IS COMMON KNOWLEDGE OF RATIONALITY SLUGGISH?

DOV SAMET

1. Introduction

The iterative elimination of strongly dominated strategies can be justified by common knowledge of rationality. Rationality in this context means that players do not play strategies which are strongly dominated in a game they know they play. Indeed, each iteration of elimination of strongly dominated strategies seems to correspond to an iteration of mutual knowledge, that is, an iteration of "all know that". Thus, the first round of elimination is justified by rationality. The second is justified by mutual knowledge of rationality, the third, by mutual knowledge of mutual knowledge of rationality and so on.

The example in Section 2 demonstrates that for games with an infinite number of strategies the said correspondence may fail to hold. For common knowledge, only the infinitely many *finite* iterations of "all know that" are required. In contrast, elimination of strongly dominated strategies may require *transfinite* iterations, that is, eliminations made after all the finite rounds of elimination. This result is puzzling. On one hand, it seems plausible that common knowledge of rationality cannot hold if the transfinite process of elimination is not completed. On the other hand, the finite iterations of "all know that" culminate in common knowledge of rationality before the task of elimination is exhausted. Common knowledge of rationality seems to sluggishly follow the process of elimination of strongly dominated strategies, which is carried on vigorously beyond all finite rounds.[1]

One is tempted to continue transfinitely the iteration of mutual knowledge. Indeed, Barwise (1988) showed that in a non-well-founded set theoretic model, common knowledge is reached after transfinite iterations of mutual knowledge. Also, Heifetz (1999) showed that in infinitary logic common belief requires a transfinite sequence of iterations of mutual belief. However, the puzzling phenomenon demonstrated here and in Lipman (1994) is presented in standard set theoretic terms and simple finitary language. In such a setup mutual knowledge of rationality of all finite order *does* define common knowledge of rationality, and no transfinite steps are required.

We show that the failure to derive the transfinite process of iterative elimination of strongly dominated strategies by the finite iterations of "all know that" is due to the inadequacy of the proof and not to the notion of common knowledge. Common knowledge of rationality is perfectly captured by the finite iterations of mutual knowledge on one hand, and on the other hand it does imply the transfinite process of elimination of strongly dominated strategies.

[1]Lipman (1994) presented a similar phenomenon in the more elaborate probabilistic context of rationalizability, where strategies that are not best response are eliminated iteratively (Bernheim, 1984; Pearce, 1984).

	0	1	2	3	4	$\cdots$	a	b
0	0	0	0	0	0	$\cdots$	0	0
1	1	1	1	1	1	$\cdots$	1	1
2	0	2	2	2	2	$\cdots$	2	2
3	0	0	3	3	3	$\cdots$	3	3
4	0	0	0	4	4	$\cdots$	4	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	1	2	3	4	5	$\cdots$	0	0
b	0	1	2	3	4	$\cdots$	1	1

FIGURE 1. Player 1's payoff matrix in a two-player symmetric game

2. An example

Consider player 1's payoff matrix in a two-player symmetric game depicted in Figure 1, where the set of strategies of each player is $\{0, 1, 3, \dots, a, b\}$.[2] For $i \geq 0$, let $S^i = \{i, i+1, \dots, a, b\}$. Note, that the matrix of the game with a set of strategy profiles $S^i \times S^i$ can be obtained from the original matrix by adding i to each of the payoffs except for the payoffs of a player when she plays a or b, which remain the same.

The only dominated strategies in the game are the strategies 0 of both players, as these strategies are strongly dominated by the strategies 1. When the 0 strategies are eliminated we are left with the matrix $(S^1)^2$. By the comment in the previous paragraph, the only dominated strategies in this matrix are strategies 1 of each player. Repeating this elimination successively we eliminate all strategies $i \in \{0, 1, \dots\}$ of both players. The remaining matrix is $\{a, b\}^2$. Here, strategy a of both players is strongly dominated by b. Hence, the process continues for one more round, round ω, which comes after all finite rounds. The process ends in round ω with the profile $\{b\}^2$.

