

# Common Knowledge of Rationality is Self-Contradictory

Herbert Gintis\*

February 25, 2012

## Abstract

The conditions under which rational agents play a Nash equilibrium are extremely demanding and often implausible. Common knowledge of rationality (CKR), by contrast, in many cases implies agents play a Nash equilibrium. Game theorists routinely assume CKR when it is convenient to do so, considering CKR to be an unproblematic strengthening of the rationality assumption. Yet examples show that CKR is not a legitimate epistemic condition. In some cases rationality implies that CKR is false, so CKR is self-contradictory. Thus unless CKR can be justified in a particular situation by a consistent set of more basic epistemic principles, CKR cannot be assumed. The failure of CKR is related to some well-known antinomies of modal logics.

The conditions under which rational agents play a Nash equilibrium are extremely demanding and often implausible (Aumann and Brandenburger 1995). Common knowledge of rationality (CKR), by contrast, in many cases implies agents play a Nash equilibrium. Game theorists routinely assume CKR when it is convenient to do so, considering CKR to be an unproblematic strengthening of the rationality assumption. Yet in some cases rationality implies the absence of common knowledge of rationality, so CKR is self-contradictory. Thus unless CKR in a particular situation can be justified by a consistent set of more basic epistemic principles, CKR cannot be assumed. The failure of CKR, we shall see, is related to some well-known antinomies of modal logics.

## 1 Epistemic Game Theory

An *epistemic game*  $\mathcal{G}$  consists of a normal form game with players  $i = 1, \dots, n$  and a finite pure-strategy set  $S_i$  for each player  $i$ , so  $S = \prod_{i=1}^n S_i$  is the set of

---

\*Santa Fe Institute and Central European University.

pure-strategy profiles for  $\mathcal{G}$ , with payoffs  $\pi_i: S \rightarrow \mathbf{R}$  for each player  $i$ . In addition,  $\mathcal{G}$  includes a set of possible states  $\Omega$  of the game, a knowledge partition  $\mathcal{P}_i$  of  $\Omega$  for each player  $i$ , and a subjective prior  $p_i(\cdot; \omega)$  over  $\Omega$  for each player  $i$  that is a function of the current state  $\omega$ . A state  $\omega$  specifies, possibly among other aspects of the game, the strategy profile  $s$  used in the game. We write this  $s = \mathbf{s}(\omega)$ . Similarly, we write  $s_i = \mathbf{s}_i(\omega)$  and  $s_{-i} = \mathbf{s}_{-i}(\omega)$ .

The subjective prior  $p_i(\cdot; \omega)$  represents  $i$ 's beliefs concerning the state of the game, including the choices of the other players, when the actual state is  $\omega$ . Thus,  $p_i(\omega'; \omega)$  is the probability  $i$  places on the current state being  $\omega'$  when the actual state is  $\omega$ . We write the cell of the partition  $\mathcal{P}_i$  containing state  $\omega$  as  $\mathbf{P}_i\omega$ , and we interpret  $\mathbf{P}_i\omega \in \mathcal{P}_i$  as the set of states that  $i$  considers possible (that is, the set of states among which  $i$  cannot distinguish) when the actual state is  $\omega$ . Therefore, we require that  $\mathbf{P}_i\omega = \{\omega' \in \Omega \mid p_i(\omega'; \omega) > 0\}$ . Because  $i$  cannot distinguish among states in the cell  $\mathbf{P}_i\omega$  of his knowledge partition  $\mathcal{P}_i$ , his subjective prior must satisfy  $p_i(\omega''; \omega) = p_i(\omega''; \omega')$  for all  $\omega'' \in \Omega$  and all  $\omega' \in \mathbf{P}_i\omega$ . Moreover, we assume a player believes the actual state is possible, so  $p_i(\omega; \omega) > 0$  for all  $\omega \in \Omega$ .

