Topics in Theoretical Economics

Volume 4, *Issue* 1 2004 *Article* 8

Counterfactual Reasoning and Common Knowledge of Rationality in Normal Form Games

Eduardo Zambrano*

*University of Notre Dame, ezambran@nd.edu

Copyright ©2004 by the authors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress, which has been given certain exclusive rights by the author. *Topics in Theoretical Economics* is one of *The B.E. Journals in Theoretical Economics*, produced by The Berkeley Electronic Press (bepress). http://www.bepress.com/bejte.

Counterfactual Reasoning and Common Knowledge of Rationality in Normal Form Games*

Eduardo Zambrano

Abstract

When evaluating the rationality of a player in a game one has to examine counterfactuals such as "what would happen if the player were to do what he does not do?" In this paper I develop a model of a normal form game where counterfactuals of this sort are evaluated as in the philosophical literature (cf. Lewis, 1973; Stalnaker, 1968). According to this method one evaluates a statement like "what would the player believe if he were to do what he does not do" at the world that is closest to the actual world where the hypothetical deviation occurs. I show that in this model common knowledge of rationality need not lead to rationalizability. I also present assumptions that allow rationalizability to follow from common knowledge of rationality. These assumptions suggest that rationalizability may not rely on weaker assumptions about belief consistency than Nash equilibrium.

KEYWORDS: Common knowledge, counterfactual reasoning, interactive epistemology, rationalizability.

*I would like to thank David Easley, Larry Blume, Joseph Halpern, Edward O'Donoghue, Kaushik Basu, participants at a Cornell seminar, the editor and two anonymous referees for their comments on a previous version of this paper. All errors are naturally my own.

1 Introduction

There is now a rich literature that explores noncooperative games in terms of the rationality of the players and their epistemic state: what they know about the game and about each other's rationality, strategies, knowledge and beliefs.[1] The main goal of this literature is to determine exactly what assumptions on the decision theoretic problem of each player would justify game theoretic solution concepts such as rationalizability, Nash equilibrium and backward induction, among others.

It is known that from an epistemic point of view not all solution concepts are created equal. For example, while virtually everyone in the literature agrees that common knowledge of rationality leads to rationalizability in normal form games, strong disagreement persists regarding whether common knowledge of rationality leads to backward induction in perfect information games. The source of this disagreement seems to be the serious difficulties that arise in the extensive form when formulating the relevant concepts (knowledge, rationality, etc.) because of the need to address counterfactual statements like "if the player were to reach a certain node, which he *knows* he won't, he would be *rational* from there on."

The purpose of this paper is to point out that the epistemic justification of solution concepts in normal form games can be problematic for reasons that are identical to those that complicate the justification of backward induction in extensive games. The key to the problem is that, to justify the rationality of a player at a given state one has to consider *what would the player believe if he were to do what he actually does not do.* The problem arises because, since the state specifies the strategy chosen and each player knows his own strategy, there is no state that the agent considers possible in which he deviates, and therefore, given the deviation, beliefs about strategies are not well-defined.

A way out of this situation is simply to ignore the fact that a deviation is inconsistent with the player's knowledge and to define beliefs given a deviation to be as if he weren't deviating. This is the traditional manner in which this situation is handled in the game theoretic literature and, to be sure, one in which common knowledge of rationality leads to rationalizability. This characterization leaves us, however, wondering just what assumptions are being made implicitly about the treatment of counterfactuals in this for-

[1] For excellent surveys of the literature see Binmore and Brandenburger [7], Geanakoplos [12], Dekel and Gul [11], and Battigalli and Bonanno [8].

mulation and whether there are other ways around this problem that are just as sensible from a decision theoretic standpoint.

To study this carefully, in this paper I use the method by which counterfactuals are captured in the philosophical literature after the work of Lewis [14] and Stalnaker [18]. The aim of this method, as Stalnaker [19] puts it, is generality: "to make, in the definition of a model, as few substantive assumptions as possible about the epistemic states and behavior of players of a game in order that substantive assumptions can be made explicit as conditions that distinguish some models from others."[2]

According to this method one evaluates a statement like "what would the player believe if he were to do what he actually does not do" at the state that is *closest* to the actual state *in which the hypothetical deviation actually occurs*. I define beliefs given a deviation in this manner (that is, with respect to the state that is closest to the actual state in which the deviation occurs) and then rationality as payoff maximization given those beliefs. I then show that common knowledge of rationality in this model with counterfactuals *does not* lead to rationalizability.

This reveals that the traditional model makes assumptions about the treatment of counterfactuals that are far from innocuous in that they are critical in generating the connections between rationality, knowledge and rationalizability. In this paper I also provide a statement of what those assumptions are. What they reveal is very interesting: that to justify the notion of rationalizability in epistemic models one needs to make assumptions about beliefs "off the equilibrium path," (that is, given deviations from the strategies prescribed at a given state) that are very similar to those assumptions underlying refinements of Nash equilibrium such as subgame perfection in extensive games. This implies that it may be misleading to believe that, from an epistemic point of view, rationalizability relies on weaker assumptions about belief consistency than Nash equilibrium. I believe that the points that I make here may be known to some (for example, Brandenburger and Dekel [10] reached a similar conclusion in a very different setup). Nevertheless, there does not appear to be a careful discussion of this subtle issue regarding counterfactual reasoning in normal form games in the literature.

The rest of this paper is aimed at presenting a coherent formulation and proof of the claims made above. In Section 2 I present the formal statement of the results and their proofs. In Section 3 I discuss the results. I conclude

[2] Stalnaker [19, p. 140].

in Section 4. Although I present all the results for the case of two players I do not foresee any difficulties in replicating the results for games with any finite number of players when the relevant solution concept is that of *correlated* rationalizability.

2 The Results

In what follows I use the standard notation and definitions as presented, for example, in Osborne and Rubinstein [16], Aumann [2] and Halpern [13]. Let a two-player game in normal form $G = \langle S_1, S_2, U_1, U_2 \rangle$ be given where, for each player i, S_i is the set of strategies that are open to player i and U_i is a function that represents player i's preference relation over $S = S_1 \times S_2$. A *model* of G is a tuple $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{s})$, where Ω is a set of states of the world, $\mathcal{K}_i$ is player i's information partition of Ω, $\mathbf{p}_i(\cdot;\omega)$ maps each state $\omega \in \Omega$ to a probability distribution on the cell $\mathcal{K}_i(\omega)$ in partition $\mathcal{K}_i$ that includes ω. The interpretation is that player i has probabilistic beliefs on Ω at state ω given by $\mathbf{p}_i(\cdot;\omega)$. Finally, $\mathbf{s}$ maps each state $\omega \in \Omega$ to a strategy profile $\mathbf{s}(\omega) = (s_1, s_2)$. I write $\mathbf{s}_i(\omega)$ for s_i. The interpretation is that player i chooses strategy $\mathbf{s}_i(\omega)$ at state ω.

As usual, I assume that every player i knows his or her own type, that is, $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$ and $\mathbf{p}_i(\cdot;\omega) = \mathbf{p}_i(\cdot;\omega')$ whenever $\omega' \in \mathcal{K}_i(\omega)$.

