Int J Game Theory (1999) 28:435–442

International Journal of Game Theory
© Springer Verlag 1999

How canonical is the canonical model? A comment on Aumann's interactive epistemology*

Aviad Heifetz

The School of Economics, Tel Aviv University, Tel Aviv 69978, Israel
(e-mail: aviad@econ.tau.ac.il)

Abstract. Aumann (1989) argued that the natural partitions on the space of all maximally consistent sets of formulas in multi-player S5 logic are necessarily "commonly known" by the players. We show, however, that there are many other sets of partitions on this space that conform with the formulas that build the states – as many as there are subsets of the continuum! Thus, assuming a set of partitions on this space is "common knowledge" is an informal but meaningful meta-assumption.

Key words: Common knowledge, epistemic logic

In 1976 Aumann suggested modeling uncertainty in game theory with partition spaces. A model of this kind consists of a space of possible states Ω, together with a partition Π^i of Ω for each player i. The member $\Pi^i(\omega)$ of the partition Π^i that contains ω is the set of states player i considers as possible when ω prevails. So for every event $E \subseteq \Omega$

$$K^i(E) = \{\omega \in \Omega : \Pi^i(\omega) \subseteq E\}$$

is the event where player i knows for sure that E occurs. $K^j K^i(E)$ is the event where j knows that i knows E, and so forth. Thus, the knowledge operators enable to unfold the mutual knowledge and uncertainties of the players in the model.

Aumann was well aware of the fact that while using these models, one assumes that the players may be uncertain about the true state of the world, but

* This work is part of my PhD dissertation in Tel Aviv university. I am grateful to my advisors, Dov Samet and Ehud Lehrer.

not about the partitions of their fellows. Justifying this assumption, Aumann (1976) wrote:

"The implicit assumption that the information partitions ... are themselves common knowledge ... constitutes no loss of generality. Indeed in the full description of a state ω of the world is the manner in which information is imparted to the two persons. This implies that the information sets $\Pi^1(\omega)$ and $\Pi^2(\omega)$ ... are indeed defined unambiguously as functions of ω, and that these functions are known to both players."

In his 1989 notes Aumann elaborated further: "When we come to interpret the model ... an inevitable question is, 'what do the participants know about the model itself?' Does each 'know' the information partitions Π^i of the others? Are the Π^i themselves in some sense 'common knowledge'? If so, how does the model reflect each individual's information – or lack of information – about the others' partitions? To do this right, doesn't one need to superpose another such model over the current one, to deal with knowledge (or uncertainty about) the Π^i? But then, wouldn't one need another such model, without end even in the transfinite domain?

... The most convincing way to remove all these questions and doubts is to construct Ω and the Π^i – or equivalently, the K^i – in an explicit, canonical manner, so that it is clear that the knowledge operators are 'common knowledge' in the appropriate sense."

Then, Aumann introduces the canonical model Ω of all maximally consistent sets of formulas in S5 epistemic logic, that has a knowledge operator k^i for each player i. (In the appendix we recall the details of this construction, as well as the other logical ingredients we use.) The partitions of the players in this model are indeed defined by the structure of the states:

$$\Pi^i(\omega) = \{\omega' \in \Omega : k^i\varphi \in \omega \Leftrightarrow k^i\varphi \in \omega'\}.$$

If we denote by $[\psi]$ the set of states to which the formula ψ belongs, one can prove that for every $\omega \in \Omega$

$$\omega \in K^i[\psi] \quad \Leftrightarrow \quad k^i\psi \in \omega. \tag{1}$$

This implies that with the partitions Π^i each state in Ω is a model of all the formulas it contains.

Aumann (1989) continues: "Thus the question becomes, does each individual 'know' the operators k^i of the others (in addition, of course, to his own)?

"The answer is 'yes'. The operator k^i operates on formulas; it takes each formula f to another formula. Which other formula? What is the result of operating on f with the operator k^i? Well, it is simply the formula k^if. 'Knowing' the operator k^i just means knowing this definition. Intuitively, for an individual j to 'know' k^i means that j knows what it means for i to know something. It does not imply that j knows any specific formula k^if.

