3

Robert Aumann

Professor of Mathematics

Center for the Study of Rationality

and the Department of Mathematics

The Hebrew University of Jerusalem, Israel

Interviewed on January 23rd, 2010, in Bonn, Germany. The editors would like to thank Prof. Dr. Hildenbrand for his hospitality on this occasion.

O.R.:

Professor Aumann, many thanks for accepting our invitation to give us an interview. First question: What initially drew you to epistemic logic?

R.A.:

It was a natural outgrowth of work in game theory. The beginning came with work on common knowledge, the Agreement Theorem. This states that when two people have common priors and common knowledge about each other's probability assessments, then these probability assessments must be the same. The moment you start talking about common knowledge you get drawn into all kinds of epistemic considerations. The question is, what led to *that*? What led to the interest in common knowledge?

I wrote a paper in the very early '70s called 'Subjectivity and Correlation in Randomized Strategies'.[1] Randomised strategies go back to von Neumann and Morgenstern. These are mixed strategies based on some kind of coin toss, roll of the dice or spin of the roulette wheel. On the other hand there was the work of Savage, going back to Ramsey and de Finetti. This concerned *subjective*

[1] *Journal of Mathematical Economics*, 1(1), pp. 67-96, 1974.

probabilities, which started to bloom in the '50s and early '60s. So the next step seemed obvious: what happens when you try to apply subjective – rather than objective – probabilities to randomisation? This question led to the paper about subjectivity and correlation in randomised strategies, which was also the beginning of *correlated* equilibria. In this paper, I show that you can have subjective mixing and that subjective mixing leads to different equilibria, called in the paper *subjective* equilibria. You even get different pure equilibria from those you get with just objective mixing. At the end of the paper I discuss how this can occur. Why would people have different subjective evaluations? At that time it was an open question.

One summer I was at the Stanford Institute for Mathematical Studies in the Social Sciences. Sitting with Kenneth Arrow and Frank Hahn in Frank's office, we were discussing the question and I recall that the idea somehow arose from that discussion. I went back to my office and started thinking about it and that's how the agreement theorem was born. At the time it actually seemed too simple to publish. As you know, the proof only takes two lines; but I think the main thing is the *formulation* of the idea: it's not the proof itself. The proof seemed so simple that it somehow seemed beneath my dignity to publish it. I explained the theorem to Arrow and Hahn, and at first they didn't want to believe it. Finally they became convinced, and then actually convinced me that this was worth publishing. It became one of the two papers in my oeuvre that are cited most widely, and that was the beginning of my interest in epistemic logic.

Then one tries to develop, to get at the foundations, and this is how I became especially interested in the *syntactic* part. One question that is very puzzling to people who read the Agreeing to Disagree paper is: where does the partition *come from*? Somehow the knowledge partitions are assumed to be known to the players. In fact they're *common* knowledge to the players. This is built into the system. Where does it come from? These are the questions that most often arise with regard to that paper. How do people know each other's partitions? Why is this sort of given? Why is the partition a *given*? And that got me interested in the syntactic angle. In some sense, when you talk about partition spaces, when you talk about a semantic representation in epistemic logic, you want to say: Hey, let's put it in English! You are making some kind of assertion about agreement or whatever it is that you're talking about. Establish this assertion and don't give me partition spaces and state spaces and stuff like that. Say it in English! And that's what the syntactic approach is all about: saying it in English.

I became obsessed with the idea of *deriving* a semantic representation from *plain English*, which is what syntactics is all about. 'Syntactic epistemic logic' is just saying it in English. Dov Samet, an outstanding expert in epistemic logic, especially interactive epistemic logic, who did a PhD with me, put it this way, very picturesquely: If you want to explain it to your mother, you had better say it syntactically. Your mother won't understand if you try to say it semantically.

That sounds right. In semantics there's more grease, which allows you to work more easily. But you don't know where the foundations come from. Syntactic logic is more awkward, but it is simply logic. You derive one thing from another. You have certain axioms and you can argue about the axioms if you wish, but at least one thing follows from another. You don't have these sets and states of the world. What is a state of the world, anyway?

This is where the idea arose to view the semantic approach as *deriving* from the syntactic approach. You build a semantic approach *from* the syntactic approach.

This is how the whole thing started. I never see something and say: hey, this looks interesting, let's work on that. That's not the way it works. Rather, you have some problem and you say: we need to develop a tool to solve it. It is not 'this looks interesting and let's work on it'. People often ask me, why did I go into game theory? Well, I was faced with a problem, a specific problem that called for the use of game theory. So I said 'We have to study games.' It was a very specific problem. It's not like when you're a student at the university, you say: this subject looks attractive, that subject looks attractive, let's go into that subject. At least it was never that way in *my* research. Research is always very specifically purpose-oriented; one thing leads to another.