An event E is a set of states. If E is an event define the operator K_i on events by

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

$K_i(E)$ is the event that player i knows E. Set $K(E) = K_1(E) \cap K_2(E)$ and

$$CK(E) = K(E) \cap K(K(E)) \cap K(K(K(E))) \cap \cdots$$

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

An event F is *self-evident* if $\mathcal{K}_i(\omega) \subseteq F$ for every $\omega \in F$ and every $i = 1, 2$. It is well known (cf. Osborne and Rubinstein [16, Ch. 4]) that an event E is common knowledge if and only if there is a self evident event F such that $F \subseteq E$.

I will denote a point mass probability distribution on ω by 1_ω and a probability distribution q on a set $\{\omega_1, \ldots, \omega_n\}$ by $q_1\omega_1 + \cdots + q_n\omega_n$.

To abbreviate I will say that strategy s_i is a best response to a probability q_i on Ω whenever s_i is a best response to the distribution on S_j induced by q_i and conditioned on player i's information.

A strategy $s_i \in S_i$ is *rationalizable* if for every player $j = 1, 2$ there is a set $Z_j \subseteq S_j$ such that

- $s_i \in Z_i$, and
- every strategy $s_j \in Z_j$ is a best response to a belief q_j of player j whose support is a subset of Z_i.

2.1 A-Rationality

Whether a player is rational at a given state ω depends both on the strategy that the player uses at that state ω and the belief that the player maintains about the opponents when considering deviations. However, as mentioned in the Introduction, it is not completely clear what beliefs should the player hold when considering deviations from the strategy prescribed in state ω because his own knowledge is contradicted in the face of such deviation: at every state that he considers possible he *does not* deviate!

The traditional notion of rationality in the epistemic literature ignores this and defines beliefs given a deviation to be as if the player weren't deviating. Formally, one says that player i is *A-rational* at ω if there is no $s_i \neq \mathbf{s}_i(\omega)$ such that

$$\sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; \omega) U_i(s_i, \mathbf{s}_j(\omega')) > \sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; \omega) U_i(\mathbf{s}_i(\omega), \mathbf{s}_j(\omega')).$$

In words, player i is A-rational at ω if his strategy maximixes his expected payoff given his belief at ω.[3]

Using these definitions it can be easily shown following traditional methods that if at a state ω A-rationality is common knowledge then the strategy profile $\mathbf{s}(\omega)$ is rationalizable. Let $A\text{-}RAT$ consist of all the states where all the players are A-rational. Let R consist of all the states where the strategies chosen by the players at those states are rationalizable. The formal statement is

Theorem 1 $CK(\text{A-RAT}) \subseteq R$.

Proof. See Osborne and Rubinstein [16, p. 80]. ■

[3] The "A" in A-rationality is for Robert Aumann, following notation used in Halpern [13].

2.2 W-Rationality and S-Rationality

I now present two alternative notions of rationality that provide an explicit treatment of counterfactuals such as "what would the player believe if he were to do what he actually does not do?" As mentioned in the Introduction, I use definitions that are close to the ones used by Stalnaker and Halpern in their discussion of backward induction. Informally, player i is *W-rational* at ω if there is no deviation $s_i \neq \mathbf{s}_i(\omega)$ such that strategy s_i is preferred to $\mathbf{s}_i(\omega)$ given the belief that player player i holds *at the state closest to* ω *in which* i *deviates to* s_i. Alternatively, player i is *S-rational* at ω if *at each state closest to* ω *in which* i *deviates to every possible* $s_i \neq \mathbf{s}_i(\omega)$ player i holds belief p_i and $\mathbf{s}_i(\omega)$ is a best response to p_i.[4] To make these definitions precise we must specify what it means for a state to be the "closest" state to ω.

To formalize I add an additional component to the traditional definition of a model of G. An *extended model* of G is a tuple $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{s}, f_1, f_2)$ where $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{s})$ is a model of G and for every player i the function f_i (the selection function for player i) maps states and strategies for player i into states, that is, $f_i : \Omega \times S_i \to \Omega$. Intuitively, if $f_i(\omega, s_i) = \omega'$, then the state ω' is the state closest to ω, according to player i, in which player i deviates from the strategy prescribed by ω and, instead, plays s_i. To capture this intuition I assume that for every player i the function f_i satisfies the following conditions:

F1. The deviation s_i takes place in $f_i(\omega, s_i)$, that is, $\mathbf{s}_i(f_i(\omega, s_i)) = s_i$, and

F2. Player i is the *only* one that deviates from $\mathbf{s}(\omega)$ in $f_i(\omega, s_i)$, that is, $\mathbf{s}_j(f_i(\omega, s_i)) = \mathbf{s}_j(\omega)$ for all $s_i \in S_i$.

F1 guarantees that the deviation s_i takes place in $f_i(\omega, s_i)$ while F2 is intended to capture the intuitive meaning of an unilateral deviation: at the closest state to ω in which player i contemplates a particular deviation, player j still plays $\mathbf{s}_j(\omega)$. Notice that F1 and F2 imply that, if s_i is chosen at ω, the strategy profile chosen at the closest state to ω where i chooses s_i is also s.[5]

[4] The "W" in W-rationality is for "weakly." The "S" in S-rationality is for Robert Stalnaker, following notation used in Halpern [13].

[5] For a discussion and other applications of selection functions in the epistemic literature see Halpern [13], Stalnaker [19] and Stalnaker [20].

The stage is set to define what rationality means in this context. Formally, one says that player i is *W-rational* at ω if there is no $s_i \neq \mathbf{s}_i(\omega)$ such that

$$\sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; f(\omega, s_i)) U_i(s_i, \mathbf{s}_j(\omega')) > \sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; f(\omega, s_i)) U_i(\mathbf{s}_i(\omega), \mathbf{s}_j(\omega')).$$

In words, player i is W-rational if there is no deviation $s_i \neq \mathbf{s}_i(\omega)$ such that strategy s_i is preferred to $\mathbf{s}_i(\omega)$ given the belief that player player i holds *at the state closest to ω in which i deviates to s_i*. The interpretation is that the rationality of choosing strategy $\mathbf{s}_i(\omega)$ at state ω against a deviation $s_i \neq \mathbf{s}_i(\omega)$ is determined with respect to beliefs that arise at the closest state to ω in which s_i is actually chosen, that is, with respect to beliefs at $f(\omega, s_i)$.

W-rationality is designed to be a notion of what it means for a strategy to be rational when beliefs facing a deviation are consistent with the deviation rather than with the actual strategy chosen. It is a notion of rationality that is weaker than the one implied by traditional expected utility theory, yet it is consistent with it. This consistency can be captured formally. Let U consist of all the states where the strategies chosen by the players at those states are not strictly dominated by a mixed strategy and $U^P \supseteq U$ consist of all the states where the strategies chosen by the players at those states are not strictly dominated by a pure strategy. Let *W-RAT* consist of all the states where all the players are W-rational. The formal statement is

Lemma 2 W-RAT $\subseteq U^P$. *Moreover, for every normal form game G there is an extended model in which the selection functions satisfy F1-F2 such that* $U \subseteq$ W-RAT.