"... Thus the assertion that each individual 'knows' the knowledge operators of all individuals has no real substance; it is part of the framework. If j did not 'know' the operators k^i, he would be unable even to consider formulas in the language..., to say nothing of knowing or not knowing them."

we are in perfect agreement with Aumann's analysis in the last paragraph, but not with the leap forward that immediately follows: "From this we conclude that all individuals indeed 'know' … the partitions Π^i."

True, the partitions Π^i of the players on the canonical model Ω are defined very naturally in terms of the formulas of the form $k^i\varphi$ that constitute every state. In effect, these partitions are the coarsest possible partitions on Ω such that each state models all the formulas that belong to it. (it is easily verifiable that no state can be added to any partition member without spoiling this property.) *However, there are many finer partitions that do the job, as we demonstrate below. This means that it is not trivial at all to assume it is "common knowledge" which of these partitions every player has.*

To see there are finer partitions on the canonical model that are coherent with the formulas that build them, consider the space[1]

$$V = \{0,1\} \times [0,1) \times [0,1).$$

The set of players is $I = \{a,b\}$. A typical state in V will be of the form

$$v = (r, 0.a_1a_2\ldots, 0.b_1b_2\ldots), \tag{2}$$

where the numbers in $[0,1)$ are written as binary expansions, and where dyadic numbers are written in their terminating version (with zeroes from some stage on). The first of these numbers will be called the type of player a in v, and the second the type of player b. Let there be a unique atomic formula φ in the language, that holds exactly in those states where $r = 1$.

Adopt the convention that player j is the opponent of player i. The partition of a player $i \in I$ will be

$$\begin{aligned} &P^i((r, 0.a_1a_2\ldots, 0.b_1b_2\ldots)) \\ &\quad = \{i_1r, 1 - i_1(1-r)\} \times \{0.i_1i_2\ldots\} \\ &\qquad \times \bigcap_{n=1}^{\infty} \bigcup_{k=1}^{2^{n-1}} \left[\frac{2(k-1) + i_{n+1}j_n}{2^n}, \frac{2k - i_{n+1}(1-j_n)}{2^n}\right). \end{aligned}$$

This is a closed form but non-transparent way to say, that the equivalence relation $\sim_i$ that defines the partition of player i on V is

$$\begin{aligned} &(r, 0.a_1a_2\ldots, 0.b_1b_2\ldots) \sim_i (r', 0.a_1'a_1'\ldots, 0.b_1'b_2'\ldots) \\ &\Leftrightarrow \quad i_n' = i_n \ \forall n \in N, \quad i_1 = 1 \Rightarrow r' = r, \quad i_{n+1} = 1 \Rightarrow j_n' = j_n, n \in N. \end{aligned}$$

In words: player i knows his own type; he knows whether φ holds iff his first digit is 1; he knows the n-th digit of his opponent iff his $n+1$-th digit is 1.

For example, when $a_2 = 1$, player a knows whether b_1 is 1 or 0. $b_1 = 1$ exactly when player b knows whether r is 1 or 0, i.e. whether φ holds or not. Inductively, $a_n = 1$ exactly when player a knows whether b knows whether a

[1] The same construction was used for a different purpose in Hart, Heifetz and Samet (1993), and a related one in Heifetz and Samet (1993).

knows ... (n levels) whether φ holds. In any case, player a always knows whether the above assertion is true or not. Notice that had we added even one extra state to a partition member of player a, one of the last two sentences would no longer hold for some $n \in N$, and similarly for player b.

The partition P^i tells us in what states player i knows φ. The partition P^j then tells us in what states player j knows, for instance, $\varphi \wedge k^i\varphi$, and so on. By induction on the structure of formulas in the language, we can tell what formulas hold in what states. Thus, the observation in the last paragraph implies that had we added even one extra state to a partition member of a player, this would have affected the formulas that hold at least in some of the states. Notice further that if the depth of a formula ψ is n, i.e. the nesting of the operators k^i and k^j inside ψ is of depth n, the truth of ψ in a state v as in (2) is determined only by r and by the first n digits in the types of the players (this is easily verifiable by induction).

