability, strategy profiles $\mathbf{s}(\omega')$ and $\mathbf{s}(\widetilde{\omega})$ coincide at $\widetilde{v}$, in particular, $\mathbf{s}(\omega')$ differs from BI at $\widetilde{v}$.

Therefore, at and below $\widetilde{v}$, $BI = (\mathbf{s}_{-i}(\omega'), BI_i)$. Since backward induction at $\widetilde{v}$ chooses the best move for i given BI-moves below $\widetilde{v}$, and the game tree is generic, the payoff of $\mathbf{s}(\omega')$ at $\widetilde{v}$ is strictly worse than that of BI:

$$h_i^{\widetilde{v}}(\mathbf{s}(\omega')) < h_i^{\widetilde{v}}(\mathbf{s}_{-i}(\omega'), BI_i).$$

By Definition 1, i is not rational in $\widetilde{\omega}$ at $\widetilde{v}$, which contradicts the robust knowledge of rationality assumption. □

In addition, we can make an immediate existence observation: the epistemic conditions of Theorem 1 are possible for any game tree.

Theorem 2 *For an arbitrary game tree, there exists an extended Aumann model such that robust knowledge of rationality holds in a true situation.*[5]

Proof. Given a game tree $\mathcal{T}$ with set of nodes N and set of terminal nodes Z, define an extended Aumann model $\mathcal{M}$ such that

$\Omega = \{\omega_v \mid v \in N \setminus Z\}$, i.e., one state for each nonterminal node;

$\mathcal{K}_1 = \mathcal{K}_2 = \ldots = \mathcal{K}_n = \{\{\omega_v\} \mid v \in N \setminus Z\}$;

$\mathbf{s}(\omega_v)$ is a (unique) profile the leads from root r to v and coincides with the backward induction profile BI at all other nodes;

$$f(\omega_u, v) = \begin{cases} \omega_u & \text{if } v \preceq u; \\ \omega_v & \text{otherwise.} \end{cases} \tag{2}$$

It is easy to observe that properties F1–F3 of the selection function are valid.

In model $\mathcal{M}$, the set RS of situations relevant to true situation (ω_r, r) is

$$\{(\omega_v, v) \mid v \in N \setminus Z\}.$$

Indeed, since reachability in $\mathcal{M}$ is trivial, RS is generated by belief revision only: $(\omega', v') \in RS$ *and* $v' \preceq v''$ *yields* $(f(\omega', v''), v'') \in RS$. Belief revision always applies selection function f to a future node, hence only the second clause of the definition of f (2) is applicable: since $u \prec v$, $(f(\omega_u, v), v) = (\omega_v, v)$.

In each situation (ω_v, v), player $P(v)$ is rational since profile $\mathbf{s}(\omega_v)$ coincides with BI at and below v, and BI suggests the best response for $P(v)$ at v. □

[5]Theorem 2 and its proof were added on May 30, 2011.

6 Discussion

1. Coherence of knowledge and revision. It makes sense to consider some additional properties of extended Aumann models. For example, it seems natural to assume that if a vertex is reached, all players know this, i.e., if v is reached in some state in $\mathcal{K}_i(\omega)$, then v is reached in all states in $\mathcal{K}_i(\omega)$. This property is met in the Stalnaker example, model $\mathcal{A}$.

2. Error tolerance levels. For games with a 'small' number of irrational moves, robust knowledge of rationality can be justified by a strong reputation for the rationality of players, their history of rational behavior, etc. An isolated irrational move can be viewed as a technical error. However, trust in rationality can fade with each irrational move and given a 'large' number of such moves, robust knowledge of rationality could become unfeasible. More realistic models of robust rationality could include an error-tolerance level, i.e., a bound on the number of errors (e.g., one) allowed for each player.

7 Acknowledgments

The author is greatly indebted to Robert Aumann for his interest and encouraging support of this work. The author is deeply grateful to Adam Brandenburger for his guidance and inspiring discussions. This paper substantially benefited from insightful comments by Christian Bach. Many thanks also to Vladimir Krupski, Elena Nogina, and Çağıl Taşdemir for numerous useful suggestions.

Special thanks to Karen Kletter for editing this paper.

References

[Aumann, 1987] R. Aumann. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55(1):1–18, 1987.

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

[Aumann, 2010] R. Aumann. Epistemic Logic: 5 Questions. Vincent F. Hendricks and Olivier Roy, eds. Automatic Press/VIP, pp. 21-33, 2010.

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

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

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

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

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