Since each state  $\omega$  in epistemic game  $\mathcal{G}$  specifies the players' pure strategy choices  $\mathbf{s}(\omega) = (\mathbf{s}_1(\omega), \dots, \mathbf{s}_n(\omega)) \in S$ , the players' subjective priors specify their beliefs  $\phi_1^\omega, \dots, \phi_n^\omega$  concerning the choices of the other players. We call  $\phi_i^\omega$  player  $i$ 's *conjecture* concerning the behavior of the other players in state  $\omega$ . Player  $i$ 's conjecture is derived from  $i$ 's subjective prior. To see this, if  $\xi$  is any truth function on  $\Omega$  we define  $[\xi]$  to be  $\{\omega \in \Omega \mid \xi(\omega)\}$ . Note that  $[s_{-i}] =_{\text{def}} [s_{-i}(\omega) = s_{-i}]$  is an event, so we define  $\phi_i^\omega(s_{-i}) = p_i([s_{-i}]; \omega)$ , where  $[s_{-i}] \subset \Omega$  is the event that the other players choose strategy profile  $s_{-i}$ . Thus, at state  $\omega$ , each player  $i$  takes the action  $\mathbf{s}_i(\omega) \in S_i$  and has the subjective prior probability distribution  $\phi_i^\omega$  over  $S_{-i}$ . A player  $i$  is deemed *Bayesian rational* at  $\omega$  if  $\mathbf{s}_i(\omega)$  maximizes  $\pi_i(s_i, \phi_i^\omega)$ , where

$$\pi_i(s_i, \phi_i^\omega) =_{\text{def}} \sum_{s_{-i} \in S_{-i}} \phi_i^\omega(s_{-i}) \pi_i(s_i, s_{-i}). \quad (1)$$

In other words, player  $i$  is Bayesian rational in epistemic game  $\mathcal{G}$  if his pure-strategy choice  $\mathbf{s}_i(\omega) \in S_i$  for every state  $\omega \in \Omega$  satisfies

$$\pi_i(\mathbf{s}_i(\omega), \phi_i^\omega) \geq \pi_i(s_i, \phi_i^\omega) \quad \text{for } s_i \in S_i. \quad (2)$$

## 2 Rationalizability and CKR

Let  $\mathcal{G}$  be an epistemic game. We denote the set of mixed strategies with support in  $S$  as  $\Delta^*S = \prod_{i=1}^n \Delta S_i$ , where  $\Delta S_i$  is the set of mixed strategies for player  $i$ . We denote the mixed strategy profiles of all  $j \neq i$  by  $\Delta^*S_{-i}$ .

In epistemic game  $\mathcal{G}$ ,  $X_i \subseteq S_i$ , is a *best response set* for player  $i$  if, for each  $x_i \in X_i$ ,  $i$  has a conjecture  $\phi_{-i} \in \Delta X_{-i}$  such that  $x_i$  is a best response to  $\phi_{-i}$ , as defined by (2). Because a union of best response sets is also a best response set, there is a maximal best response set  $X_i^*$ . We define a strategy to be *rationalizable* if it is a member of  $X_* = \prod_{i=1}^n X_i^*$ .

The set of rationalizable strategies in a normal form game is the same as the set of strategies eliminated by the iterated elimination of strongly dominated strategies (Bernheim 1984, Pearce 1984). It is clear that a strongly dominated strategy will be eliminated in the first round of the rationalizability construction if and only if it is eliminated in the first round of the iterated elimination of strongly dominated strategies. This observation can be extended to each successive stage in the construction of rationalizable strategies, which shows that all strategies that survive the iterated elimination of strongly dominated strategies are rationalizable. Are there other strategies that are rationalizable? The answer is that strongly dominated strategies exhaust the set of rationalizable strategies. For details, see Bernheim (1984) or Pearce (1984).