From now on, *identify each $v \in V$ with the maximally consistent set of formulas that hold in it.* Not every maximally consistent set of formulas ω that is a state in the canonical model Ω appears also in V. But different partition members of P^i map into different partition members of Π^i, since as we saw, in both spaces we can not add extra states to the members of these partitions without changing the formulas that obtain in states. Thus, the restriction of Π^i to V yields the partition P^i.

Now, we claim that in every $v \in V$, $\Pi^i(v)$ can be refined without changing the formulas that hold in the states of Ω. In the restriction of Π^i to V there is a continuum of members (one for each type in $[0,1)$), so we can decide independently whether or not to refine each such member. Then we can carry a similar procedure for player j. This gives us:

Theorem 1. *With one atomic formula φ and two players a and b, the space Ω of S5 maximally consistent sets of formulas admits $|2^{[0,1)}|$ different pairs of partitions, such that for every formula ψ and every player i*

$$\omega \in K^i[\psi] \Longleftrightarrow k^i\psi \in \omega,$$

where K^i is the knowledge operator of player i in any one of these pairs.

Proof. Fix a player i and a state

$$v = (r, 0.a_1a_2\ldots, 0.b_1b_2\ldots) \in V.$$

Let $\{v_k\}_{k=0}^{\infty} \in N$ be the sequence such that $i_{v_k+1} = 0$ for all $k \geq 0$. This means that player i can not tell in v the coordinates $\{j_{v_k}\}_{k=1}^{\infty}$ of his opponent. Since we chose terminating expansions, player i has infinitely many zeroes along his type. Denote

$$C_j = \{v = (r, 0.a_1a_2\ldots, 0.b_1b_2\ldots) \in V : \text{the sequence } \{j_{v_k}\}_{k=1}^{\infty} \text{ converges}\}$$

(Of course, $\{j_{v_k}\}_{k=1}^{\infty}$ converges when it is eventually constantly 0 or constantly 1).

Now, refine $\Pi^i(v)$ in Ω with the event C_j: Reveal to player i whether the sequence of digits in the type of j that he can not observe converges or not. This means splitting $\Pi^i(v)$ to

$$X = \Pi^i(v) \cap C_j$$

and

$$Y = \Pi^i(v) \backslash X.$$

We will be done once we prove that with the new partition each $\omega \in \Omega$ is still a model for every formula ψ it contains. We prove this claim by induction on the structure of ψ, verifying for every fixed ψ that the claim holds $\forall \omega \in \Omega$. Without loss of generality we can assume that ψ is in disjunctive normal form (i.e., that ψ is a disjunction of conjunctions, where each conjunct is either the atomic formula φ or its negation, or the formulas $k^a\psi'$, $k^b\psi'$ or their negations, where ψ' is also in disjunctive normal form. It is straightforward to check that every formula $\bar{\psi}$ is equivalent to a formula ψ in disjunctive normal form, and that $\bar{\psi}$ holds in a state if and only if ψ does).

The claim clearly holds for the atomic formula φ and its negation. Assume, by induction, that the claim holds for ψ and ψ'. Then it clearly holds for $\psi \wedge \psi'$ and $\psi \vee \psi'$ as well. It also holds for $k^j\psi$ and $\neg k^j\psi$ for the other player j, since we did not alter his partition on Ω. If $\omega \notin \Pi^i(v)$, the claim also holds for $k^i\psi$ and $\neg k^i\psi$, since we did not change the partition member of player i in ω.

It remains to check the case $\omega \in \Pi^i(v) = X \cup Y$. If $k^i\psi \in \omega$, then ψ belongs to every $\omega' \in \Pi^i(\omega) = \Pi^i(v)$, so this certainly remains true for every $\omega' \in X$ and $\omega' \in Y$. Finally, if $\neg k^i\psi \in \omega$, then also $\neg k^i\psi \in v$. This means that there is a $v' \in P^i(v)$ in V such that $\neg\psi \in v'$. If the depth of $\neg\psi$ is n then $\neg\psi$ belongs also to all the other

$$v'' \in P^i(v)$$

such that in v' and v'' the first coordinate r is the same and the types of the players have the same first n digits. *The question whether v'' is in C_j is not determined by any finite number of digits in the type of player j in v''*; and the type of player i in v has infinitely many 0-s, so some of the v'' he considers as possible are in C_j, i.e. in X, and some not in C_j, i.e. in Y. Hence, if ω is either in X or in Y, in both cases there is a state v'' in the new partition member that contains ω to which $\neg\psi$ belongs. This completes our proof.