Proof. See the Appendix. ■

As seen, W-rationality has almost as much bite as the traditional notion of domination. A W-rational player will never play a strategy that is strictly dominated by another pure strategy. Moreover, when a player chooses a strategy that is strictly dominated by a mixed strategy, this can be seen as a consequence of the player not being W-rational.

Example 1 *Consider the game in Figure 1 (A Prisoner's Dilemma). Because* W-RAT$\subseteq U^P$ *a W-rational player will never play "cooperate" in this*

game.

		Bob	
		D	C
Ann	D	$3,3$	$10,2$
	C	$2,10$	$5,5$

(1)

Figure 1. *A 'Prisoner's Dilemma'*

Despite this consistency, it is important to notice that common knowledge of W-rationality does not lead to rationalizability.

Example 2 *Consider the game in Figure 2, taken from Osborne and Rubinstein [16, p. 61].*

		Bob	
		L	R
	T	$3,0$	$0,1$
Ann	M	$0,0$	$3,1$
	B	$1,1$	$1,0$

(2)

Figure 2.

Notice that the unique rationalizable strategy profile for this game is MR.

Consider the following extended model $(\Omega, \mathcal{K}_{Ann}, \mathcal{K}_{Bob}, \mathbf{p}_{Ann}, \mathbf{p}_{Bob}, \mathbf{s}, f_{Ann}, f_{Bob})$ of this game, where

- $\Omega = \{\omega_1, \ldots, \omega_6\}$,
- $\mathcal{K}_{Ann}(\omega_l) = \{\omega_l\}$ for $l \neq 3,4$; $\mathcal{K}_{Ann}(\omega_l) = \{\omega_3, \omega_4\}$ for $l = 3,4$;
- $\mathcal{K}_{Bob}(\omega) = \{\omega\}$
- $\mathbf{p}_{Ann}(\cdot \mid \omega_l) = 1_{\omega_l}$ for $l \neq 3,4$; $\mathbf{p}_{Ann}(\cdot \mid \omega_l) = \frac{1}{4}\omega_3 + \frac{3}{4}\omega_4$ for $l = 3,4$;
- $\mathbf{p}_{Bob}(\cdot \mid \omega) = 1_{\omega}$;
- $\mathbf{s}(\omega_l) = s^l$ for $l = 1, \ldots, 6$, where s^l is the $l-th$ element in the following enumeration of S: $\{BL, ML, TL, TR, MR, BR\}$, and
- f_{Ann} and f_{Bob} are the unique selection functions satisfying F1-F2.

In this extended model of the game in Figure 2 the selection functions satisfy F1-F2 and yet $CK(W\text{-}RAT) \nsubseteq R$.

To see this I wish to show that in this extended model Ann is W-rational at ω_1. Ann's choice at ω_1 is B. Consider Ann's deviation to M. What belief does Ann hold when she deviates to M? The closest state to ω_1 in which Ann chooses M is ω_2. At state ω_2 Ann believes that Bob chooses L with probability one, and Ann prefers B to M given this belief. Now consider Ann's deviation to T. The closest state to ω_1 in which Ann chooses T is ω_3. At state ω_3 Ann believes that Bob chooses R with probability $\frac{3}{4}$, and Ann prefers B to T given this belief. Hence, Ann is W-rational at ω_1. I now want to show that Bob is W-rational at ω_1. Bob's choice at ω_1 is L. Consider Bob's deviation to R. The closest state to ω_1 in which Bob chooses R is ω_6. At state ω_6 Bob believes that Ann chooses B with probability one, and Bob prefers L to R given this belief. Hence, Bob is W-rational at ω_1. It follows that W-rationality is common knowledge at ω_1, yet the strategies chosen at ω_1 are not rationalizable for this game.

Remark 1 *Notice that this does not contradict Theorem 1 since at state ω_1 neither player is A-rational, and therefore A-rationality is not common knowledge at ω_1. Notice also that the extended model is consistent with the content of Theorem 1 as well: at state ω_5 A-rationality is common knowledge and the strategies chosen at ω_5 are rationalizable. Moreover, since ω_5 is the only state where rationalizable strategies are played, A-rationality is not common knowledge at any state other than ω_5.*

For some, W-rationality may be too large a departure from A-rationality in the sense that $\mathbf{s}_i(\omega)$ may only be rational against deviations $s_i \neq \mathbf{s}_i(\omega)$ for a belief that depends on the deviation being considered. In other words: a player may be W-rational at ω yet $\mathbf{s}_i(\omega)$ need not be a best response to any belief about the opponent. For example, Ann was W-rational at ω_1 in the game discusses above yet her choice of strategy is strictly dominated by the mixed strategy $\frac{1}{2}T + \frac{1}{2}M$ and therefore $\mathbf{s}_{Ann}(\omega_1)$ is not a best response to any belief about Bob's choice.

For this reason I consider a stronger notion of rationality; one in which the same belief p_i rationalizes a strategy $\mathbf{s}_i(\omega)$ against *any* deviation, that is: (a) player i has belief p_i at all the closest worlds to ω in which deviations from $\mathbf{s}_i(\omega)$ actually occur, and (b) player i is W-rational at ω. This is the notion of S-rationality.

One says that player i is *S-rational* at ω if

1. for every $s_i, s_i' \neq \mathbf{s}_i(\omega)$, $\text{marg}_{S_j}\mathbf{p}_i(\cdot; f_i(\omega, s_i)) = \text{marg}_{S_j}\mathbf{p}_i(\cdot; f_i(\omega, s_i'))$, and

2. player i is W-rational at ω.

The notion of S-rationality takes a minimal departure from A-rationality in the sense that both involve the same standard notion of best-responding to beliefs. They differ on their treament of the player's beliefs but only given the occurrence of events that are considered impossible by the players. Both definitions are therefore strongly consistent with traditional subjective expected utility theories. This strong consistency can be captured formally.

Lemma 3 A-RAT $\subseteq U$. *Similarly,* S-RAT $\subseteq U$ *in any extended model in which the selection functions satisfy F1-F2. Moreover, for every normal form game G there is a model such that $U \subseteq$* A-RAT *and an extended model in which the selection functions satisfy F1-F2 such that $U \subseteq$* S-RAT.

Proof. See the Appendix. ■

This lemma shows that without any assumptions on what the players believe both definitions of rationality place equivalent restrictions on outcomes: those restrictions that arise from one round of deletion of strictly dominated strategies. To illustrate consider the Prisoner's Dilemma in Figure 1. Because *S-RAT*$\subseteq U$, a S-rational player will never play "cooperate" in this game. Another example of the restrictions imposed by S-rationality arises in the game depicted in Figure 2. We saw before that B is strictly dominated by the mixed strategy $\frac{1}{2}T + \frac{1}{2}M$. As a consequence, Ann can never be S-rational at any state where she chooses B. This is so despite the fact that Ann can be W-rational at a state where she chooses B. The model of this game used in the proof of Theorem 2 is an example of this: at state ω_1 Ann is W-rational and she chooses B.