In many games, all rationalizable strategies are Nash equilibria. For instance, in the repeated prisoner's dilemma, which we discuss below, there is a unique rationalizable strategy in which both players defect on the first round. This of course is not all how real-world players behave in a many-stage finite prisoner's dilemma (McKelvey and Palfrey 1992). It is often assumed that this means either players are irrational or they believe their partner is irrational (Kreps et al. 1982). In fact, CKR implies rationalizability, but rationality does not.

We can derive rationalizability assuming CKR, as follows. Let  $s_1, \dots, s_n$  be the strategy profile chosen when  $\phi_1, \dots, \phi_n$  are the players' conjectures. The rationality of player  $i$  requires that  $s_i$  maximize  $i$ 's expected payoff, given  $\phi_i$ . Moreover, because  $i$  knows that  $j$  is rational, he knows that  $s_j$  is a best response, given some probability distribution over  $S_{-j}$ —namely,  $s_j$  is a best response to  $\phi_j$ . We say  $\phi_i$  is *first-order consistent* if  $\phi_i$  places positive probability only on pure strategies of  $j$  that have the property of being best responses, given some probability distribution over  $S_{-j}$ . By the same reasoning, if  $i$  places positive probability on the pair  $s_j, s_k$ , because  $i$  knows that  $j$  knows that  $k$  is rational,  $i$  knows that  $j$ 's conjecture is first-order consistent, and hence  $i$  places positive probability only on pairs  $s_j, s_k$  where  $j$  is first-order consistent and  $j$  places positive probability on  $s_k$ . When this is the case, we say that  $i$ 's conjecture is *second-order consistent*. Clearly, we can define consistency of order  $r$  for all positive integers  $r$ , and a conjecture that is  $r$ -consistent for all  $r$  is simply called *consistent*. We say  $s_1, \dots, s_n$  is rationalizable if there is some consistent set of conjectures  $\phi_1, \dots, \phi_n$  that places positive probability on  $s_1, \dots, s_n$ .

### 3 The Power of CKR

For an example of the power of CKR, consider the following game  $G_n$ , known as the *Traveler's Dilemma* (Basu 1994). Two business executives pay bridge tolls while on a trip but do not have receipts. Their superior tells each of them to report independently an integral number of dollars between 2 and  $n$  on their expense sheets. If they report the same number, each will receive this much back. If they report different numbers, each will get the smaller amount, plus the low reporter will get an additional \$2 (for being honest) and the high reporter will lose \$2 (for trying to cheat).

	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$	$s_9$
$s_2$	2,2	4,0	4,0	4,0	4,0	4,0	4,0	4,0
$s_3$	0,4	3,3	5,1	5,1	5,1	5,1	5,1	5,1
$s_4$	0,4	1,5	4,4	6,2	6,2	6,2	6,2	6,2
$s_5$	0,4	1,5	2,6	5,5	7,3	7,3	7,3	7,3
$s_6$	0,4	1,5	2,6	3,7	6,6	8,4	8,4	8,4
$s_7$	0,4	1,5	2,6	3,7	4,8	7,7	9,5	9,5
$s_8$	0,4	1,5	2,6	3,7	4,8	5,9	8,8	10,6
$s_9$	0,4	1,5	2,6	3,7	4,8	5,9	6,10	9,9

**Figure 1:** The Traveler's Dilemma

Let  $s_k$  be strategy Report  $k$ . Figure 1 illustrates the game  $G_9$ . Note first that  $s_9$  is only weakly dominated by  $s_8$ , but a mixed strategy  $\epsilon s_2 + (1 - \epsilon)s_8$  strongly dominates  $s_9$  whenever  $1/6 > \epsilon > 0$ . When we eliminate  $s_9$  for both players,  $s_7$  only weakly dominates  $s_8$ , but a mixed strategy  $\epsilon s_2 + (1 - \epsilon)s_7$  strongly dominates  $s_8$  for sufficiently small  $\epsilon > 0$ . We thus eliminate  $s_8$ , and continue in the same manner until only  $s_2$  remains. Hence  $(s_2, s_2)$  is the only strategy pair that survives the iterated elimination of strongly dominated strategies. It follows that  $s_2$  is the only rationalizable strategy, and the only Nash equilibrium as well.