O.R.:
How did the semantic representation of epistemic logic come about, at the very beginning, at the origin of the Agreeing to Disagree paper?

R.A.:
I don't remember *exactly* the genesis of it, but it comes from Harsanyi's work. A type structure is essentially a partition structure. I think Harsanyi didn't think of it in *quite* that way, in terms of a state space with a partition. I don't know exactly how I came across that. I think it's already in the paper about subjec-

tivity and correlation, in the first issue of JME.[2] So I think it was '72, or thereabouts, fooling around with the Harsanyi type-space structure, that led to the structure of states and partitions. That's where it comes from; it's already in there. 'Agreeing to Disagree', as I said, sort of grew out of the problem left open in that paper about subjectivity and correlation. So it was a natural progression. It's a *restatement* of the Harsanyi structure. Not a very deep restatement but some sort of restatement. A type determines what each person thinks about the other person's types and payoffs. So everything is an n-tuple of types, and a slight restatement gives you a partition structure.

O.R.:
After Agreeing to Disagree, there's been a convergence between the work that you initiated on the semantic side about partition structures, and the work that was done by David Lewis in philosophy, for example. How did this convergence come about?

R.A.:
The David Lewis matter is really very, very interesting. I'll tell you my side of the story.

I wrote this paper and called this concept 'common knowledge'. This was published in 1976. Now around 1979 I ran across a paper in a philosophical journal which quoted my paper, I think. I'm not sure, though. It certainly quoted Lewis's book *Convention*,[3] which was published in '69. Perhaps it discussed a citation from my paper, too: I don't remember now, but it certainly quoted Lewis's book. I opened my eyes and said: 'Hey! What's going on here? This chap had the concept of common knowledge already in '69?' And, yes, it turned out that he did! I went back and bought the book or took it out of the library, and there it was. In 1969! And the amazing thing was that we used the *same word* for it, the same word that I thought I had invented. So from then on, when I wrote about this I started quoting Lewis.

Now, a year or two later Lewis sent a letter to the provost of Princeton University. In his PhD thesis, Sergio Ribeira DaCosta-Werlang had quoted me on common knowledge. Somehow the thesis was passed on to David Lewis as a reader. Lewis was furious because the thesis didn't quote him at all. It made believe that common knowledge was my concept. So I think Lewis wrote a let-

[2] *Journal of Mathematical Economics*, 1(1), pp. 67-96, 1974.

[3] *Convention: A Philosophical Study*, Cambridge, MA: Harvard UP, 1969.

ter to the provost in which he said ... well, he complained about this in fairly strong language. I don't remember the exact language. He sent a copy to Hugo Sonnenschein and a copy to me. So I got this copy and immediately wrote back saying: 'you know, Prof. Lewis, this is your concept. No question about it. I was not *aware* of your contribution to this when writing the '76 paper. I simply did not know. You might find this difficult to believe because we use the *same word*, we use the word 'common knowledge'. But it's true. And as soon as I became aware, which is a year or two ago now, I started citing you. Please accept my abject apologies. I'm not even saying that this is independent work. You can talk about independent work when you're talking about something that is done at approximately the same time or even a year or two later, but not when there's a hiatus of seven years. Because, you know, people talk at lunch or they talk in seminars and things filter through, very often without attribution. Both the idea itself and the name could have filtered through somehow without my being aware of it. So I'm not claiming independence. It's your contribution. The idea of common knowledge is your contribution and there's no question about it. This is all yours. I've been saying it for a while now, and I'll say it again. And let's meet.'

I was in Stony Brook at that time. I went down to Princeton, and I met Lewis and Kripke. We had a nice conversation. By that time there was a kosher dining club in Princeton. I'll have to tell you some stories about Princeton when I was originally there in the fifties. It was a hotbed of anti-Semitism. But you know, the world progresses and by 1981–82 there was a kosher dining club and in fact Hugo Sonnenschein became provost of Princeton. He's Jewish, of course. And then afterwards there was even a Shapiro who became President of Princeton University. He is also Jewish. So we're making progress. I had a nice day in Princeton and made it up with Lewis and Kripke, and everything was fine. And at every opportunity, including this one, I've been acknowledging that the idea of common knowledge is Lewis's idea. Period. No independence, nothing.

However, let me just add this: Lewis had the idea of common knowledge, and this is very clear. What he did *not* have is the agreement theorem. He did not have the agreement theorem, he had no glimpse of the agreement theorem, nothing like that. So I still lay claim to the agreement theorem.

