Substantive Rationality and Backward Induction

Joseph Y. Halpern*
Cornell University
Computer Science Department
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

November 24, 1998

Abstract

Aumann has proved that common knowledge of substantive rationality implies the backwards induction solution in games of perfect information. Stalnaker has proved that it does not. Roughly speaking, a player is *substantively rational* if, for all vertices v, if the player were to reach vertex v, then the player would be rational at vertex v". It is shown here that the key difference between Aumann and Stalnaker lies in how they interpret this counterfactual. A formal model is presented that lets us capture this difference, in which both Aumann's result and Stalnaker's result are true (under appropriate assumptions).

Starting with the work of Bicchieri [1988, 1989], Binmore [1987], and Reny [1992], there has been intense scrutiny of the assumption of common knowledge of rationality, the use of counterfactual reasoning in games, and the role of common knowledge and counterfactuals in the arguments for backward

*Supported in part by NSF under grant IRI-96-25901.

induction in games of perfect information. Startlingly different conclusions were reached by different authors.

These differences were clearly brought out during a 2.5 hour round table discussion on "Common knowledge of rationality and the backward induction solution for games of perfect information" at the recent TARK (Theoretical Aspects of Rationality and Knowledge) conference. During the discussion, Robert Aumann and Robert Stalnaker stated the following theorems:

Aumann's Theorem (informal version): Common knowledge of substantive rationality implies the backwards induction solution in games of perfect information.

Stalnaker's Theorem (informal version): Common knowledge of substantive rationality does not imply the backwards induction solution in games of perfect information.

The discussion during the round table was lively, but focused on more philosophical, high-level issues. My goal in this short note is to explain the technical differences between the framework of Aumann and Stalnaker that lead to the different results. Aumann proves his theorem in [Aumann 1995]. I show here what changes need to be made to Aumann's framework to get Stalnaker's result.[1] I believe that the points that I make here are well known to some (and certainly were made informally during the discussion). Nevertheless, there does not appear to be a careful comparison of the differences between the models in the literature. I hope this note will help to clarify a few issues and put the debate on a more rational footing.

There are three terms in the theorems that need clarification:

- (common) knowledge
- rationality
- *substantive* rationality

[1]The model that I use to prove Stalnaker's result is a variant of the model Stalnaker uses in [Stalnaker 1996], designed to be as similar as possible to Aumann's model, to bring out the key differences. This, I believe, is essentially the model that Stalnaker had in mind at the round table.

I claim that Stalnaker's result can be obtained using exactly the same definition of (common) knowledge and rationality as the one Aumann used in [Aumann 1995]. The definition of knowledge is the standard one, given in terms of partitions. (I stress this point because Stalnaker [1996] has argued that probability-1 belief is more appropriate than knowledge when considering games.) The definition of rationality is that a player who uses strategy s is rational at vertex v if there is no other strategy that he knows will give him a better payoff, conditional on being at vertex v. Both Aumann and Stalnaker give substantive rationality the same reading: "rationality at all vertices v in the game tree". They further agree that this involves a counterfactual statement: "for all vertices v, if the player were to reach vertex v, then the player would be rational at vertex v". The key difference between Aumann and Stalnaker lies in how they interpret this counterfactual. In the rest of this note, I try to make this difference more precise.

The Details

I start by considering Aumann's model. Fix a game Γ of perfect information for n players. As usual, we think of Γ as a tree. Because Γ is a game of perfect information, the players always know which vertex in the tree describes the current situation in the game. The nonleaf vertices in Γ are partitioned into n sets, $G_1, \ldots, G_n$, one for each player. The vertices in G_i are said to belong to i; these are the ones where player i must move. A *model of* Γ is a tuple $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$, where Ω of states of the world, $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are partitions, one for each player $i = 1, \ldots, n$ ($\mathcal{K}_i$ is i's information partition), and $\mathbf{s}$ maps each world $\omega \in \Omega$ to a strategy profile $\mathbf{s}(\omega) = (s_1, \ldots, s_n)$; s_i is i's strategy in game Γ at state ω. As usual, a strategy for i in Γ is just a mapping from i's vertices in Γ to actions. I write $\mathbf{s}_i(\omega)$ for s_i.

Let $\mathcal{K}_i(\omega)$ denote the cell in partition $\mathcal{K}_i$ that includes ω. Define the operator K_i on events as usual:

$$K_i(E) = \{\omega : \mathcal{K}_i(\omega) \subseteq E\}.$$

$K_i(E)$ is the event that i knows E. Let $A(E) = K_1(E) \cap \ldots \cap K_n(E)$. $A(E)$ is the event that everyone (*all* the players) know E. Let

$$CK(E) = A(E) \cap A(A(E)) \cap A(A(A(E))) \cap \ldots$$

$CK(E)$ is the event that E is common knowledge.