It is worth noting that this result is not due to the use of mixed strategies that dominate pure strategies. By perturbing the payoffs a small amount, a similar analysis holds using only pure strategies (Gintis 2009, §4.11).

It is clear that in real life, the two business executives, assuming they place no value on truthful reporting, would not engage in the recursive elimination of dominated strategies, but would rather simply choose  $s_k$  for some large  $k$ . This observation is widely interpreted as implying that sometimes irrationality pays off. For instance, the Science News Web edition of December 5, 2008 bears the headline

“Traveler’s Dilemma: When it’s Smart to be Dumb. Some game theory paradoxes can be resolved by assuming that people...aren’t rational.” Of course, it is not rationality but rather CKR that produces this result. If the business executives are rational, they will recognize the arbitrariness of the recursive elimination of strictly dominated strategies and take recourse to other decision principles.

#### 4 Common Knowledge of Rationality and Subgame Perfection

Consider a finite generic extensive form epistemic game of perfect information  $\mathcal{G}$  (a game is generic if, for each player, no two payoffs at terminal nodes are equal). A pure-strategy profile  $s$  assigns an action  $s^v$  at each nonterminal node  $v$ . Indeed, if  $s_i$  is the pure-strategy profile of player  $i$  and if  $i$  moves at  $v$ , then  $s^v = s_i^v$ . We denote by  $b$  the unique backward induction strategy profile. Thus, if player  $i$  moves at node  $v$ , then

$$\pi_i^v(b) > \pi_i^v(b/a^v) \quad \text{for } a^v \neq b^v, \quad (3)$$

where  $\pi_i^v(s)$  is the payoff of strategy profile  $s$  to player  $i$ , starting from node  $v$  (even if, starting from the beginning of the game,  $v$  would not be reached), and  $s/t^v$  denotes the strategy profile  $s$  for the player who chooses at  $v$ , replacing his action  $s_i^v$  with action  $t$  at  $v$ .

To specify rationality in this framework, suppose players choose pure strategy profile  $\mathbf{s}(\omega)$  in state  $\omega$ . We then say  $i$  is rational if, for every node  $v$  at which  $i$  chooses and for every pure strategy  $t_i \in S_i$ , we have

$$R_i \subseteq \neg \mathbf{K}_i \{ \omega \in \Omega \mid \pi_i^v(\mathbf{s}/t_i) > \pi_i^v(\mathbf{s}) \}; \quad (4)$$

i.e.,  $i$  does not know that there is a better strategy than  $\mathbf{s}_i(\omega)$  at  $v$ . Common knowledge of rationality, which we write as CKR, means  $R_i$  is common knowledge for all players  $i$ . Note that this definition is somewhat weaker than Bayesian rationality, which requires that agents have subjective priors over events and maximize utility subject to these priors.

Let  $I^v \subseteq \Omega$  be the event that  $b^v$  is chosen at node  $v$ . Thus

$$I^v = \{ \omega \in \Omega \mid \mathbf{s}(\omega)^v = b^v \}, \quad (5)$$

so the event  $I$  that the backward induction path is chosen is simply

$$I = \bigcap_v I^v.$$

The assertion that common knowledge of rationality implies backward induction is then simply expressed as (Aumann 1995)

THEOREM 1.  $CKR \subseteq I$ .

This theorem does not claim that rational agents will always play the subgame perfect equilibrium. Rather, it claims that if a player makes a move to a node that is not along the backward induction path of play, then common knowledge of rationality cannot obtain at that node or at any subsequent node of the game tree. There is nothing irrational about a player making such a move, as he may have some notion as to how rational agents will play the game based on considerations other than CKR. An example follows.