O.R.:

You met both Lewis and Kripke. Were you already interested at that time in the syntactic part of epistemic logic?

R.A.:

I'm not sure I got into the syntactic thing at that point. No, I don't think so, no. I was bothered by this question of where the partition structure comes from. That was in the early '80s: it bothered me even then. But I sort of finessed it. I said, well, *descriptively*, not *formally*: a state of the world includes everything that is relevant about that state, including what people know about that state. It's rather like a dictionary. It's nothing substantive, it's like a dictionary. And that statement occurs in my '87 paper 'Correlated Equilibria as an Expression of Bayesian Rationality'.[4] But actually to *describe* that dictionary, to do that and carry through this idea formally, I didn't see how to do it. I didn't see how to do it and therefore I can't say I was actively looking for it. I don't think so.

And then came a paper by Samet,[5] whom I have already mentioned. That's where I got this idea that we have a syntax, we have a set of statements, a list of statements; this list has to be complete and consistent. By complete we mean that every sentence – either the sentence itself or its negation – is in the list. Consistent means there are no logical contradictions in the list. A beautiful idea. You take a language and you say everything that can be said in that language. Everything. And then you look at lists, complete and consistent in this sense. And you say: that list *is* a state of the world. It's a beautiful idea. And this came out of Samet's paper.

I took that, developed it and formalised it. In some respects it was already formal. The fruition of this was the paper which was published in the *International Journal of Game Theory* in 1999, 'Interactive Epistemology',[6] which takes that and actually builds a semantic space out of this 'dictionary'. I was at Yale University for six weeks in '89, where I gave a course on interactive epistemology, and presented that idea. It took another ten years to publish it because I'm very slow to publish. Other stuff is going on and there's family and grandchildren, all kinds of stuff happening. It took ten years, but eventually I did publish, basically, the notes from the six weeks in '89 at Yale on interactive epistemology. So

[4] *Econometrica*, 55(1), pp. 1-18, 1987.

[5] D. Samet. Ignoring Ignorance and Agreeing to Disagree. *Journal of Economic Theory*, 52, pp. 190-207, 1990.

[6] *International Journal of Game Theory*, 28(3), pp. 263-300, 1999

that's how that came about. I'm convinced now that semantic game theory is fine, but the syntactic is more basic. Say it in English! If you have something to say, say it in English.

O.R.:
What is the relevance of epistemic logic? To what kind of topic is it best geared to contribute?

R.A.:
It's for exploring the foundations of game theory. You want to ask yourself: why do we look only at the objective mixed equilibria? Then one thing leads to another, you get the agreement theorem and then you have to ask yourself where this comes from. It's foundational work. It also develops a toolkit that enables you to think precisely about these things. It allows you to formulate precise questions.

Let me give you another application. I published this paper in 1995, 'Backward Induction and Common Knowledge of Rationality'.[7] People raised their eyebrows about it, and said: this is open to question, because according to Aumann's definition of rationality, people also have to behave rationally at nodes of the game tree that they themselves have excluded by previous actions. So they know they're not going to get there, but they still have to behave rationally. I wrote a lengthy discussion in that paper and I do think it makes sense and I think people should read the discussion part very, very carefully. Part of what ails the world is that people only look superficially at a paper. I really thought carefully about these problems. And I think the criticism is *not* justified.

But I always like to look at the other side of the coin. I said, OK, people don't like this idea that you have to behave rationally at nodes of the game tree that you yourself have excluded. They call these 'counterfactual conditionals'. In order to avoid using counterfactual conditionals, Adam Brandenburger and I replaced the idea of *knowledge* of something by *belief*, by the idea of *strong* belief. This means I continue to believe that everybody is rational until it becomes *logically* impossible to do so, until it's obvious we've reached a node of the game tree where somebody had to behave irrationally. And then we developed this idea of common strong belief of rationality. It seemed intuitive that this should lead to backward induction. Adam and I worked on this for five or six years and we couldn't prove it.

[7] *Games and Economic Behavior*, 8(1), pp. 6-19, 1995.

Then one day a paper crossed my desk written by Battigalli and Siniscalchi.[8] They did exactly that. Fantastic! It is a wonderful piece of work; it's very deep and it's very roundabout, but they did it! It's very good.

Battigalli and Siniscalchi did it using a very complex *semantic* model. In a semantic model it's difficult to express what it means for something to be logically impossible. In a semantic model, something that is logically impossible is represented by an empty set; but if an event is represented by an empty set, that doesn't mean it's logically impossible. You have to have a *universal* semantic model to say that, and these models are large and clumsy.