Despite this equivalence between A-rationality and S-rationality, it is important to notice that, just as with W-rationality, common knowledge of S-rationality does not lead to rationalizability.

Example 3 *Consider the game described in Figure 3, which is a variant of a game discussed in Basu [4].*

		Bob		
		L	M	H
	L	$3,3$	$10,2$	$10,1$
Ann	M	$2,10$	$5,5$	$11,4$
	H	$1,10$	$4,11$	$10,10$

(3)

Figure 3. *A 'Traveler's Dilemma'*

Notice that the unique rationalizable strategy profile for this game is LL.

Consider the following extended model $(\Omega, \mathcal{K}_{Ann}, \mathcal{K}_{Bob}, \mathbf{s}, f_{Ann}, f_{Bob})$ of this game, where

- $\Omega = \{\omega_1, \ldots, \omega_9\}$,
- $\mathcal{K}_{Ann}(\omega_l) = \{\omega_l\}$ for $l \neq 2,3,6,9$; $\mathcal{K}_{Ann}(\omega_l) = \{\omega_2, \omega_3\}$ for $l = 2,3$; $\mathcal{K}_{Ann}(\omega_l) = \{\omega_6, \omega_9\}$ for $l = 6,9$;
- $\mathcal{K}_{Bob}(\omega_l) = \{\omega_l\}$ for $l \neq 4,7,8,9$; $\mathcal{K}_{Bob}(\omega_l) = \{\omega_4, \omega_7\}$ for $l = 4,7$; $\mathcal{K}_{Bob}(\omega_l) = \{\omega_8, \omega_9\}$ for $l = 8,9$;
- $\mathbf{p}_{Ann}(\cdot \mid \omega_l) = 1_{\omega_l}$ for $l \neq 2,3,6,9$; $\mathbf{p}_{Ann}(\cdot \mid \omega_l) = \frac{1}{10}\omega_2 + \frac{9}{10}\omega_3$ for $l = 2,3$; $\mathbf{p}_{Ann}(\cdot \mid \omega_l) = \frac{1}{10}\omega_6 + \frac{9}{10}\omega_9$ for $l = 6,9$;
- $\mathbf{p}_{Bob}(\cdot \mid \omega_l) = 1_{\omega_l}$ for $l \neq 4,7,8,9$; $\mathbf{p}_{Ann}(\cdot \mid \omega_l) = \frac{1}{10}\omega_4 + \frac{9}{10}\omega_7$ for $l = 4,7$; $\mathbf{p}_{Bob}(\cdot \mid \omega_l) = \frac{1}{10}\omega_8 + \frac{9}{10}\omega_9$ for $l = 8,9$;
- $\mathbf{s}(\omega_l) = s^l$ for $l = 1, \ldots, 9$, where s^l is the $l-th$ element in the following enumeration of S: $\{MM, LM, LH, ML, LL, HM, HL, MH, HH\}$.
- f_{Ann} and f_{Bob} are the unique selection functions satisfying F1-F2.

In this extended model of the game in Figure 3 the selection functions satisfy F1-F2 and yet $CK(S\text{-}RAT) \nsubseteq R$.

To see this I wish to show that in this extended model Ann is S-rational at ω_1. Ann's choice at ω_1 is M. Notice that M is a best response for Ann if she believes that Bob chooses H with probability $\frac{9}{10}$ and M with probability $\frac{1}{10}$. I now want to show that this belief is held at every closest state to ω_1 where Ann deviates. Consider the closest state to ω_1 where Ann deviates from M and chooses L. Such state is ω_2. At state ω_2 Bob chooses M, but

Ann cannot distinguish between ω_2 and ω_3. Moreover, Ann believes that ω_2 occurs with probability $\frac{1}{10}$ and ω_3 with probability $\frac{9}{10}$. Therefore, Ann's belief about Bob's choice at ω_2 is $\frac{1}{10}M + \frac{9}{10}H$. Now consider the closest state to ω_1 where Ann deviates from M and chooses H. Such state is ω_6. At state ω_6 Bob chooses M, but Ann cannot distinguish between ω_6 and ω_9 and Bob chooses H at ω_9. Moreover, Ann believes that ω_6 occurs with probability $\frac{1}{10}$ and ω_9 with probability $\frac{9}{10}$. Therefore, Ann's belief about Bob's choice at ω_6 is $\frac{1}{10}M + \frac{9}{10}H$. Hence, Ann is S-rational at ω_1. The argument that shows that Bob is S-rational at ω_1 is identical. It follows that S-rationality is common knowledge at ω_1, yet the strategies chosen at ω_1 are not rationalizable for this game.

Remark 2 *The remark to Example 2 applies to this result without change.*

Remark 3 *Notice that the game in Figure 3 can be obtained from the Prisoner's Dilemma by adding for each player a strictly dominated strategy, H. Cooperation (that is, strategy profile MM which occurs at state ω_1) in the Prisoner's Dilemma can be a consequence of common knowledge of S-rationality in the extended game by making M a best response to the added strategy, and having beliefs at the closest states to ω_1 where the players deviate to be that the other player chooses H with high probability. Such beliefs are inconsistent with the strategies actually chosen by the players, but that is just as in the definition of rationalizability.*

The theorem above shows that common knowledge of S-rationality does not lead to rationalizability in normal form games. A question that naturally arises is: what extra restrictions does common knowledge of S-rationality impose on the set of outcomes relative to those imposed by S-rationality alone? The answer is: without further restrictions on the players beliefs, none. The formal statement is

Lemma 4 $CK(\text{S-RAT}) \subseteq U$ *in any extended model in which the selection functions satisfy F1-F2. Moreover, for every game G there is an extended model in which the selection functions satisfy F1-F2 such that $U \subseteq CK(\text{S-RAT})$.*

Proof. See the Appendix. ■

2.3 Beliefs "off the equilibrium path"

We know that there is a sharp difference between A-rationality and W-rationality but, how to characterize the difference between A-rationality and S-rationality? An examination of both definitions reveals that they differ exactly in which kinds of beliefs the players are supposed to have at certain states that are considered impossible by the players. In particular, A-rationality does not allow the players to change their beliefs when doing hypothetical reasoning whereas S-rationality does. This suggests a condition on selection functions that guarantees that beliefs do not change at all when doing hypothetical reasoning in the extended model.

One says that player i is *B-rational* at ω if

1. for every strategy $s_i \in S_i$,$\text{marg}_{S_j}\mathbf{p}_i(\cdot;\omega) = \text{marg}_{S_j}\mathbf{p}_i(\cdot; f_i(\omega, s_i))$, and
2. player i is W-rational at ω.