We now review the process in terms of common knowledge of rationality. When players are rational they do not play strategy 0. Thus, the game played is given by the matrix $(S^1)^2$. If the players know that they are rational, then they know that the game is given by this matrix, and being rational they do not play strategy 2, and the game played is given by the matrix $(S^2)^2$. If they know that they know that they are rational, then they know that this is the matrix of the game, and being rational they do not play strategy 3. Common knowledge of rationality means that the players know that they are rational (and hence they are rational), and that they know that they know that they are rational and so on. Thus common knowledge of rationality implies the successive elimination of all strategies $\{0, 1, 2, \dots\}$. Indeed, each iteration of mutual knowledge, in the definition of common knowledge, corresponds, as we have seen, to an iteration of elimination of strongly dominated strategies.

The argument that accompanies the process shows that when rationality is common knowledge then the players face the game with the matrix $\{a, b\}^2$. The finite iterations of mutual knowledge that amount to common knowledge of rationality

[2]The payoffs in this game are unbounded. But the process of elimination will be the same if each payoff i is replaced by a number a_i such that the sequence a_i is strictly increasing and bounded.

seem to be out of sync with the iterations of the elimination of strongly dominated strategies. Common knowledge is achieved after all the finite iterations of "all know that", while the elimination of strongly dominated strategies requires one more round of elimination, round ω, which comes after all the finite rounds. It is hard to accept that common knowledge of rationality implies *only* that players know that they are playing either a or b. If they know, they surely will not play a. However, common knowledge of rationality seems to be achieved before a is eliminated.

One is tempted to say that this example shows that common knowledge of rationality is *not* attained after all the finite iterations of knowledge and one more iteration is required. But as we will see, when knowledge is formally defined, common knowledge *is* achieved after all finite iterations of mutual knowledge. The failure to prove that common knowledge implies the elimination of a is a failure of the proof, not of the non-transfinite nature of common knowledge. We will prove that common knowledge of rationality does imply the elimination of a, and explain why the suggested proof above fails to show it.

3. The puzzle formally reproduced

We now examine the relation between common knowledge of rationality and the iterative elimination of strongly dominated strategies in a formal setup.

Iterative elimination of strongly dominated strategies. We consider a two-player game with strategy sets S_1 and S_2. The analysis of games with many players is similar. We refer to any set of strategy profiles $S' = S'_1 \times S'_2$, where $S'_i \subseteq S_i$ for $i = 1, 2$, as a game or a subgame of S.

Consider a weakly decreasing sequence of games $S^\alpha = S_1^\alpha \times S_2^\alpha$ starting with $S^0 = S_1 \times S_2$, where α ranges over all ordinals with cardinality which does not exceed the cardinality of the set of strategy profiles. Such a sequence is a *process of iterative elimination of strongly dominated strategies* if it satisfies the following conditions. For each ordinal β, if the game $S^{<\beta} = (\cap_{\alpha<\beta} S^\alpha)$ has dominated strategies then S^β is obtained by eliminating one or more such strategies from $S^{<\beta}$.[3] If the game $S^{<\beta}$ does not have any dominated strategy, then $S^\beta = S^{<\beta}$. The game $\cap_\alpha S^\alpha$ is the game that survives the whole sequence of eliminations and it is called the *terminal game* of the process.[4]

Models of knowledge. In order to formalize notions of knowledge, we consider a state space the subsets of which, called *events*, corresponding to sentences that are used to discuss the game. We assume that for each state there is a specification of the strategy played at the state by each of the players. Thus, the event that the strategy profiles played by the players are in the game S' is well defined. We

[3] If β is not a limit ordinal, then $\beta = \gamma + 1$, and due to the monotonicity of the sequence, the game $S^{<\beta}$ is $S_1^\gamma \times S_2^\gamma$.

[4] The order of elimination in finite games is known to be independent of the order of elimination. But in infinite games the order does matter. Consider, for example, a game in which player 1's strategy set is $\{0, 1, 2, \dots\}$, and suppose that when she plays n her payoff is n independently of the second player. We can eliminate in the first round all strategies of player 1, but one of them. The strategy which is not eliminated is now the terminal game. This example shows also that unlike finite games, in infinite games the terminal game may be empty. We can eliminate in the first round all of player 1's strategy. Dufwenberg and Stegeman (2002) provided conditions that guarantee order independence of such processes.

denote this event by $[S']$. Knowledge is introduced by operators K_i for each player i, such that for any event E, K_iE is the event that i knows E.[5] We skip the details of the construction of the state space, which is documented in numerous publications.[6] Recall the following two properties of knowledge operators that we use in the sequel.[7]

Distributivity: Knowledge distributes over conjunction. That is, for any family of events $(E_x)_{x\in X}$, $\mathrm{K}_i(\cap_x E_x) = \cap_x \mathrm{K}_i E_x$.

