Φ

Name (to appear in publication): Yoav Shoham

Position (to appear in publication): Professor of Computer Science

Affiliation (to appear in publication): Stanford University

The 5 Questions

1. Why were you initially drawn to epistemic logic?

"Why" is hard but "when" is easy. It was in 1986, close to the end of my graduate studies in computer science at Yale. I was invited to the first TARK conference (it then stood for Theoretical Aspects of Reasoning about Knowledge; we later changed it to Theoretical Aspects of Rationality and Knowledge, reflecting the broader scope that emerged but retaining the TARK brand…). In time I became quite involved with TARK, including serving as program chair in 1994, an invited speaker in 2009, and member its board of directors for much of its existence. But like a first love, that first conference was a formative experience for me. It was wonderful to see philosophers, computer scientists, game theorists, psychologists and other sorts convene around a shared topic and a nascent theory. The beauty of the theory left a deep impression on me, as did some disturbing questions it raised. Both the beauty and the puzzles kept me explicitly actively for about a decade, during which time I published papers on knowledge, belief, and nonmonotonic belief revision. Apparently the issues continued to have a grip on me later as well, since after a hiatus of a decade, during which I devoted much of my time to non-epistemic issues at the interface of computer science and game theory, I now find myself drawn back to that material.

If I were to take a stab at the "why" question, I would mention two factors. The first is the magic of having three quite disparate fields – philosophy, game theory, and computer science – converge on the same formal notion of knowledge. Of course when you look closely at magic you see the sleight of hand; and clearly, computer science (and in particular artificial intelligence, or AI, where the logic of knowledge first appeared) explicitly borrowed from philosophical logic. But still the convergence is breathtaking, even in retrospect.

The other factor is the concrete distributed-systems interpretation of what otherwise are quite abstract models. A "possible world" is a global state of a system, a snapshot of it within an execution history. The global state of the system is made up of the local state of each of the processors. Two possible worlds are accessible from the perspective of a given processor if the process has the same local state in them. This obviously gives rise to the partition model of knowledge, and the S5 logic. Never mind that, on a closer look, S5 seems quite problematic as a model of knowledge (more on this later); it was remarkable to me that the "standard" logic of knowledge, which was proposed based on quite abstract reasoning, it fact

arose naturally out of a very concrete and straightforward computational model. I will return to this point shortly.

2. What example(s) from your work, or work of others, illustrates the relevance of epistemic logic?

As a computer scientist I suppose it is incumbent on me to describe computer science applications, so let me do it, briefly. Two applications immediately come to mind. The first I already discussed – distributed systems. Computer scientists have always searched for good models with which to reason about distributed systems. The typical questions revolve around robustness to failures of various kinds, including failure by a node in the system to correctly follow the prescribed protocol, and failure of the network to transmit messages (or transmitting it with unbounded delay). Intuitively, people would make statements such as "Well, processor A doesn't know whether processer B received its message, but processor B knows that processor A doesn't know, and so…" Logics of knowledge proved an excellent way of making such reasoning precise, and specifically the formal construct of "common knowledge" (everyone knows, everyone knows that everyone knows, …) completely characterized conditions necessary for coordinated action in the face of such system malfunctions.

Elsewhere (and earlier) in computer science the logic of knowledge was put into use in the context of AI planning. AI has a long tradition of planning research. In a typical scenario of so-called "classical" AI planning, the planner has some information about the state of the world, and some actions at its disposal, and must concoct a plan to change the world into some desired state by appropriate sequencing of the actions. Traditionally, each action has some preconditions (for example, you cannot grasp an object if you are already holding another object), which makes the problem computationally hard. Thirty years ago it was recognized that some preconditions are epistemic in nature; in order to open a safe you must know the combination of its lock. And, again, logics of knowledge proved just the ticket for modeling these preconditions. Later epistemic reasoning in planning included reasoning about epistemic post-conditions (after running a litmus test you know whether or not the substance is acidic), and coordinating the action of multiple robots with limited sensing capability.

3. What is the proper role of epistemic logic in relation to other disciplines, for instance mainstream epistemology, game theory, computer sciences or linguistics?

I will merge my comments on this topic into the answer to the next question. See discussion there of "the rules of the game".