The notion of B-rationality requires that player i holds the same beliefs at world ω as he does at the closest worlds to ω in which he deviates. It is a restriction on beliefs "off the equilibrium path," (that is, given deviations from the strategies prescribed at a given state) in the sense that, at states that cannot occur from the point of view of the players, beliefs cannot be arbitrary: they are restricted by what cannot occur at the worlds believed to be possible by the players. It is the same kind of restriction on beliefs that underlies refinements of Nash equilibrium such as subgame perfection in extensive games. As with A-rationality and S-rationality, one can show that B-rationality places equivalent restrictions on outcomes as A-rationality: those imposed by one round of deletion of strictly dominated strategies. Let *B-RAT* consist of all the states where all the players are rational. The formal statement is

Lemma 5 B-RAT $\subseteq U$ *in any extended model in which the selection functions satisfy F1-F2. Moreover, for every normal form game G there is an extended model in which the selection functions satisfy F1-F2 such that* $U \subseteq$ B-RAT.

Proof. The proof of Lemma 3 applies here if one replaces in the proof the terms *S-RAT* and S-rationality with *B-RAT* and B-rationality, respectively. ■

Despite this equivalence, the first condition in the definition of B-rationality captures the key difference between A-rationality and S-rationality in that for

every extended model of G in which the selection functions satisfy F1-F2 we have that common knowledge of B-rationality *does* lead to rationalizability. The formal statement is

Theorem 6 *CK(B-RAT) $\subseteq R$ in any extended model in which the selection functions satisfy F1-F2.*

Proof. The proof follows the one used by Osborne and Rubinstein [16, p. 80] suitably modified to incorporate the role of the selection functions into the analysis. Pick $\omega \in CK(B\text{-}RAT)$. Then there is a self-evident event F such that $\omega \in F \subseteq B\text{-}RAT$. Let $Z_i = \{s_i(\omega') \in S_i : \omega' \in F\}$ for $i = 1, 2$. Note that $\mathbf{s}_i(\omega) \in Z_i$ because $\omega \in F$ and that for each $\omega' \in F$ there is a belief p_i that arises at every closest state to ω' where player i deviates from $\mathbf{s}_i(\omega')$ such that $\mathbf{s}_i(\omega')$ is a best response to p_i. Such belief is given by $\text{marg}_{S_j}\mathbf{p}_i(\cdot; f_i(\omega', s_i))$.

It remains to be shown that the support of p_i is contained in Z_j. By the first condition in the definition of B-rationality, $\text{marg}_{S_j}\mathbf{p}_i(\cdot;\omega') = p_i$. Recall that, by definition, the support of $\mathbf{p}_i(\cdot;\omega')$ is a subset of $\mathcal{K}_i(\omega')$. Moreover, since F is self-evident, we have $\mathcal{K}_i(\omega') \subseteq F$ and therefore the support of p_i is contained in Z_j. ■

The final result is that common knowledge of B-rationality is indeed possible in an extended model in which the selection functions satisfies F1-F2. The formal statement is

Theorem 7 *For every normal form game G there is an extended model in which the selection functions satisfy F1-F2 such that $R \subseteq CK$(B-RAT).*

Proof. Fix a game G. Let $Z_1 \times Z_2$ be the set of rationalizable strategy profiles. For every $i = 1, 2$ and every rationalizable strategy s_i let $p_i(\cdot; s_i)$ be a selection among the probability distributions on Z_j such that s_i is a best response to $p_i(\cdot; s_i)$. Let the extended model $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{s}, \mathbf{p}_1, \mathbf{p}_2, f_1, f_2)$ be defined by:

- $\Omega = \{\omega_s^t : s, t \in S\}$ and, for every player $i = 1, 2$
- $\mathcal{K}_i(\omega) = \{\omega' \in R : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$ if $\omega \in R$ and $\mathcal{K}_i(\omega) = \{\omega' \in \Omega : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$ otherwise;
- $\text{marg}_{S_j}\mathbf{p}_i(\cdot;\omega) = p_i(\cdot; \mathbf{s}_i(\omega'))$ if $\omega \in R$ or $\omega = f_i(\omega', \mathbf{s}_i(\omega))$ for some $\omega' \in R$ and $\mathbf{p}_i(\cdot;\omega) = 1_\omega$ otherwise;

- $\mathbf{s}\left(\omega_s^t\right) = s$;
- $f_1\left(\omega_s^t, s_1\right) = \omega^s_{s_1, \mathbf{s}_2(\omega_s^t)}$, and
- $f_2\left(\omega_s^t, s_2\right) = \omega^s_{\mathbf{s}_1(\omega_s^t), s_2}$.

Functions f_1 and f_2 are defined as in the proof to Lemma 2 and therefore satisfy F1-F2. To see that $R \subseteq CK(B\text{-}RAT)$ pick $\omega \in R$. This means that $\mathbf{s}_i(\omega)$ is a best response to a belief $p_i(\cdot; \mathbf{s}_i(\omega))$ whose support is contained in Z_j. I now want to show that $\text{marg}_{S_j}\mathbf{p}_i(\cdot; f_i(\omega, s_i)) = p_i(\cdot; \mathbf{s}_i(\omega)) = \text{marg}_{S_j}\mathbf{p}_i(\cdot; \omega)$ for every $s_i \in S_i$. This will show that player i is B-rational at ω.

If $\omega \in R$ then $\text{marg}_{S_j}\mathbf{p}_i(\cdot; \omega) = p_i(\cdot; \mathbf{s}_i(\omega))$. But for every $s_i \in S_i$ there is a state ω' with $\mathbf{s}_i(\omega') = s_i$ such that $\omega' = f_i(\omega, s_i)$ and hence $\text{marg}_{S_j}\mathbf{p}_i(\cdot; f_i(\omega, s_i)) = p_i(\cdot; \mathbf{s}_i(\omega))$. Hence, player i is B-rational at ω. This shows that $R \subseteq B\text{-}RAT$ in this extended model. Now notice that by the construction of $\mathcal{K}_i$, if $\omega \in R$ then $\mathcal{K}_i(\omega) \subseteq R$ for $i = 1, 2$ and therefore R is a self-evident event. This means that if $\omega \in R$ then B-rationality is common knowledge at ω, that is, $\omega \in CK(B\text{-}RAT)$. ■

3 Discussion

a. The Literature. All the literature on the subject of this paper (Bernheim [5], Pearce [17], Brandenburger and Dekel [10] and Tan and Werlang [21]) takes the position that common knowledge of rationality leads to rationalizability in normal form games.[6] The point of view that the present paper presents goes against this conventional wisdom by showing that one can come up with notions of rationality that are very much like the traditional notion (in the sense that they place equivalent restrictions on outcomes as the traditional notion) such that common knowledge of rationality does not lead to rationalizability.

Nevertheless, there is a very clear, compelling intuition that relates common knowledge of rationality and rationalizability, and I make it clear how this intuition relates to the notions of rationality that I develop. My aim with this is not to criticize or disprove any of the previous work on the literature but to simply point out in a precise manner the strong belief restrictions

[6] Although there are quite natural conditions, not involving common knowledge of anything, that lead to rationalizability. See Zambrano [22].

underlying the apparently weak notion of rationalizability. In this sense this paper shares the main goal of the paper by Brandenburger and Dekel [10].