Truth: What is known is true. That is, for each E, $\mathrm{K}_iE \subseteq A$.

By applying distributivity to two events $E \subseteq F$ we conclude straightforwardly that $\mathrm{K}_iE \subseteq \mathrm{K}_iF$. Thus, knowledge operators are *monotonic*.

Common knowledge. We define an operator K of mutual knowledge which corresponds to the claim that all players know. Thus, for each E, $\mathrm{K}E = \mathrm{K}_1E \cap \mathrm{K}_2E$. The properties of distributivity, truth, and monotonicity of the individual knowledge operators K_i are trivially inherited by the operator K.

Powers of K mean iteration of the phrase "all know". Thus, K^1E is simply $\mathrm{K}E$, and $\mathrm{K}^{n+1}E = \mathrm{K}\mathrm{K}^nE$. The event that E is *common knowledge* is $\mathrm{C}E = \cap_{n\geq 1}\mathrm{K}^nE$.[8]

Rationality. For our purposes, rationality means that when all players know that they are playing a game S', and $S'' \subseteq S'$ is obtained by eliminating strategies in S' that are strongly dominated in this game, then the strategy profile they play is in S''. Thus, rationality requires that either it is not the case that all the players know that S' is played, or else they must play a strategy profile in S''. That is, for rationality, $\neg\mathrm{K}[S'] \cup [S'']$ must hold for each such pair S' and S'', and thus the event that the players are rational is $R = \cap_{S',S''}\neg\mathrm{K}[S'] \cup [S'']$, where the intersection is over all pairs S' and S'' such that the S'' is obtained from S' by elimination of some strategies that are strongly dominated in S'.[9]

---

[5]We make all the definitions, claims, and proofs that follow, as close as possible to their formulation in natural language or in a syntax of a well defined language. This is done by not using the term "state". Thus, although events and knowledge operators are set theoretic semantic notions, they are used here in a way that makes it possible to translate them to natural language.

[6]Kripke models serve as set theoretic semantics of knowledge. In such models a set of states is endowed with the binary relationship of accessibility for each player. Knowledge operators are easily defined in Kripke models. The event K_iE consists of each state from which only states in E can be accessed. The partition model in Aumann (1976) is a Kripke model for a set of axioms in the syntax, called S5, in which the accessibility relations are equivalence relations. There is another class of models, based on Boolean algebras with operators. A set theoretic model with operators K_i, as discussed here, is a Boolean algebra model for knowledge, where the algebra consists of all subsets of the state space. A discussion of such models can be found in Samet (2010). Boolean algebra models are more general than Kripke models and they are important for the study of the relation between knowledge and belief (Halpern *et al*, 2009a,b).

[7]Knowledge operators in partition models also have two properties of introspection, positive and negative. These properties are not required for the analysis of common knowledge that we carry out here.

[8]Aumann (1976) introduced common knowledge in a set theoretic semantic model of partitions. The event that E is common knowledge is the union of all elements of the meet of the players' partitions that are contained in E.

[9]Rationality is defined here as the event R that all players are rational. Hillas and Samet (2014) define the event R_i that player i is rational for each i, and define rationality as the event $R = \cap_i R_i$. This definition of rationality of all players implies the definition given here.

This formal presentation helps us to follow closely the argument made in the example in Section 2, and reproduces the same puzzle: Common knowledge of rationality seems to imply only the finite rounds of elimination of strongly dominated strategies.

Proposition 1. *Common knowledge of rationality implies that the game being played is the intersection of all finitely indexed games in the sequence. That is,*

$$\mathrm{C}R \subseteq [\cap_{n<\omega} S^n].$$

Proof. It is enough to show that $\mathrm{K}^n R \subseteq [S^{n+1}]$ for each finite n. We prove this by induction on n.[10] First, note that $R \subseteq [S^1]$. This follows since $R \subseteq \neg\mathrm{K}[S^0] \cup [S^1]$ and $\neg\mathrm{K}[S^0] = \emptyset$.