	<i>C</i>	<i>D</i>
<i>C</i>	3,3	0,4
<i>D</i>	4,0	1,1

**Figure 2:** Bob and Alice Play the Prisoner's Dilemma

For example, suppose Alice and Bob play the Prisoner's Dilemma, one stage of which is shown in Figure 2, 100 times, with the condition that the first time either player defects, the game terminates. Common sense tells us that players will cooperate for at least 95 rounds, and this is indeed supported by experimental evidence (Andreoni and Miller 1993). However, a backward induction argument indicates that players will defect in the very first round. To see this, note that the players will surely defect in round 100. But then, nothing they do in round 99 can help prolong the game, so they will both defect in round 99. Repeating this argument 99 times, we see that they will both defect on round one. It follows that CKR implies that defection will take place on round one of this game.

Suppose however, that Alice and Bob are rational, have subjective priors concerning each other's play, and maximize their expected payoffs subject to these priors. Specifically, suppose Alice believes that Bob will cooperate up to round  $k$  and then defect, with probability  $g_k$ . Then, Alice will choose a round  $m$  to defect in that maximizes the expression

$$\pi_m = \sum_{i=1}^{m-1} 3(i-1)g_i + (3(m-1)+1)g_m + (3(m-1)+4)(1-G_m), \quad (6)$$

where  $G_m = g_1 + \dots + g_m$ . The first term in this expression represents the payoff if Bob defects first, the second term if both players defect simultaneously, and the

final term if Alice defects first. Maximizing this expression suggests cooperating for many rounds for all plausible probability distributions. For instance, suppose  $g_k$  is uniformly distributed in the rounds  $m = 1, \dots, 99$ . Then it is a best response to cooperate up to round 98. Indeed, suppose Alice expects Bob to defect in round one with probability 0.95 and otherwise defect with equal probability on any round from two to 99. Then it is still optimal for her to defect in round 98. Clearly, the backward induction assumption is not plausible unless Alice believes Bob is highly likely to be an obdurate backward inductor. Few in fact are.

## 5 The Paradoxes of Common Knowledge of Rationality

Consider the following game. Bob places three slips of paper, on each of which is written a whole number between 1 and 1000. The three numbers must be distinct. Alice chooses one of these three slips of paper with equal probability. After looking at the number on her slip, she can either Play or Pass. If she Plays and she has chosen the largest of the three numbers, Bob pays her \$10. If she Passes, she pays Bob \$1. If she Plays and her number is not the highest of the three, she pays Bob \$10000.

Let Bob's Random Strategy be to deploy a publicly observable randomizing device that randomly samples the integers  $1, \dots, 1000$  without replacement. The device gives Alice three folded slips of paper on which are written these three numbers, hidden from view. Alice unfolds one of the three slips, chosen at random, and observes the number written on it. We do not assume that Bob's Random Strategy is an optimal strategy, but it is a strategy available to him. It is easy to show that Alice's best response to Bob's Random Strategy is to Pass unless the number on her slip is 1000. To see this, note that obviously if it is a best response to Pass if her slip shows 999, then it is also a best response to pass if her slip shows any number lower than 999. So let us assume her slip says 999. Alice loses choosing Play only if Bob randomly chose the three numbers  $m, 999, 1000$ , where  $1 \leq m \leq 998$ . Conditional on the fact that Alice chose the 999 slip, the probability that Bob chooses  $m, 999, 1000$  is  $p = 2/999$ . Alice's expected payoff to is then  $-\$10000p + \$10(1 - p) = -\$10.04$ . Thus Alice's optimal response is to Play if her slip shows 1000, and to pass otherwise. The probability that Alice chooses Play is then  $(1 - 3/1000)(1/3) = 0.001$ . Thus the payoff to Bob from using the Random Strategy is  $(0.999) \times \$1 - (0.001) \times \$10 = \$0.989$ .