The present paper owes a great intellectual debt to the work in Halpern [13] and Basu [4]. Halpern [13] first used Stalnaker's approach to counterfactual reasoning to substantiate Stalnaker's informal argument that common knowledge of rationality need not lead to backward induction. The methods used in the present paper closely follow those in Halpern [13]. The difference is that here they are used to evaluate the epistemic justification of a solution concept for normal form games while he focuses on perfect information games. *The difference is important because the position in the literature has been that the paradoxes that counterfactuals create in games are inherently an extensive form phenomenon.* An exception to this point of view is presented in Basu [4] who shows by example the problematic nature of rationalizability in relation to the problematic nature of backward induction. Those papers are direct precursors of the present work.

It is now understood that an adequate treatment of counterfactuals is key to understanding the paradoxes that arise in extensive form games. In this paper I argue that the situation is no different in normal form games, and that the treatment of counterfactuals pioneered by Lewis [14] and Stalnaker [18] in the philosophical literature can be very useful in undertanding the interplay between knowledge, rationality, and rationalizability in normal form games.

To place this point of view in the context of the existing literature note that, as far as the normal form is concerned, Dekel and Gul [11, p. 123] acknowledged the counterfactual nature of choice in this type of games but asserted that no elaborate theory of nearby states was necessary to deal with its subtleties. Binmore [6, pp. 220-225], in turn, argues for a revision of the traditional view about counterfactuals that game theorists have, even in normal form games, but argues against using the "closest" state approach by saying that it is not clear what "closest" ought to mean in the game theoretic context. Stalnaker [19] has an explicit treatment of counterfactual worlds in terms of selection functions, as in the present paper. His work differs from mine in that in his work beliefs in the counterfactual possible state $f_i(\omega, s_i)$ for player i must be identical to those that the player holds at state ω. It is precisely the relaxation of this assumption that which allows me to distinguish between common knowledge of A-rationality and common knowledge of S-rationality despite the fact that these two definitions of rationality, *per se,* are equivalent from a purely decision theoretic standpoint.

In another direction, a number of papers have been devoted to provid-

ing the decision theoretic foundations of other solution concepts such as Nash equilibrium (Aumann and Brandenburger [3]) and correlated equilibrium (Aumann [1]). The techniques developed in the present paper can be used to investigate what exactly is being implicitly assumed in those papers regarding "off-path" belief restrictions by allowing an explicit treatment of counterfactuals in the analysis.

b. Counterfactuals. Key to the results presented above is that, because the player knows his own strategy, and a state encodes the strategies chosen by the players at that state, it is not clear what the player's beliefs should be if he *were* to deviate, because there is no state that the player considers possible where he *does* deviate. A statement like *what would the player believe if he were to do what he actually will not do* is a counterfactual. The reader is warned not to treat counterfactuals of this sort lightly, for "one really cannot discuss rationality, or indeed decision making, without substantive conditionals and counterfactuals. Making a decision means choosing among alternatives. Thus one *must* consider hypothetical situations −what would happen if one did something different from what one actually does."[7]

Furthermore, such hypothetical situations are not easy to interpret and understand. Consider, for example, the counterfactual discussed in Aumann [2, p.15]: If Hitler had crossed the channel after Dunkirk, he would have won the war. Such statements are problematic because "If Hitler had crossed the channel, the world would have been different in a myriad of ways. To assign meaning to such a conditional, one must be more specific about the hypothetical world created by the crossing. That is a nontrivial task, even in principle."[8] For more on this see the excellent treatment of the role of counterfactual reasoning in games found in Stalnaker [19] and [20].

c. Rationality. Rationality in this paper is either defined as some form of not playing dominated strategies (as in W-rationality) or as best responding to some beliefs (as in S-rationality and B-rationality) and key to the differences between the notions is how beliefs for a given player differ at state ω and at the states closest to ω where he deviates. It is important at this point to recall the *descriptive* nature of epistemic exercises such as the one carried out in the present paper. The epistemic view addresses "not *why* the players do what they do, not what *should* they do; just what *do* they do, what *do* they believe (...)Not that i does a *because* he believes b; simply that he does

[7] Aumann [2, p. 15]
[8] Aumann [2, p. 15]

a, and believes b."[3, pp. 1174-1175] Consistent with this point of view, the intended interpretation in the models developed above of a statement like "player i believes b were he to deviate from a" is not that i believes b *because* he deviates from a; *simply that he deviates from a, and believes b.* For more on the interpretation of epistemic exercises see below, as well as Aumann [1], Stalnaker [20] and Aumann and Brandenburger [3].

d. The Analyst vs. The Player. Researchers in game theory usually operate under what Myerson [15] calls the *intelligence assumption.* In Myerson's words: "if we [analysts] develop a theory that describes the behavior of intelligent players in some game and we believe that this theory is correct, then we must assume that each player in the game will also understand this theory and its predictions."[9] Wittingly or not, the assumption motivates much of what is done in modern epistemic research. Clear evidence of this can be found in Brandenburger [9]: "Unless we want to accord the [analyst] a 'privileged' position that is somehow denied to the players, it is only natural to ask what happens if a player can think about the game the same way."[10]

Were there no conflation between the levels at which the analyst and the players operate when reasoning about the game, the results presented in this paper would not be of much relevance. When the epistemic model is "primarily a convenient framework to enable us-the analysts-to discuss the things we want to discuss [about] actions, payoffs, beliefs, rationality, equilibrium, and so on,"[11] there are no counterfactuals in the normal form, and rationality must be defined as it has always been (that is, as A-rationality in this paper). But when we adopt the *intelligence assumption* and think of the players as being able to "talk about the things that [the analysts] want to talk about,"[12] issues regarding counterfactuals become important and should be dealt with explicitly in our theoretical exercises.

So we reach an interesting point in the discussion: one where the relevance of the results presented in this paper depends on subtle aspects of interpretation regarding the level (that of the analyst vs. that of the player) at which the epistemic assumptions are supposed to be sound. I do not wish to be judge on this matter. I prefer to leave it to the passage of time, once people have had the chance to read and absorb the ideas in this paper, for a conclusion regarding these matters of interpretation to be reached. I

[9] Myerson [15, p. 4].
[10] Brandenburger [9, p. 22].
[11] Aumann and Brandenburger [3, p. 1175]
[12] Aumann and Brandenburger [3, p. 1175].

hope this paper will contribute in one way or another to the settling of this important conceptual matter.[13]

4 Conclusion

The purpose of this paper is to point out that the epistemic justification of solution concepts in normal form games can be problematic for reasons that are identical to those that complicate the justification of backward induction in extensive games. The key to the problem is that to justify the rationality of a player at a given state of the state one has to consider *what would the player believe if he were to do what he actually will not do.*

The traditional manner in which this situation is handled is to define beliefs given a deviation to be as if the player weren't deviating. Common knowledge of rationality given this definition of beliefs leads to rationalizability. In this paper I present an alternative approach inspired on the method employed by Halpern [13] to capture counterfactuals in a perfect information game. According to this method one evaluates a statement like "what would the player believe (at a given state) if he were to reach a node which he knows that will never be reached?" at the state that is *closest* to the actual state *in which the node is actually reached.* I define beliefs given a deviation in this manner (i.e., at the state that is closest to the actual state *in which the deviation actually occurs)* and then show that common knowledge of rationality given beliefs defined in this way need not lead to rationalizability.