Aumann and Stalnaker (and everyone else who has written on this subject that I am aware of) assume that the players know their strategies. Formally, that means that if $\omega' \in \mathcal{K}_i(\omega)$, then $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$; that is, i uses the same strategy at all the states in a cell of $\mathcal{K}_i$.

Next we need to define rationality. Note that a strategy profile s and vertex v uniquely determine a path in Γ that would be followed if s were played starting at v. Let $h_i^v(s)$ denote i's payoff if this path is followed. Informally, i is rational at vertex v if there is no strategy that i could have used that i knows would net him a higher payoff than the strategy he actually uses. More precisely, *i is rational at vertex v in ω* if, for all strategies $s^i \neq \mathbf{s}_i(\omega)$, $h_i^v(\mathbf{s}(\omega')) \geq h_i^v((\mathbf{s}_{-i}(\omega'), s^i)$ for some $\omega' \in \mathcal{K}_i(\omega)$. That is, i cannot do better by using s^i than $\mathbf{s}_i(\omega)$ against all the strategy profiles of the other players that he considers possible at ω. This is a weak notion of rationality (which is certainly satisfied by expected utility maximization). By taking such a weak notion, Aumann's Theorem becomes stronger. As will be clear from the example, Stalnaker's Theorem holds even if we strengthen the requirements of rationality (to require strict inequality, for example).

Aumann then defines *substantive rationality* to mean rationality at all vertices in the game tree. That is, i is substantively rational in state ω if i is rational at vertex v in ω for every vertex $v \in G_i$. For future reference, I call Aumann's notion of substantive rationality A-rationality.

Using these definitions, Aumann can and does prove his theorem (using a straightforward backward induction argument).

Stalnaker's definition of substantive rationality is different from Aumann's although, as I indicated above, he is trying to capture the same intuition. His definition tries to enforce the intuition that, for every vertex $v \in G_i$, *if* i were to actually reach v, then what he would do in that case would be rational. The key point is that, according to Stalnaker's definition, in order to evaluate, at state ω, whether i is being rational at vertex v by performing the action dictated by his strategy at ω, we must consider i's beliefs in the state "closest" to ω according to i where v is actually reached.

To formalize this, we must add one more component to Aumann's model: for each player i, we must have a selection function f mapping states and i's vertices to states. An *extended model of* Γ is a tuple $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$, where $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$ is a model of Γ and $f : \Omega \times G_i \to \Omega$. Intuitively, if

$f(\omega, v) = \omega'$, then state ω' is the state closest to ω where vertex v is reached.[2]

Given this intuition, we may want to assume that f satisfies some constraints such as the following:

F1. v is reached in $f(\omega, v)$ (that is, v is on the path determined by $\mathbf{s}(f(\omega, v))$).

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 Γ below v.

F1 guarantees that v is actually reached in $f(\omega, v)$, while F2 says that if v is actually reached in ω, then ω is the closest state to itself where v is reached. F3 is intended to capture the intuitive meaning of a strategy. If, according to $\mathbf{s}_i(\omega)$, player i performs action a at vertex v, it seems reasonable to expect that at the closest world where v is actually reached, player i does in fact perform a. This follows from F3. However, F3 says more than this. It says that at all the vertices below v, all the players also perform the actions dictated by $\mathbf{s}(\omega)$. This extra requirement arguably makes F3 too strong. However, as I shall show, Stalnaker's Theorem continues to hold even with this strong assumption.[3]

According to Stalnaker, i is substantively rational in state ω if i is rational at vertex v in $f(\omega, v)$ for every vertex $v \in G_i$. Let us call this notion *S-rationality*. Thus, the crucial (and, in fact, only) difference between Aumann's approach and Stalnaker's approach is that A-rationality requires i to be rational at vertex v in ω and S-rationality requires i to be rational at vertex v in $f(\omega, v)$.

The difference can perhaps be best understood by considering the game described in Figure 1, which is a variant of a game introduced by Aumann [1995].

[2] Again, I should stress that this is not exactly the model that Stalnaker uses in [Stalnaker 1996], but it suffices for my purposes. I remark that in [Halpern 1998], I use selection functions indexed by the agents, so that agent 1 may have a different selection function than agent 2. I do not need this greater generality here, so I consider the simpler model where all agents use the same selection function.