Assuming CKR, we can show that the payoff to the game for Alice is strictly positive, and since this is a zero sum game, the payoff to Bob is strictly negative. Since Bob can choose the three numbers any way he wishes, we can as if a rational Bob would ever include 1000 in his three numbers. If Bob does, then Alice will

win the \$10 with probability  $1/3$  and lose \$1 with probability  $2/3$ , giving an expected payoff of \$8. Thus Bob's payoff to including 1000 is strictly negative, and including 1000 is dominated by Bob's Random Strategy. Because Alice knows that Bob is rational, she knows he will not include 1000 among his three numbers. But Bob knows that Alice knows he is rational, so if he includes 999 among the three numbers, he knows Alice will know that if she picks 999, she will guess that it is the highest, so including 999 among the three numbers is dominated by the Rational Strategy. Continuing to iterate this argument, he must choose numbers 1, 2, and 3. But then Alice knows this, so if her slip says 3, she will guess correctly that it is the highest number. Thus Bob knows that  $\{1, 2, 3\}$  is dominated by his Random Strategy.

It follows that if we assume CKR in this context, then Bob cannot play his Random Strategy, and hence his strategy choice does not maximize his payoff. But this contradicts Bob's rationality, and hence also CKR. The implication of this reasoning is that CKR is self-contradictory.

Note that I am not saying that CKR is simply a "very strong" condition that will hold only under "ideal" circumstances. Rather, I am asserting that CKR does not describe a condition of knowledge that obtains among individuals. CKR does not fail because agents lack information about other agents, or that they are uncertain about the rationality of other players, but rather because an agent can rationally act in a manner that contradicts CKR.

## 6 Antinomies of Common Knowledge of Rationality

How might CKR be defended? The most straightforward way using a bit of modal logic is to assume that if  $s$  is true, then there is a possible world in which all agents know that  $s$  is true. Let  $s_1$  be the statement that there is a possible world in which  $s$  is true. Then there is a possible world in which all agents know  $s_1$ . Let  $s_2$  be the statement that there is a possible world in which all agents know  $s_1$ . We continue the argument similarly, showing the existence of a sequence of possible worlds in which  $s, s_1, s_2, \dots$  are known. The existence of this sequence is CKR.

We know from the Bob and Alice example, which shows that CKR is not true, that the premise "if  $s$  is true, then there is a possible world in which all agents know  $s$ " must be false. Philosophers have, in fact, long known the impermissibility of this premise. For instance, consider the famous Moore paradox where  $s$  is the statement "It is raining outside but I don't know it" (Green and Williams 2007). It may be true that it is raining outside and I do not know that it is raining outside, and everyone in the world might know this except me. but I cannot know this truth.

The following situation, known as the surprise exam problem, shows how back-

ward induction type arguments of the type common when CKR is assumed can be erroneous when truth does not imply the possibility of knowing the truth. Consider a class of philosophy students taking a logic course that meets each week from Monday through Friday. The instructor announces that there will an exam one day next week, but the particular day it is given will be a surprise. One student thinks to himself, “The exam cannot be given on Friday because then it would not be a surprise.” He then noted that a similar argument shows that the exam could not be given on Thursday. And so on. He concludes that a surprise exam cannot be given. On the next Wednesday, the instructor gives an exam, and all the students are surprised. A rational student, anticipating this event, will conclude that it is true that the exam will be given one day next week, but he himself cannot know this fact in a formal sense that allows the deployment of the laws of logic.

For an overview of the many proposed solutions to the surprise examination problem (it has several other names) by philosophers and logicians see Margalit and Bar-Hillel (1983) and Chow (1998). Interpretations vary widely, and there is no single accepted solution. There are a number of cogent analyses using standard logic and modal logic to show that the instructor’s statement is impermissively self-referential or self-contradictory, and because a false statement can validly imply anything, there is no paradox in the instructor’s prediction being correct. However, it is clear that my treatment is valid, provided I can offer a rigorous logical model in which there are unknowable truths.