This shows that the traditional model makes assumptions about the treatment of counterfactuals that are critical in generating the connections between rationality, knowledge and rationalizability. In this paper I also provide a formal statement of what those assumptions are and they reveal that to justify the notion of rationalizability in epistemic models one needs to make assumptions about beliefs *off the equilibrium path*, (that is, given deviations from the strategies prescribed at a given state) that are very similar to those assumptions underlying refinements of Nash equilibrium such as subgame perfection in extensive games. This implies that it may be misleading to believe that, from an epistemic point of view, rationalizability relies on weaker assumptions about belief consistency than Nash equilibrium.

[13] I am very grateful to an anonymous referee for his valuable insights on these matters of interpretation.

References

[1] Aumann, R., "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica 55*, 1-18, 1987.

[2] _____, "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior 8*, 6-19, 1995.

[3] Aumann, R., and A. Brandenburger, "Epistemic Conditions for Nash Equilibrium," *Econometrica 63*, 1161-1180, 1995.

[4] Basu, K., "The Traveler's Dilemma: Paradoxes of Rationality in Game Theory," *American Economic Review 84*, 391-395, 1994.

[5] Bernheim, B. D., "Rationalizable Strategic Behavior," *Econometrica 52*, 107-1028, 1984.

[6] Binmore, K., *Game Theory and the Social Contract Volume I: Playing Fair*, Cambridge: MIT Press, 1994.

[7] Binmore, K. and A. Brandenburger, "Common Knowledge and Game Theory," in K. Binmore (ed.), *Essays on the Foundations of Game Theory*, Oxford: Blackwell, 1990.

[8] Battigalli, P. and G. Bonanno, "Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory," *Research in Economics 53*, 149-225, 1999.

[9] Brandenburger, A., "The Power of Paradox: Some Recent Developments in Interactive Epistemology," *manuscript*, Harvard Business School, 2001.

[10] Brandenburger, A. and E. Dekel, "Rationalizability and Correlated Equilibria," *Econometrica 55*, 1391-1402, 1987.

[11] Dekel, E. and F. Gul, "Rationality and Knowledge in Game Theory," in D. Kreps and K. Wallis (eds.), *Advances in Economics and Econometrics: Theory and Applications, Volume 1*, Cambridge: Cambridge University Press, 1997.

[12] Geanakoplos, J., "Common Knowledge," in R. Aumann and S. Hart (eds.), *Handbook of Game Theory, Volume 2*, Amsterdam: North-Holland, 1993.

[13] Halpern, J., "Substantive Rationality and Backward Induction," *Games and Economic Behavior 37*, 425-435, 2001.

[14] Lewis, D., *Counterfactuals*, Cambridge: Harvard University Press, 1973.

[15] Myerson, R., *Game Theory: Analysis of Conflict*, Cambridge: Harvard University Press, 1991.

[16] Osborne, M. and A. Rubinstein, *A Course in Game Theory*, Cambridge: MIT Press, 1994.

[17] Pearce, D., "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica 52*, 1029-1050, 1984.

[18] Stalnaker, R., "A Theory of Conditionals," in N. Rescher (ed.), *Studies in Logical Theory, Number 2*, Oxford: Blackwell, 1968.

[19] _____, "Knowledge, Belief and Counterfactual Reasoning in Games," *Economics and Philosophy 12*, 133-163, 1996.

[20] _____, "Belief Revision in Games: Forward and Backward Induction," *Mathematical Social Sciences 36*, 31-56, 1998.

[21] Tan, T. and S. Werlang, "The Bayesian Foundations of Solution Concepts of Games," *Journal of Economic Theory 45*, 370-391, 1988.

[22] Zambrano, E., "Epistemic Conditions for Rationalizability," *manuscript*, University of Notre Dame, 2001.

5 Appendix:

5.1 Proof of Lemma 2

Assume $\omega \notin U^P$. Then there is $s_i \neq \mathbf{s}_i(\omega)$ such that, for all $s_j \in S_j$, $U_i(s_i, s_j) > U_i(\mathbf{s}_i(\omega), s_j)$. This in fact implies that for all $\omega' \in \Omega$, $U_i(s_i, \mathbf{s}_j(\omega')) > U_i(\mathbf{s}_i(\omega), \mathbf{s}_j(\omega'))$,and, therefore, that for all $\omega' \in \Omega$,

$$\mathbf{p}_i(\omega'; f(\omega, s_i))\, U_i(s_i, \mathbf{s}_j(\omega')) > \mathbf{p}_i(\omega'; f(\omega, s_i))\, U_i(\mathbf{s}_i(\omega), \mathbf{s}_j(\omega')).$$

Summing across all $\omega' \in \Omega$ we get that there is $s_i \neq \mathbf{s}_i(\omega)$ such that

$$\sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; f(\omega, s_i)) U_i(s_i, s_j) > \sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; f(\omega, s_i)) U_i(\mathbf{s}_i(\omega), s_j),$$

and hence player i is not W-rational at ω.

Now for the second part. Pick a normal form game G. For every $i = 1, 2$ and every strategy s_i not dominated by a mixed strategy let $p_i(\cdot; s_i)$ be a selection among the probability distributions on S_j such that s_i is a best response to $p_i(\cdot; s_i)$.

Consider the the extended model $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{s}, f_1, f_2)$ where

- $\Omega = \{\omega_s^t : s, t \in S\}$ and, for every player $i = 1, 2$
- $\mathcal{K}_i(\omega) = \{\omega' \in \Omega : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$,
- $\text{marg}_{S_j} \mathbf{p}_i(\cdot; \omega) = p_i(\cdot; \mathbf{s}_i(\omega'))$ if $\omega \in U$ or $\omega = f_i(\omega', \mathbf{s}_i(\omega))$ for some $\omega' \in U$ and $\mathbf{p}_i(\cdot; \omega) = 1_\omega$ otherwise.
- $\mathbf{s}(\omega_s^t) = s$,
- $f_1(\omega_s^t, s_1) = \omega^s_{s_1, \mathbf{s}_2(\omega_s^t)}$, and
- $f_2(\omega_s^t, s_2) = \omega^s_{\mathbf{s}_1(\omega_s^t), s_2}$.

It is not hard to see that the functions f_1 and f_2 satisfy F1-F2. I will show this for player 1. The argument for player 2 is identical. Notice that $\mathbf{s}_1(f_1(\omega_s^t, s_1)) = \mathbf{s}_1\left(\omega^s_{s_1, \mathbf{s}_2(\omega_s^t)}\right) = s_1$, so property F1 holds. Now notice that for any $s_1 \in S_1$, $\mathbf{s}_2(f_1(\omega_s^t, s_1)) = \mathbf{s}_2\left(\omega^s_{s_1, \mathbf{s}_2(\omega_s^t)}\right) = \mathbf{s}_2(\omega)$ and property F2 holds.