Now I have a student who is just completing his doctorate. His name is Itai Arieli, a very bright and deep young man. He managed to overcome this difficulty with a syntactic proof. 'Logically impossible' has a very transparent meaning in a syntactic model. Itai has a straightforward – well, straightforward may not be the right word – but he has a largely syntactic proof of this which solves the problem without encountering the conceptual difficulties of the semantic approach.

How does this relate to your question? The syntactic and the semantic models provide a *toolkit* for thinking about these problems. You want to fix some system, you need the screwdrivers. This is the toolkit for dealing with these problems and for attacking them, and thinking about them logically.

O.R.:
So if I understand correctly, you see this work in the characterisation of solution concepts as foundational for game theory. Could you think of it in the same way as the work in the foundations of decision theory or subjective probability, for example?

R.A.:
I think maybe it's less important in one-person decision theory. It does not play much of a role there. Savage,[9] Anscombe-Aumann,[10] it's not *essential*, I don't see applications there.

But let me say it's not just foundational stuff. I wrote a paper with Jacques Dreze[11] about rational expectations in games.

[8] P. Battigalli and M. Siniscalchi, Strong Belief and Forward Induction Reasoning. *Journal of Economic Theory*, 106(2), pp. 356-391, 2002.

[9] L. J. Savage, *The Foundations of Statistics*, New York: John Wiley, 1954.

[10] F. J. Anscombe and R. J. Aumann. A Definition of Subjective Probability. *The Annals of Mathematical Statistics*, 34(1), pp. 199-205, 1963.

[11] R. J. Aumann and J. H. Dreze, Rational Expectations in Games, *Amer-*

In some sense this grew out of the epistemic approach, which as I say started with the paper in the JME about subjectivity and correlation. This is more or less the latest development there; not the latest development in foundational terms but in *practical* terms.

In this paper with Jacques what we're really trying to promote is a revolution. We haven't been too successful yet. We're saying: Listen! Equilibrium is not the way to look at games. Nash equilibrium is *king* in game theory. Absolutely king. We say: No! Nash equilibrium is an interesting concept, it's an important concept, but it's not the most basic concept. The most basic concept should be: *to maximise your utility given your information*, in a game just like in any other situation. What does that imply when you make strong assumptions like common priors and common knowledge of rationality? Does it imply Nash equilibrium? No, it does not! It implies something else, and we can write formulas for it. That's the way to go.

In some sense the epistemic world of ideas led to this. This work started with a remark Jacques made. Jacques visited Israel in '96, and we went on a trip to Petra in Jordan with a whole group of economists who were visiting Israel at that time. I was sitting next to Jacques in the bus and he said to me: 'Bob, how would you define values in games that are not zero-sum?' We started thinking about it and a year or two later I went to Louvain-la-Neuve. We thought about it some more and eventually it grew into this paper. So in one sense the *question* didn't come from epistemic roots, but the *answer* came from epistemic roots. In two-person zero-sum games, common knowledge of rationality and common priors imply the value. That's it. You don't need the equilibrium concept. This gives a solid foundation to the value of two-person zero-sum games, which was missing before. Then you take that and you extend it to other games. I think that should replace Nash equilibrium as the fundamental concept of game theory. But it hasn't yet...

So epistemic considerations also lead to conclusions *outside* their own domain, which is the test of any important scientific theory or development. It's got to lead some place outside, not just inside.

O.R.:
In this paper with Dreze you make a strong case for analysing interaction in terms of common priors and common knowledge of rationality. Do you think these are fundamental assumptions?

ican Economic Review, 98(1), pp. 72-86, 2008.

R.A.:
Yes, I do. Let me put it this way: common priors is in my opinion a very reasonable assumption, *more* reasonable than common knowledge of rationality. People usually take it the other way around. They think that common knowledge of rationality is something they can buy; it's a strong assumption but it's something they can buy. But common priors? They ask 'Where did these come from?' *I* think common priors are more reasonable, because in some very basic sense you want somehow to account for differences in probability assessments. Why do people differ in their probability assessments? It seems reasonable that at *some* level it's because they had different inputs. So they got to know different things, and this is what accounts for the different posteriors. When you say that differences in probability assessments are due to different information, that is the common prior assumption. So this is really something very basic and very important.