4. Which topics and/or contributions should have deserved more attention in late 20th century epistemic logic?

I will list two technical topics, and one nontechnical. The technical topics include the connection between knowledge and belief, and iterated belief revision. The nontechnical item is a meta-discussion about the rules of the game. It is not that these topics did not receive attention; they did. But the discussion around them seemed to subside long before closure was reached, which makes it hard to build on those

foundations as we try to apply them and/or extend the domain of discourse.

There are two common slogans concerning the connection between knowledge and belief, both reducing the former to the latter.[1] The first is about knowledge being "justified, true belief". Critique of this slogan (which I do not share) notwithstanding, I am not aware of a successful, generally accepted technical embodiment of this slogan (the problem is in capturing the notion of justification; as a side note, I believe that this avenue has been under-explored, and that the hasty rejection of this slogan was based on a an example, due to Gettier, which adopted a sense of justification that is inappropriate in this context).

The other slogan is of knowledge as "stable belief", or, more verbosely, "belief that is stable with respect to the truth". The idea is that one has beliefs, which over time get revised by new (correct) information that is learned. Those beliefs that are never forced out by new (correct) evidence are elevated to the status of knowledge. Several related technical proposals were made to capture this intuition; a typical one posited a total preorder on possible worlds as the accessibility relation, leading to a "standard" KD45 model of belief, but a model of knowledge which is weaker than S5. Specifically, the negative introspection property of knowledge is jettisoned, and replaced by weaker conditions, such as those captured by the S4.3 logic (with S4.3 lying in-between S4 and S5).

These latter proposals seem to me quite compelling, and it's puzzling to me why the discussion around them never picked up. Even before these proposals, S5 had been critiqued as a model of knowledge, and it was proposed that somewhat weaker logics be considered, specifically those lying in the neighborhood of S4.2 and S4.3. The fact that an independently motivated model happens to produce independently argued-for properties of knowledge is already striking. But there are additional attractive properties of these models. Beside inducing appealing properties on both knowledge and belief, they induce quite natural connections between them (knowledge being stronger than belief, and various natural hybrid introspective properties involving both knowledge and belief; for example, if you know something you believe that you do). And furthermore, the total-preorder structure had been considered earlier in the context of belief revision and nonmonotonic logics. So it is remarkable that a model with all these attractive properties and correlation with other independent proposals did not take hold more strongly than it did, especially since (to my knowledge) no alternative connection between knowledge and belief was proposed. It seems to me this reflects fatigue and apathy on the part of the research community more than disagreement.

Belief revision has certainly received tremendous attention. In this attempt to capture the dynamics of beliefs, researchers have investigated normative theories of revising a logical theory with a newly arrived piece of evidence. The interesting part of the theory concerns the case in which the new evidence is inconsistent with the original theory. The dominant (though not uncontested) theory in this area consists of the AGM axioms, which, as is well known, are closely related to the total-preorder structure discussed above (or, equivalently, Groves' original "system of spheres"). The limitations of the AGM theory become apparent when one tries to iterate the revision operator. The AGM axioms have nothing to say about it, and it is not obvious how to extend the model theoretic analysis. Briefly, the model theoretic account of AGM takes as input two arguments, a belief state (essentially, a total preorder on worlds) and a belief (a set of worlds). For the static theory, it is enough that the operator produce another set of worlds; that is enough to define the new beliefs. But if one wants to iterate the operation, the operator must produce an entire new belief state. There has been a continuous strand of literature on this topic, though not nearly as voluminous as in the 1980s. Opinions diverge on the state-of-the-art here. Some (mostly from AI) have argued that we have reached a good understanding of iterated belief revision. Others (including a well known philosophical logician) have argued that we have not. I tend to side with the latter. In any event, in such situations the nay sayers have an unfair advantage; if some people claim there

[1] For the record, I am aware of at least one suggestion in the opposite direction, defining belief as "defeasible knowledge", and an accompanying technical proposal, but I will not dwell on it.

is universal agreement, and others claim there is not, then there is not.

Finally, a word about the rules of the game, and about methodological differences among fields. Let me focus on philosophy and computer science. There is in certain philosophical circles an established tradition, when considering formal theories of everyday concepts, of relying on particularly instructive examples as litmus tests. The "morning star -- evening star" example catalyzed discussion of cross-world identity in first-order modal logic. Closer to home, the example of believing that you will win the lottery and coincidentally later actually winning it served to disqualify the definition of knowledge as true belief. A similar example purported to disqualify defining knowledge as *justified* true belief (though, as I mentioned earlier, I don't think it did).