We now prove the induction hypothesis for $n = 1$. As $R \subseteq [S^1]$, it follows by monotonicity that $\mathrm{K}R \subseteq \mathrm{K}[S^1]$. Thus, by truth, $\mathrm{K}R \subseteq R \cap \mathrm{K}[S^1] \subseteq (\neg\mathrm{K}[S^1] \cup [S^2]) \cap \mathrm{K}[S^1] = [S^2]$. Suppose the hypothesis is proved for $n/$ Then $\mathrm{K}^{n+1}R = \mathrm{K}\mathrm{K}^n R \subseteq R \cap \mathrm{K}[S^{n+1}] \subseteq (\neg\mathrm{K}[S^{n+1}] \cup [S^{n+2}]) \cap \mathrm{K}[S^{n+1}] = [S^{n+2}]$. □

4. The solution

Evidence. We say that event E is an *evidence* of F if E implies that all know F, that is, $E \subseteq \mathrm{K}F$. An event E is *self evident* if it is an evidence of itself, that is $E \subseteq \mathrm{K}E$. Note that by the truth property, $\mathrm{K}E \subseteq E$, and therefore E is self evident when $\mathrm{K}E = E$. That is, E is a fixed point of K, or equivalently, it is a fixed point of all the individual knowledge operators K_i.[11] Self evident events are formal rendering of situations in which all the players are present and each one of them knows the description of the situation. We are interested in self evident situations in which some fact becomes known to the players, like an announcement of some fact in the presence of all players, such that they can describe the situation. Formally such a situation is described by an event E which is self evident and is an evidence of F, that is $E \subseteq \mathrm{K}E \cap \mathrm{K}F$.

The following proposition shows that a self evident event E which is an evidence of F implies that F is commonly known. Such events seem to describe the simplest situations, and perhaps the only ones that make a statement commonly known. This proposition also shows that the event that F is commonly known is self evident and it is an evidence of F. However, there can be other self evident events that are evidence of F. Common knowledge of F is the largest event with this property.[12]

Proposition 2. *A self evident event E which is an evidence of F implies common knowledge of F, that is $E \subseteq \mathrm{C}F$. Moreover, $\mathrm{C}F$ itself is self evident and an evidence of E, and thus it is the largest, and hence the least informative event with these properties.*

---

[10] For simplicity, we assume that if the game is finite, the sequence S^n is defined for all $n < \omega$. Of course, this sequence is constant after some finite round.

[11] The semantic definition of common knowledge in Aumann (1976) is in the spirit of the fixed point definition. An element of the meet is an event P which is the union of elements or each of the players' partitions. Hence $P \subseteq \mathrm{K}P$.

[12] When knowledge is generated by a partition, then the event that F is commonly known is the union of *all* the elements of the meet which are contained in F. A self evident event which is an evidence of F is *any* union of elements of the meet which are contained in E.

Proof. Assume that $E \subseteq \mathrm{K}E \cap \mathrm{K}F$. We prove by induction on n that $E \subseteq \mathrm{K}^n F$, and thus $P \subseteq \mathrm{C}E$. For $n = 1$ this follows from our assumption. Suppose that $E \subseteq \mathrm{K}^n F$, then by monotonicity, $\mathrm{K}E \subseteq \mathrm{K}^{n+1} F$, and since per our assumption $E \subseteq \mathrm{K}E$ it follows that, $E \subseteq \mathrm{K}^{n+1} F$.

Obviously $\mathrm{C}F$ is an evidence of F, as $\mathrm{C}F \subseteq \mathrm{K}F$. To see that it is self evident, we note that by distributivity, $\mathrm{K}(\mathrm{C}F) = \cap_{n\geq 2}\mathrm{K}^n F$. However, as $\mathrm{K}F \subseteq F$, $\mathrm{K}^2 E \subseteq \mathrm{K}E$, and thus $\cap_{n\geq 2}\mathrm{K}^n F = \cap_{n\geq 1}\mathrm{K}^n F$. We conclude that $\mathrm{C}E = \mathrm{K}(\mathrm{C}E)$. □

The transfinite process of elimination of strongly dominated strategies is fully captured when we view common knowledge of rationality as a self evident event which is an evidence of rationality.