Common knowledge of rationality is not that kind of thing at all. It's a very far-fetched assumption. I don't even know if *I'm* rational, but let's say I am. I'm willing to assume you're rational, but do I *know* that you *know* that I'm rational? It's not that common knowledge of rationality is something basic; it's some sort of *benchmark.*

The way we put it in the paper is: tic-tac-toe is a draw. What does that mean, that it *is* a draw? What it means is that under common knowledge of rationality the result will be a draw. It doesn't mean that a smart kid would not lose the game, because a smart kid might play for a win, or try, because that kid does not assume the other one knows that he is rational and so might play for some kind of trap and might lose as a result. Common knowledge of rationality is not something about the real world, it's a sort of a benchmark. It says: this is the way things work out when there's no irrationality in the system. If there is no irrationality in the system at all, this is how things work out. Everything else is calculated as a *departure* from the benchmark. It's like a no-friction assumption in physical systems, or the no-air-resistance assumption. When you have the formula for acceleration of a body under gravity, you let something fall, it assumes no air resistance. But there *is* air resistance. Does that make the formula for no air resistance unimportant? No, it's very important; you start out with the assumption that under *ideal* conditions, you have no air resistance, and then from there you can go and calculate what happens *with* air resistance. I don't think common knowledge of rationality is at all a reasonable assumption about the real world. I believe common priors *is* a reasonable assumption about the real world. Common priors is much more basic than CKR. But if you don't buy CKR, you won't get the value of a two-person zero-sum game.

So in order to answer Jacques' question, we do need CKR because he wants to generalize the value. The value of the game is the outcome when people play *correctly.* So when people play correctly, you have CKR in some sense. That's the *definition* of correct play.

O.R.:
You mention in that paper that the notion of value was something that was more or less forgotten but it returns when you look at a game from the perspective of common priors and common knowledge of rationality. Do you think there are other topics, especially in epistemic logic, that have been somewhat relegated to the background and should be given more prominence?

R.A.:
One should probably try to look at relaxing these conditions of CKR or of common priors. Shmuel Zamir spoke with me about 'approximately common priors'. One could try to look at rational expectations in specific classes of games. I can't think at the moment about something really foundational but my research, as far as I've experienced, never worked that way. Something seizes you and then you go in that direction. If you don't, then it hasn't seized you. Some kind of approximation to these ideal conditions is one question that can be asked. It's important to keep the application in mind, keep your purpose in mind. But to point to something specific, no, I don't know.

O.R.:
You mentioned the common knowledge of rationality assumption as a benchmark. Would you think the same of the assumption or the fact that agents in most epistemic logic systems are so-called logically omniscient?

R.A.:
Logical omniscience, that is a very important problem. It's a problem I've thought about a little. Not that I've gotten anywhere with it. I really don't know what to do with it. It is a very important problem, which does need treatment, but I don't know what to do with it.

O.R.:
Do you think recent work on unawareness could...

R.A.:
Unawareness, yes, those two: logical omniscience and unawareness, those are the two big open problems as far as I know. I think there's been work on them recently, though, but I haven't studied it. *Mea culpa.* I cannot venture an opinion as to whether the recent work really gets to the foundations of the problem. But those *are* the two outstanding problems. Maybe somebody has solved them in the meantime. Unawareness and logical omniscience: even though I am not omniscient – I do not claim to be omniscient – I'm unaware of satisfactory solutions to those two problems. But they are two very important problems, which do cry out for satisfactory treatment.

In that connection maybe I should mention something else which is maybe not particularly logical, but it has to do with logic. That is the problem of how much to calculate when you're optimising something. Most work that I know of, all the work I know of, does not take the cost of calculation into account. So the question is, how much do you calculate? If you really want to optimise the solution to something, how much should you calculate in order to find that solution? You want to get the cheapest possible solution without knowing what to expect. In order to even calculate the *time* it would take to optimise, you have to do a calculation, and that calculation in itself may not be worthwhile. When you're playing chess and the clock is ticking, how much do you calculate? It's not at all clear. Deep Blue had to solve this problem somehow. I don't know how they solved it. This is a very important problem, which has not been solved satisfactorily. It's in the same league as unawareness and logical omniscience.

O.R.:
So you say that the big issue or the big challenges for us now are about relaxing the fundamental assumptions?

R.A.:
No, I think logical omniscience and unawareness and how much to calculate, those are the three big outstanding issues at this time. The matter of relaxing, I don't think it's *that* important. It's not that basic. Those other things are more basic and I think they'll be more difficult to achieve. Relaxing is also relaxing the focus, in some sense: you see things more blurred. I like focus.

O.R.:
Are you optimistic about progress on these three questions?

R.A.:
Who knows? *Que sera, sera.* Those are Hilbert's three problems for epistemic game theory or logic. The third question is not really epistemic but it's also there and it has philosophical sides to it. But the unawareness and the logical omniscience, absolutely. Whether I'm optimistic, I don't know. It will be very nice if somebody solves them.