[3] There are certainly other reasonable properties we could require of the selection function. For example, we might want to require that if v is reached in some state in $\mathcal{K}_i(\omega)$, then $f(\omega, v) \in \mathcal{K}_i(\omega)$. I believe that it is worth trying to characterize the properties we expect the selection function should have, but this issue would take us too far afield here. Note that F1–F3 are properties that seem reasonable for arbitrary games, not just games of perfect information.

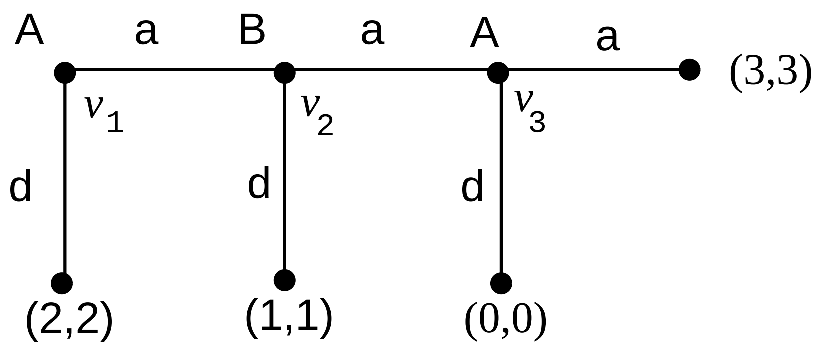

Figure 1: A variant of Aumann's Ann-Bob game.

Consider the following five strategy profiles:

- s^1 is the strategy profile (da, d) (i.e., Ann goes down at v_1 and across at v_3 and Bob goes down at v_2);
- s^2 is the strategy profile (aa, d);
- s_3 is the strategy profile (ad, d);
- s^4 is the strategy profile (aa, a);
- s_5 is the strategy profile (ad, a).

Note that s^4 is the backward induction solution.

Now consider the extended model $(\omega, \mathcal{K}_{Ann}, K_{Bob}, \mathbf{s}, f)$ of this game, where

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$;
- $\mathcal{K}_{Ann}(\omega_i) = \{\omega_i\}$, for $i = 1, \ldots, 5$;
- $\mathcal{K}_{Bob}(\omega_i) = \{\omega_i\}$ for $i = 1, 4, 5$; $\mathcal{K}_{Bob}(\omega_2) = \mathcal{K}_{Bob}(\omega_3) = \{\omega_2, \omega_3\}$;
- $\mathbf{s}(\omega_i) = s^i$, for $i = 1, \ldots, 5$;
- f is the unique selection function satisfying F1–F3.

It is easy to check that, at ω_1, it is common knowledge that strategy profile s^i is being used, for $i = 1, 2$. It is also common knowledge at ω_1 that, if vertex v_2 were reached, Bob would play down.

In this extended model, clearly Bob is not rational at vertex v_2 in ω_1, since he plays down. This means that we do not have A-rationality at ω_1 (and, *a fortiori*, we do not have common knowledge of A-rationality at ω_1). On the other hand, Bob is rational at vertex v_2 in ω_2, since Bob considers it possible that Alice may go down at v_3 (since $\mathcal{K}_{Bob}(\omega_2) = \{\omega_2, \omega_3\}$). Similarly, Alice is rational at vertex v_3 in ω_4. Since $f(\omega_1, v_2) = \omega_2$ and $f(\omega_1, v_3) = \omega_4$, it follows that we have S-rationality at ω_1, and hence common knowledge of S-rationality at ω_1.

This example is an instance of Stalnaker's Theorem: we have common knowledge of substantive rationality in the sense of S-rationality at ω_1, yet the backward induction solution is not played at ω_1. Nevertheless, it does not contradict Aumann's Theorem, since we do not have common knowledge of A-rationality.

With this machinery, we can now state Aumann's Theorem and Stalnaker's Theorem more formally. Let *S-RAT* consist of all states where all the players are S-rational; let *A-RAT* consist of all states where all the players are A-rational; let *BI* consist of all states where the backward induction solution is played.

Aumann's Theorem: If Γ is a nondegenerate[4] game of perfect information, then in all models of Γ, we have $CK(A\text{-}RAT) \subseteq BI$.

Stalnaker's Theorem: There exists a nondegenerate game Γ of perfect information and an extended model of Γ in which the selection function satisfies F1–F3 such that $CK(S\text{-}RAT) \not\subseteq BI$.