Remark. Under the assumption in the theorem, Ω has the cardinality of the continuum (see Aumann (1989), or a simpler proof in Hart, Heifetz and Samet (1996)), so there are altogether $|2^{[0,1]}|$ different pairs of partitions on Ω. The theorem shows that the cardinality of the partition pairs that are coherent with the structure of the states has the very same cardinality, which is hence the largest possible.

What may we conclude from this state of affairs? Could not one construct a canonical model where the partitions of the players would be uniquely determined by the inner structure of the states? For Harsanyi (1967–68) type spaces, where the uncertainty of the players takes the form of a σ-additive probability distribution, such a universal space does exist (Mertens and Zamir (1985)). Intuitively, the source of the difference between the two cases is that with σ-additive beliefs, the beliefs regarding "limit" or "tail" events like C_j are already determined by the beliefs regarding finite-order mutual uncertainties. Partition spaces, on the other hand, lack such a strong connection between the limit uncertainties and the finite-order uncertainties.[2]

All is not lost, though. Heifetz and Samet (1993) show that every (non-redundant) partition space may be isomorphically embedded as a subspace in a suitable canonical space, where the restriction of the players' partitions to this subspace is indeed uniquely determined by the inner structure of the states. In this canonical construction one may need to specify inside the states *transfinite* levels of mutual uncertainties of the players. In Heifetz (1994) we show that this is essentially equivalent to constructing canonical models for extended S5 logic, that allows for infinite conjunctions and disjunctions in the language.

Both approaches show, however, that there is no a priori ordinal bound on the number of levels of mutual uncertainties that have to be explicitly specified inside the states, in order to yield unique partitions that agree with that inner structure. Furthermore, for every given ordinal level of specification, such uniqueness can only be attained in a subspace of the construction. Partition spaces are, therefore, complex indeed.

Appendix: A reminder on multi-player S5 logic

The material below is standard in the literature, and may be found e.g. in Aumann (1989), Halpern and Moses (1992) or Chellas (1980, for the "one-player" case).

In what follows I is a fixed set of players, and A is a fixed set of atomic formulas. The logical language is the least collection that contains A, such that if φ and ψ are formulas in the language, so are $\neg\varphi$, $\varphi \wedge \psi$ and $k^i\varphi\ \forall i \in I$. k^i is the knowledge modality of player i, so $k^i\varphi$ is the formula "player i knows φ". As usual, $\varphi \vee \psi$ stands for $\neg(\neg\varphi \wedge \neg\psi)$, and $\varphi \to \psi$ for $\neg\varphi \vee \psi$.

The depth of formulas, denoted $dp(\varphi)$, is defined inductively by the following rules:

1) $dp(\varphi) = 0$ for atomic φ.
2) $dp(\neg\varphi) = dp(\varphi)$
3) $dp(\varphi \wedge \psi) = \max(dp(\varphi), dp(\psi))$
4) $dp(k^i\varphi) = dp(\varphi) + 1$

The axioms and inference rules of the logic consist of any axiomatization of the propositional calculus including Modus Ponens

$$\frac{\varphi,\ \varphi \to \psi}{\psi}, \tag{MP}$$

together with the following axiom schemes and inference rule:

[2] This was also observed by Fagin, Halpern and Vardi (1991).

$$k^i\varphi \to \varphi \tag{T}$$

$$k^i(\varphi \to \psi) \to (k^i\varphi \to k^i\psi) \tag{K}$$

$$k^i\varphi \to k^ik^i\varphi \tag{4}$$

$$\neg k^i\varphi \to k^i\neg k^i\varphi \tag{5}$$

$$\frac{\varphi}{k^i\varphi} \tag{RN}$$

A *proof* of a formula φ is a finite sequence of formulas that terminates with φ, each of whose elements is either an axiom or inferred from previous formulas by an inference rule. In such a case we say that φ is a *theorem* and write $\vdash \varphi$.