I will follow Binkley (1968). Let us assume there are only two days, Monday and Tuesday. The full term argument is similar, but (much) longer. We take the case of a single student with knowledge operator  $k$ . We assume for any knowledge operator that

- A1  $kf \implies \neg k\neg f$
- A2  $kf \& k(f \implies g) \implies kg$
- A3  $kf \implies kkf$

Note that A1 is weaker than the usual assumption  $kf \implies f$ ; i.e., “what is known is possible” is weaker than “what is know is true.” We also assume the student knows all tautologies of the propositional calculus and all axioms.

Let  $k_m f$  mean “the student knows  $f$  on Monday” and let  $k_t f$  mean “the student knows  $f$  on Tuesday.” Let  $E_m$  be the event “the exam is given on Monday,” and let  $E_t$  be the event “the exam is given on Tuesday.” We assume

- A4  $\neg E_m \implies k_t \neg E_m$
- A5  $k_m f \implies k_m k_t f$

A4 says that if the exam is not given on Monday, then on Tuesday the student knows this fact, and A5 says that if the student knows something on Monday, he knows on Monday that he will continue to know it on Tuesday. The instructor’s

assertion can be written as

$$E = (\neg E_m \iff E_t) \& (E_m \implies \neg k_m E_m) \& (E_t \implies \neg k_t E_t). \quad (7)$$

Let us assume  $k_m E$ . From A4 we have

$$k_m(\neg E_m \implies k_t \neg E_m). \quad (8)$$

From  $k_m E$  we have  $k_m(E_t \implies \neg E_m)$ , which with (8) gives

$$k_m(E_t \implies k_t \neg E_m). \quad (9)$$

Now from  $k_m E$  and A5, we have  $k_m k_t(\neg E_m \implies E_t)$ , so

$$k_m(k_t \neg E_m \implies k_t E_t). \quad (10)$$

From (9) and (10), we have

$$k_m(E_t \implies k_t E_t). \quad (11)$$

Now  $k_m E$  implies  $k_m(E_t \implies k_t \neg E_t)$ , which together with (11) implies  $k_m(\neg E_t)$ , and hence  $k_m E_m$ . This, together with  $k_m E$  gives

$$k_m \neg k_m E_m. \quad (12)$$

However,  $k_m E_m$  and A3 imply  $k_m k_m E_m$ , so by A1, we have  $\neg k_m \neg k_m E_m$ , which contradicts (12). Therefore the original assumption  $k_m E$  is false.  $E$  is thus true but it is inadmissible to assume therefore that the student knows that  $E$  is true.

It follows that CKR is self-contradictory.

## REFERENCES

- Andreoni, James and John H. Miller, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence," *Economic Journal* 103 (May 1993):570–585.
- Aumann, Robert J., "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior* 8 (1995):6–19.
- and Adam Brandenburger, "Epistemic Conditions for Nash Equilibrium," *Econometrica* 65,5 (September 1995):1161–1180.
- Basu, Kaushik, "The Traveler's Dilemma: Paradoxes of Rationality in Game Theory," *American Economic Review* 84,2 (May 1994):391–395.

- Bernheim, B. Douglas, "Rationalizable Strategic Behavior," *Econometrica* 52,4 (July 1984):1007–1028.
- Binkley, Robert, "The Surprise Examination in Modal Logic," *Journal of Philosophy* 65 (1968):127–135.
- Chow, Timothy Y., "The Surprise Examination or Unexpected Hanging Paradox," *American Mathematical Monthly* 105 (1998):41–51.
- Gintis, Herbert, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences* (Princeton: Princeton University Press, 2009).
- Green, Mitchell S. and John N. Williams, *Moore's Paradox: New Essays on Belief, Rationality and the First-Person* (Oxford: Oxford University Press, 2007).
- Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory* 27 (1982):245–252.
- Margalit, Avishai and Maya Bar-Hillel, "Expecting the Unexpected," *Philosophia* 13 (1983):263–288.
- McKelvey, R. D. and T. R. Palfrey, "An Experimental Study of the Centipede Game," *Econometrica* 60 (1992):803–836.
- Pearce, David, "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* 52 (1984):1029–1050.