Note that, in an extended model of the Ann-Bob game, it is consistent for Ann to say "Although it is common knowledge that I would play across if v_3 were reached, if I were to play across at v_1, Bob would consider it possible that I would play down at v_3." This is not possible in Aumann's framework because, without selection functions, Aumann has no way of allowing the agents to revise their beliefs. (This point is essentially made by Samet [1996].) In the definition of A-rationality, for any vertex v, player i's beliefs in state ω about the possible strategies player j may be using if vertex v is reached are the same (and are determined by $\mathcal{K}_i(\omega)$). It is crucial for Aumann's result (and, I believe, a weakness in his model) that players do not (and

[4] A game is nondegenerate if the payoffs are different at all the leaves.

cannot) revise their beliefs about other player's strategies when doing such hypothetical reasoning.

It is not hard to place a condition on selection functions that guarantees that players beliefs about other player's strategies do not change whatever vertex they may be in.

F4. For all players i and vertices v, if $\omega' \in \mathcal{K}_i(f(\omega, v))$ then there exists a state $\omega'' \in \mathcal{K}_i(\omega)$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree on the subtree of Γ below v.[5]

Combined with F1–F3, F4 this gives us the properties we want. In fact, we have the following result.

Theorem 1: *If Γ is a nondegenerate game of perfect information, then for every extended model of Γ in which the selection function satisfies F1–F4, we have $CK(S\text{-}RAT) \subseteq BI$. Moreover, there is an extended model of Γ in which the selection function satisfies F1–F4.*

Proof: For the first half of the theorem, suppose $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$ is an extended model of Γ in which the selection function satisfies F1–F4. Further suppose that $\omega \in CK(S\text{-}RAT)$. We want to show $\omega \in BI$. The proof basically mimics Aumann's proof of his theorem, since F1–F4 essentially gives us his framework.

I first recall a standard characterization of common knowledge. Define the notion of *omega$''$ being reachable from ω' in k steps* inductively: *ω'' is reachable from ω' in 1 step* iff $\omega'' \in \mathcal{K}_i(\omega')$ for some $i \in \{1, \ldots, n\}$; *ω'' is reachable from ω' in $k+1$ steps* iff there exists a state ω''' that is reachable from ω' in 1 step such that ω'' is reachable from ω''' in k steps. We say that *ω'' is reachable from ω'* if ω'' is reachable from ω' in k steps for some k. It is well known [Aumann 1976] that $\omega' \in CK(E)$ iff $\omega'' \in E$ for all ω'' reachable from ω'.

I show by induction on k that for all states ω' reachable from ω, if v is a vertex which is at height k in the game tree (i.e., k moves away from a leaf),

[5] Actually, F4 says that in $f(\omega, v)$, player i considers at least as many strategies possible as at ω. To capture the fact that player i's beliefs about other player's possible strategies does not change, we would need the opposite direction of F4 as well: if $\omega' \in \mathcal{K}_i(\omega)$ then there exists a state $\omega'' \in \mathcal{K}_i(f(\omega, v))$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree on the subtree of Γ below v. I do not impose this requirement here simply because it turns out to be unnecessary for Aumann's Theorem.

the move dictated by the backward induction solution (for the subgame of Γ rooted at v) is played at v in state ω'.

For the base case, suppose v is at height 1 and ω' is reachable from ω. Since $\omega \in CK(S\text{-}RAT)$, we must have $\omega' \in S\text{-}RAT$. Suppose player i moves at ω'. Since $\omega' \in S\text{-}RAT$, player i must make the move dictated by the backwards induction solution at $f(\omega', v)$. By F3, he must do so at ω' as well.

For the inductive step, suppose that v is at height $k+1$, player i moves at v, and ω' is reachable from ω. Suppose, by way of contradiction, that a is the action indicated by the backward induction solution at v but $\mathbf{s}_i(\omega')(v) = a' \neq a$. Note that by the induction hypothesis, at every vertex below v, all the players play according to the backward induction solution in state ω'. Since $\omega' \in S\text{-}RAT$, we must have that i is rational at v in $f(\omega', v)$. By F3, it follows that i plays a' at vertex v in $f(\omega', v)$ and at every vertex below v, the players play according to the backward induction solution. Thus, there must be a state $\omega'' \in \mathcal{K}_i(f(\omega', v))$ such that by using $\mathbf{s}_i(f(\omega', v))$, player i does at least as well in ω'' as by using the backward induction strategy starting from v. By F4, there must exist some $\omega''' \in \mathcal{K}_i(\omega')$ such that $\mathbf{s}(\omega'')$ and $\mathbf{s}(\omega''')$ agree on the subtree of Γ below v. Since ω''' is reachable from ω, by the induction hypothesis, all players use the backward induction solution at vertices below v. By F3, this is true at ω'' as well. However, this means that player i does better at ω'' playing a at v than a', giving us the desired contradiction.