We say that a formula φ is *deducible* from a set of formulas Γ, written $\Gamma \vdash \varphi$, if there is a finite sequence of formulas that terminates with φ, each of whose elements is either a theorem, belongs to Γ, or inferred from previous formulas by (MP) (but not by (RN)!). We say that Γ is *consistent* if one can not deduce from it a formula and its negation.

A *model* for the above described syntax is a space Ω with a partition Π^i of Ω for every player $i \in I$, together with an *interpretation function* f from formulas to subsets of Ω, such that

$$f(\neg\varphi) = f(\varphi)^c, \quad f(\varphi \wedge \psi) = f(\varphi) \cap f(\psi) \tag{$A.1$}$$

and $\forall i \in I$

$$f(k^i\varphi) = K^i(f(\varphi)). \tag{$A.2$}$$

Here, K^i is the knowledge operator on events $E \subseteq \Omega$

$$K^i(E) = \{\omega \in \Omega : \Pi^i(\omega) \subseteq E\}.$$

The intuition behind this definition is that when ω occurs, player i considers as possible exactly the states in $\Pi^i(\omega)$. Thus, $K^i(E)$ is the event where player i is sure that the prevailing state belongs to E.

clearly, it is enough to define the interpretation function f for the atomic formulas. Inductively, (A.1) and (A.2) determine uniquely how f should be defined for every other formula in the language.

We say that a formula φ holds or obtains in $\omega \in \Omega$ when $\omega \in f(\varphi)$. We write $\models \varphi$ when φ holds in every state of every model. The S5 logic is determined by the class of partition space models:

Theorem. *For every formula* φ

$$\models \varphi \iff \vdash \varphi.$$

The standard way to prove the implication from left to right ("completeness") is by constructing the canonical model Ω whose states are the maximally consistent sets of formulas. Denote

$$[\varphi] = \{\omega \in \Omega : \varphi \in \omega\}.$$

The partition of player i on Ω is defined by

$$\Pi^i(\omega) = \{\omega' \in \Omega : k^i\varphi \in \omega \Rightarrow \varphi \in \omega'\}.$$

One then proves that this is indeed a partition, and that with the map

$$f(\varphi) = [\varphi]$$

one gets a model for the language. This means that $\models \varphi$ implies in particular that $[\varphi] = \Omega$. Hence φ is provable from the axioms – otherwise we could have built a maximally consistent set of formulas by adjoining to $\neg\varphi$ consecutively each formula in the language or its negation.

References

Aumann R (1976) Agreeing to disagree. Annals of Statistics 4:1236–1239

Aumann R (1989) Notes on interactive epistemology. Cowles Foundation for Research in Economics working paper Revised version in this issue

Chellas BF (1980) Modal logic, an Introduction. Cambridge University Press

Fagin R, Halpern YJ, Vardi MY (1991) A model-theoretic analysis of knowledge. Jour. of the ACM 91:382–428

Halpern JY, Moses YO (1992) A guide to completeness and complexity for modal logics of knowledge and beliefs. Artificial Intelligence 54:319–379

Harsanyi JC (1967–68) Games with incomplete information played by Bayesian players, parts I, II, & III. Man. Sc. 14:159–182, 320–334, 486–502

Hart S, Heifetz A, Samet D (1996) Knowing whether, knowing that, and the cardinality of state spaces. Journal of Economic Theory 70:249–256

Heifetz A (1994) Infinitary epistemic logic. In: The proc. of the 5th Conference on theoretical aspects of reasoning about knowledge, pp. 95–107, Morgan Kaufmann, California, pp. 95–107; Published in Mathematical Logic Quarterly 43:333–342

Heifetz A, Samet D (1993) Universal partition structures. IIBR working paper 26/93, Faculty of Management, Tel Aviv University. Published in two parts: Knowledge Spaces with Arbitrarily High Rank, Journal of Economic Theory 80:260–273, Hierarchies of Knowledge: An Unbounded Stairway, forthcoming in Mathematical Social Sciences

Mertens JF, Zamir S (1985) Formulation of Bayesian analysis for games with incomplete information. Int. J. Game Theory 14:1–29