Such "intuition pumps" can be highly instructive, but the question is what role they play. In the above-mentioned philosophical circles they tend to serve as necessary but insufficient conditions for a theory. They are necessary in the sense that each of them is considered sufficient grounds for disqualifying a theory (namely, a theory which does not treat the example in an intuitively satisfactory manner). And they are insufficient since new examples can always be conjured up, subjecting the theory to ever-increasing demands.

This is understandable from the standpoint of philosophy, to the extent that it attempts to capture a complex, natural notion (such as knowledge) it its full glory. But this necessary-but-insufficient interpretation means that the process of formalization is a never-ending one, since new requirements may surface at any moment.

There is an alternative, and logics of knowledge provide an example. The S5 logic of captures well certain aspects of knowledge in idealized form, but the terms "certain" and "idealized" are important here. The logic has nothing to say about belief (as opposed to knowledge), nor about the dynamic aspects of knowledge (how it changes over time). Furthermore, even with regard to the static aspects of knowledge, it is not hard to come up with everyday counterexamples to each of its axioms. And yet, as discussed, the logic proves useful to reason about certain aspects of distributed systems, and the mismatch between the properties of the modal operator K and the everyday word "know" does not get in the way, within these confines. All this changes as one switches the context. For example, if one wishes to consider cryptographic protocols, the **K** axiom (no relation to the modal operator K) – which is valid in any normal modal logic, and here represents logical omniscience – is blatantly inappropriate.

The upshot of all this is the following criterion for a formal theory of natural concepts: One should be explicit about the intended use of the theory, and within the scope of this intended use one should require that everyday intuition about the natural concepts be a useful guide in thinking about their formal counterparts. A concrete interpretation of the above principle is what elsewhere I have called the "artifactual" perspective. Artifactual theories attempt to shed light on the operation of a specific artifact, and use the natural notion almost as a mere visual aid. This is precisely the case in knowledge and distributed systems. Does this make the artifactual perspective uninteresting from the philosophical point of view? I think not, though of course that is for philosophers to decide. In any event, it is a useful from the perspective computer science, which brings me to my last observation: While we have an artifactual perspective on the standard (if somewhat discredited) "standard" logic of knowledge, we do not have such a perspective on belief. This is another direction under-explored in the late 20th century.

5. What are the most important open problems in epistemic logic and what are the prospects for progress?

The previous section laid out several issues, including two technical ones, that remain unsettled, and that are important in my view. Naturally they constitute part of the answer here. But I will focus on a different

direction, a vast open area in which many fascinating and important open problems lie. I am referring to the general program to develop logics of rational agency. These logics attempt to capture various facets of human mental state, and posit normative relationships among them (or "rational balance", to use Nilsson's term). Knowledge and belief belong to one category of metal state, which might be called "informational attitudes", capturing agents' assessment of whether this or that fact holds true. But informational attitudes constitute just one facet of mental state. In contrast, "motivational attitudes" capture something about the agents' preference structure, and "action attitudes" capture his inclination towards taking certain actions. In a typical theory, the action attitudes mediate between the informational and motivational attitudes; the agent's choice of action is dictates by his wants and beliefs. Into these two broad camps fall notions such as desire, goal, intention and plan. The literature on such attitudes is in comparison quite thin, and herein lies the opportunity.

There is a reason why progress in this area has been more limited. When one thinks about preference in isolation of other factors, things are relatively easy. There are certainly some interesting challenges -- for example, capturing "ceteris paribus" conditions in preference (I prefer wealth to poverty all other things being equal, but I prefer being healthy and poor to being sick and rich). But things become truly involved when one considers the interaction between the various types of attitude. For example, the dynamics of belief and preference are in general intertwined, with changes in beliefs leading to change in preference (and possibly vice versa). But more complex interactions are common. In a typical situation, one "intends" to take an action in service of a "goal" which was given rise to by certain "desires", and conditional on some " beliefs" (for example, intending to drive your car to the city is motivated by your belief that there is a good show there that evening, and is inconsistent with believing that the car is not in working condition). Beside the interaction between the different attitudes (belief, intention, goal, desire), the discussion involves even more basic aspects of agency, such as action and ability.

This is a complicated picture, and so it's not surprising that progress on this has been relatively slow. But slow progress does not mean no progress. The literature on so-called BDI theories (a tongue-in-cheek reference to belief, desire and intention) contains some inspiring and intriguing ideas and provides several starting points. I believe there is much low-hanging, and tasty, fruit in this area.