Proposition 3. *If E is self evident and an evidence of rationality, then*

$$E \subseteq [\cap_\alpha S^\alpha].$$

In particular,

$$\mathrm{C}R \subseteq [\cap_\alpha S^\alpha].$$

Proof. Assume that $E \subseteq \mathrm{K}E \cap \mathrm{K}R$. We prove by transfinite induction on α that for each α, $E \subseteq [S^\alpha]$. By the truth axiom and the definition of R, $E \subseteq R \subseteq \neg\mathrm{K}[S^0] \cup [S^1]$. Since $\neg\mathrm{K}[S^0] = \emptyset$ it follows that $E \subseteq [S^1]$.

Suppose that $E \subseteq [S^\beta]$ for all $\beta < \alpha$. Then $E \subseteq S^{<\alpha}$, and by monotonicity, $\mathrm{K}E \subseteq \mathrm{K}[S^{<\alpha}]$. Thus, $E \subseteq R \cap \mathrm{K}E \subseteq (\neg\mathrm{K}[S^{<\alpha}] \cup [S^\alpha]) \cap \mathrm{K}[S^{<\alpha}] = [S^\alpha]$. □

5. The process of elimination is sluggish

By Proposition 3, for each terminal game S', $\mathrm{C}R \subseteq [S']$, and thus $\mathrm{C}R \subseteq \cap[S'] = [\cap S']$, where the intersection is over all terminal games. However, the event that $\cap S'$ is played is not a sufficient condition for common knowledge of rationality. Consider, for example, the game in Section 2. The subgame $\{b\}^2$ is a terminal game, indeed the only one. However, this game is incompatible with $\mathrm{C}R$, because $\mathrm{C}R$ is necessarily empty. To see this, note that as $\mathrm{C}R \subseteq [\{b\}^2]$ and since $\mathrm{KC}R = \mathrm{C}R$, it follows by monotonicity that $\mathrm{C}R = \mathrm{KC}R \subseteq \mathrm{K}[\{b\}^2]$. By monotonicity, $\mathrm{K}[\{b\}^2] \subseteq \mathrm{K}[\{b, 2\} \times \{b\}]$. In the game $\{b, 2\} \times \{b\}$, strategy 2 strongly dominates b. Thus, $\mathrm{C}R \subseteq R \cap \mathrm{K}[\{b, 2\} \times \{b\}] \subseteq [\{2\} \times \{b\}]$. Therefore, $\mathrm{C}R$ is a subset of two disjoint events, and hence $\mathrm{C}R = \emptyset$.

We started with the puzzle that common knowledge of rationality seems to sluggishly follow the full transfinite process of elimination of strongly dominated strategies. We then showed that the self evident nature of common knowledge of rationality does imply the full process of elimination. Now, we see that it is the transfinite process of elimination which is lagging behind common knowledge of rationality. In our example, the process ends with the profile $\{b\}^2$, while common knowledge of rationality eliminates this profile too.

The reason for this is obvious. The strategies that dominate b in the terminal game $\{b\}^2$ were eliminated and cannot be used again to eliminate b. This cannot happen when the set of strategies is finite, as we explain now. If s_i is a strategy of i that was eliminated in the process, then there was a strategy s_i' that dominates it in some subgame S'. If s_i' was eliminated too, then there is a strategy s_i'' that dominates s_i' in some subgame $S'' \subseteq S'$. This sequence of eliminated strategies is finite, and therefore it should reach eventually the terminal game. This shows that no strategy s_i that was eliminated can dominate a strategy in the terminal

game. However, in the case of infinite set of strategies, the sequence of eliminated strategies can be infinite, as it is in our example. Hence, it may never reach the terminal game. The result is, that strategies that are eliminated can dominate strategies in the terminal game.

We conclude that the process of elimination of strongly dominated strategies does not capture common knowledge of rationality.[13] The example hints also to the process that does converge to a terminal game which is consistent with common knowledge of rationality. It is a process in which strategies can be eliminated even if they are dominated by strategies that were already eliminated. Chen *et al* (2007) showed that this process does capture common knowledge of rationality. Obviously, to prove their result they use the self evident nature of common knowledge rather than the iteration of mutual knowledge.