For the second half, given a nondegenerate game Γ of perfect information, let s be the strategy where, at each vertex v, the players play the move dictated the backward induction solution in the game defined by the subtree below v. For each vertex v, let s_v be the strategy where the players play the actions required to reach vertex v, and then below v, they play according to s. Note that if v is reached by s, then $s_v = s$. In particular, if r is the root of the tree, then $s_r = s$. Consider the extended model $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$ where $\Omega = \{\omega_v : v \text{ is a vertex of } \Gamma\}$, $\mathcal{K}_i(\omega_v) = \{\omega_v\}$, $\mathbf{s}(\omega_v) = s_v$, and $f(\omega, v)$ is ω if v is reached by $\mathbf{s}(\omega)$ and ω_v otherwise. I leave it to check that this gives us an extended model where the selection function satisfies F1–4. ∎

Theorem 1 and the earlier discussion suggests that one possible culprit for the confusion in the literature regarding what is required to force the backwards induction solution in games of perfect information is the notion of a *strategy*. Exactly what should it mean to say that Alice's strategy at a state ω is s? For example, consider the game in Figure 1. According to strategy s_A^1, Alice plays across at vertex v_3. But v_3 is a vertex that cannot be reached if Alice uses s_A^1, since according to this strategy, Alice plays down at v_1. The standard reading of $s_A^1(v_3) = a$ is that "*if* v_3 is reached, then Alice plays across". But this reading leaves a number of questions unanswered. How Alice plays (if she is rational) depends on her beliefs. Should we read this as "no matter what Alice's beliefs are, if v_3 is reached, Alice will play a"? Or perhaps it should be "given her current beliefs (regarding, for example, what move Bob will make), if v_3 is reached, Alice will play a". Or perhaps it should be "in the state 'closest' to the current state where v_3 is actually reached, Alice plays a". I have taken the last reading here (where 'closest' is defined by the selection function); assumption F3 essentially forces it to be equivalent to the second reading.

However, without F4, this equivalence is not maintained with regard to Bob's beliefs. That is, consider the following two statements:

- Bob currently believes that, given Alice's current beliefs, Alice will play a if v_3 is reached;
- in the state closest to the current state where v_3 is reached, Bob believes that Alice plays a at v_3.

The first statement considers Bob's beliefs at the current state; the second considers Bob's beliefs at a different state. Without F4, these beliefs might be quite different. It is this possible difference that leads to Stalnaker's Theorem.

Strategies themselves clearly involve counterfactual reasoning. If we take strategies as primitive objects (as both Aumann and Stalnaker do, and as I have done for consistency), we have two sources of counterfactuals in extended models: selection functions and strategies. Stalnaker [1996, p. 135] has argued that "To clarify the causal and epistemic concepts that interact in strategic reasoning, it is useful to break them down into their component parts." This suggests that it would be useful to have a model where strategy is *not* a primitive, but rather is defined in terms of counterfactuals. This is precisely what Samet [1996] does.[6]

[6]Samet does not use selection functions to capture counterfactual reasoning, but *hypothesis transformations*, which map cells (in the information partition) to cells. However, as I have shown [1998], we can capture what Samet is trying to do by using selection functions.

Not surprisingly, in Samet's framework, Aumann's Theorem does not hold without further assumptions. Samet shows that what he calls a *common hypothesis* of rationality implies the backward induction solution in nondegenerate games of perfect information. Although there are a number of technical differences in the setup, this result is very much in the spirit of Theorem 1.

Acknowledgment: I'd like to thank Robert Stalnaker for his many useful comments and criticisms of this paper.

References

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics 4*(6), 1236–1239.

Aumann, R. J. (1995). Backwards induction and common knowledge of rationality. *Games and Economic Behavior 8*, 6–19.

Bicchieri, C. (1988). Strategic behavior and counterfactuals. *Synthese 76*, 135–169.

Bicchieri, C. (1989). Self refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis 30*, 69–85.

Binmore, K. (1987). Modeling rational players I. *Economics and philosophy 3*, 179–214. Part II appeared ibid., *4*, 9–55.

Halpern, J. Y. (1998). Hypothetical knowledge and counterfactual reasoning. In *Theoretical Aspects of Rationality and Knowledge: Proc. Seventh Conference*, San Francisco, Calif., pp. 83–96. Morgan Kaufmann. To appear, *International Journal on Game Theory*.

Reny, P. (1992). Rationality in extensive form games. *Journal of Economic Perspectives 6*, 103–118.

Samet, D. (1996). Hypothetical knowledge and games with perfect information. *Games and Economic Behavior 17*, 230–251.

Stalnaker, R. C. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy 12*, 133–163.