Now for the final part. Pick $\omega \in U$. Then $\mathbf{s}_i(\omega)$ is not strictly dominated. By Lemma 60.1 in Osborne and Rubinstein [16, p. 60] this means that $\mathbf{s}_i(\omega)$ is a best response to $p_i(\cdot; \mathbf{s}_i(\omega))$. Then $\text{marg}_{S_j} \mathbf{p}_i(\cdot; f_i(\omega, s_i)) = p_i(\cdot; \mathbf{s}_i(\omega))$ for every $s_i \in S_i$, by the definition of $\mathbf{p}_i$. There is, therefore, no $s_i \neq \mathbf{s}_i(\omega)$ such that

$$\sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; f_i(\omega, s_i)) U_i(s_i, \mathbf{s}_j(\omega')) > \sum_{\omega' \in \Omega} \mathbf{p}_i(\omega'; f_i(\omega, s_i)) U_i(\mathbf{s}_i(\omega), \mathbf{s}_j(\omega')).$$

Hence, player i is W-rational at ω.

5.2 Proof of Lemma 3

Let $\omega \in$*A-RAT (*resp. $\omega \in$*S-RAT).* Then there is a belief $p_i = \mathbf{p}_i(\cdot;\omega)$ (resp. $p_i = \mathbf{p}_i(\cdot; f_i(\omega, s_i))$ such that, for all $s_i \neq \mathbf{s}_i(\omega)$,

$$\sum_{\omega' \in \Omega} p_i(\omega') U_i(\mathbf{s}_i(\omega), \mathbf{s}_j(\omega')) \geq \sum_{\omega' \in \Omega} p_i(\omega') U_i(s_i, \mathbf{s}_j(\omega')).$$

This means that strategy $\mathbf{s}_i(\omega)$ is a best response to the belief p_i. It is therefore not a *never best response* and, by Lemma 60.1 in Osborne and Rubinstein [16, p. 60], it is not strictly dominated. Since this is true for every player $i = 1, 2$, then $\omega \in U$.

For the second part pick a normal form game G and let $p_i(\cdot; s_i)$ and the model $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{s})$ be defined as in the proof to Lemma 2:

- $\Omega = \{\omega_s^t : s, t \in S\}$ and, for every player $i = 1, 2$
- $\mathcal{K}_i(\omega) = \{\omega' \in \Omega : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$,
- $\mathbf{p}_i(\cdot;\omega) = p_i(\cdot; \mathbf{s}_i(\omega'))$ if $\omega \in U$ or $\omega = f_i(\omega', \mathbf{s}_i(\omega))$ for some $\omega' \in U$ and $\mathbf{p}_i(\cdot;\omega) = 1_\omega$ otherwise.
- $\mathbf{s}(\omega_s^t) = s$,

Also, consider the extended model where $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{s})$ is as before and the selection functions are defined by

- $f_1(\omega_s^t, s_1) = \omega^s_{s_1, \mathbf{s}_2(\omega_s^t)}$, and
- $f_2(\omega_s^t, s_2) = \omega^s_{\mathbf{s}_1(\omega_s^t), s_2}$.

Functions f_1 and f_2 are defined as in the proof to Lemma 2 and therefore satisfy F1-F2. Now for the final part. Pick $\omega \in U$. By Lemma 2, $\omega \in$*W-RAT* and, as in the proof of Lemma 2, $\mathbf{s}_i(\omega)$ is a best response to $\mathbf{p}_i(\cdot;\omega) = p_i(\cdot; \mathbf{s}_i(\omega))$ and therefore $\omega \in$*A-RAT*. Because $\text{marg}_{S_j} \mathbf{p}_i(\cdot; f_i(\omega, s_i)) = p_i(\cdot; \mathbf{s}_i(\omega))$ for every $s_i \in S_i$, and since $\omega \in$*W-RAT,* this means that the definition of S-rationality is also satisfied at ω. Hence, $\omega \in$*S-RAT*.

5.3 Proof of Lemma 4

Lemma 3 shows that $S\text{-}RAT \subseteq U$ in any extended model in which the selection functions satisfy F1-F2. Therefore, since $CK(S\text{-}RAT) \subseteq S\text{-}RAT$, it follows that $CK(S\text{-}RAT) \subseteq U$.

For the second part fix a game G and let $p_i(\cdot; s_i)$ be as in the proof to Lemma 2 and the extended model $(\Omega, \mathcal{K}_1, \mathcal{K}_2, \mathbf{s}, \mathbf{p}_1, \mathbf{p}_2, f_1, f_2)$ be defined by :

- $\Omega = \{\omega_s^t : s, t \in S\}$ and, for every player $i = 1, 2$
- $\mathcal{K}_i(\omega) = \{\omega' \in U : \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$ if $\omega \in U$ and $\mathcal{K}_i(\omega) = \{\omega' \in \Omega : \ \mathbf{s}_i(\omega') = \mathbf{s}_i(\omega)\}$ otherwise;
- $\mathbf{p}_i(\cdot; \omega) = p_i(\cdot; \mathbf{s}_i(\omega'))$ if $\omega = f_i(\omega', \mathbf{s}_i(\omega))$ for some $\omega' \in U$ and $\mathbf{p}_i(\cdot; \omega) = 1_\omega$ otherwise;
- $\mathbf{s}(\omega_s^t) = s$;
- $f_1(\omega_s^t, s_1) = \omega^s_{s_1, \mathbf{s}_2(\omega_s^t)}$, and
- $f_2(\omega_s^t, s_2) = \omega^s_{\mathbf{s}_1(\omega_s^t), s_2}$.

Functions f_1 and f_2 are defined as in the proof to Lemma 2 and therefore satisfy F1-F2. To see that $U \subseteq CK(S\text{-}RAT)$ pick $\omega \in U$. This means that $\mathbf{s}_i(\omega)$ is a best response to a belief $p_i(\cdot; \mathbf{s}_i(\omega))$. I now want to show that $\text{marg}_{S_j} \mathbf{p}_i(\cdot; f_i(\omega, s_i)) = p_i(\cdot; \mathbf{s}_i(\omega))$ for every $s_i \in S_i$. This will show that player i is S-rational at ω.

If $\omega \in R$ then for every $s_i \in S_i$ there is a state ω' with $\mathbf{s}_i(\omega') = s_i$ such that $\omega' = f_i(\omega, s_i)$ and hence $\text{marg}_{S_j} \mathbf{p}_i(\cdot; f_i(\omega, s_i)) = p_i(\cdot; \mathbf{s}_i(\omega))$. Hence, player i is S-rational at ω. This shows that $U \subseteq S\text{-}RAT$ in this extended model. Now notice that by the construction of $\mathcal{K}_i$, if $\omega \in U$ then $\mathcal{K}_i(\omega) \subseteq U$ for $i = 1, 2$ and therefore U is a self-evident event. This means that if $\omega \in U$ then S-rationality is common knowledge at ω, that is, $\omega \in CK(S\text{-}RAT)$.