6. Discussion

Proposition 3 answers in the affirmative the question whether common knowledge of rationality implies the full process of elimination of strongly dominated strategies. The proof makes use of the fixed point definition of common knowledge, and thus decouples the two iterative processes: the one that reaches common knowledge and the one that reaches a game that does not have strongly dominated strategies.

The question still remains why these two iterative processes fail to match beyond the finite rounds. To pinpoint the reason for this, we start by emphasizing the common features of the two iterative processes. For this, consider an operator D on the set of all the subgames of S. For each game S', let $\mathrm{D}(S')$ be the game obtained by eliminating *all* strongly dominated strategies in S'. The operators K and D share the following properties.

(1) Both operators are contractions: By the truth axiom $\mathrm{K}(E) \subseteq E$, and by definition $\mathrm{D}(S') \subseteq S'$.
(2) As a result, the iterative application of both operators generates a decreasing sequences.
(3) The iterative application of K to rationality and the iterative application of D to S are clearly related, as rationality is defined here as not using strongly dominated strategies in a subgame known to the players as the game they are playing.

The reason why despite these similarities D can produce a transfinite sequence, while K is "stuck" after the finite rounds is the difference in distributivity. The distributivity of K implies that $\mathrm{K}(\mathrm{C}E) = \mathrm{C}E$ and therefore applying K after all finite rounds does not change the event. Thus, reaching the largest public announcement of rationality is achieved after all finite iterations of K. However, D does not distribute over intersections and in particular, $\mathrm{D}(\cap_n \mathrm{D}^n S)$ can be a subset of $\cap_{n>2}\mathrm{D}^n S$ rather then being the same set. Hence, reaching the terminal game may require a transfinite sequence of elimination.

Although the two definitions of common knowledge discussed here are equivalent, many students of common knowledge believe that it is the fixed point nature of common knowledge that captures the "right" aspect of the notion. The iterative

[13] Dufwenberg and Stegeman (2002) noted the many problems that afflict this process. However, since rationality and common knowledge of rationality were not formally introduced in their paper, they could not state explicitly the incompatibility of the process with common knowledge of rationality.

definition is just a way to describe the largest self evident event which is an evidence to the event which is commonly known. Moreover, even the iterative definition does not imply that common knowledge is a *process*. Rather, common knowledge is a fact; a state of minds; a statement; an event that happens to imply many other statements and events.

References

Aumann, R. (1976), Agreeing to Disagree, *The Annals of Statistics*, 4, 1236-1239.

Barwise, J. (1988), Three views of common knowledge, in *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann Publishers Inc.

Bernheim, D. (1984), Rationalizable strategic behavior, *Econometrica* 52, 1007-1028.

Chen, Y., N.V. Long, X. Luo (2007), Iterated strict dominance in general games, *Games and Economic Behavior*, 61, 299-315.

Dufwenberg, M., and M. Stegeman (2002), Existence and Uniqueness of Maximal Reductions Under Iterated Strict Dominance, *Econometrica*, 70, 2007-2023.

Halpern, J., E. Segev, and D. Samet (2009a), On definability in multimodal logic, *The Review of Symbolic Logic*, 2, 451-468.

Halpern, J., E. Segev, and D. Samet (2009b), Defining knowledge in terms of belief: The modal logic perspective, *The Review of Symbolic Logic*, 2, 469-487.

Heifetz, A. (1999), Iterative and fixed point common belief *Journal of Philosophical Logic*, 28, 61-79.

Hillas, J. and D. Samet (2014), Weak dominance: A mystery cracked, a manuscript.

Lipman, B. L. (1994), A note on the implications of common knowledge of rationality, *Games and Economic Behavior*, 6, 114–129.

Monderer, D., and D. Samet, (1989), Approximating common knowledge with common beliefs, *Games and Economic Behavior*, 1(2), 170–190.

Pearce, D. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* 52, 1029-1050.

Samet, D. (2010), S5 knowledge without partitions, *Synthese* (2010) 172, 145-155.