Epistemology: Five Questions

Joseph Y. Halpern
Professor
Computer Science Dept.
Cornell Unversity
Ithaca, NY
http://www.cs.cornell.edu/home/halpern

1. *Why were you initially drawn to epistemology (and what keeps you interested)?*

   I was initially drawn to epistemology because I enjoyed looking at puzzles that involved epistemic reasoning. Perhaps my favorite is the *muddy children puzzle*, which is a variant of the well known "wise men" or "cheating wives" puzzles [Gamow and Stern 1958; Gardner 1984; Littlewood 1953]. The following version is taken almost verbatim from [Barwise 1981].

   > Imagine n children playing together. The mother of these children has told them that if they get dirty there will be severe consequences. So, of course, each child wants to keep clean, but each would love to see the others get dirty. Now it happens during their play that some of the children, say k of them, get mud on their foreheads. Each can see the mud on others but not on his own forehead. So, of course, no one says a thing. Along comes the father, who says, "At least one of you has mud on your forehead," thus expressing a fact known to each of them before he spoke (if $k > 1$). The father then asks the following question, over and over: "Does any of you know whether you have mud on your own forehead?" Assuming that all the children are perceptive,

> intelligent, truthful, and that they answer simultaneously, what will happen?
>
> There is a "proof" that the first $k-1$ times he asks the question, they will all say "No," but then the k^{th} time the children with muddy foreheads will all answer "Yes."
>
> The "proof" is by induction on k. For $k = 1$ the result is obvious: the one child with a muddy forehead sees that no one else is muddy. Since he knows that there is at least one child with a muddy forehead, he concludes that he must be the one. Now suppose $k = 2$. So there are just two muddy children, a and b. Each answers "No" the first time, because of the mud on the other. But, when b says "No," a realizes that he must be muddy, for otherwise b would have known the mud was on his forehead and answered "Yes" the first time. Thus a answers "Yes" the second time. But b goes through the same reasoning. Now suppose $k = 3$; so there are three muddy children, a, b, c. Child a argues as follows. Assume that I do not have mud on my forehead. Then, by the $k = 2$ case, both b and c will answer "Yes" the second time. When they do not, he realizes that the assumption was false, that he is muddy, and so will answer "Yes" on the third question. Similarly for b and c.
>
> The argument in the general case proceeds along identical lines.

Let us denote the fact "at least one child has a muddy forehead" by p. Suppose that in fact there are 5 muddy children. Then even before the father speaks, each child knows p. So it would seem that the father does not provide the children with any new information. This suggests that the father he should not need to tell them that p holds when $k > 1$. But this is false. In fact, it is not hard to show that if the father does not announce p, the muddy children are never able to conclude that their foreheads are muddy (see [Fagin, Halpern, Moses, and Vardi 1995, Chapter 1] for details).

The key point here turns out to be that the father gives the children *common knowledge* of p. I found understanding the role of common knowledge, and how Kripke structures could be used to explain what was going on in this puzzle at so many levels—logical, psychological, and linguistic—absolutely fascinating.

While puzzles were perhaps what drew me into epistemology, it was the intimate connection between epistemology and other topics of arguably more practical concern, particularly distributed computing, AI, and game theory, that kept me interested. I found it wonderful how reasoning about the knowledge of agents in a system could give insight into so many problems in all these areas, from coordination to agreement to bargaining. But here too, it was thinking about puzzles that initially brought out the connection. Aumann's famous result that we can't agree to disagree [Aumann 1976] (somewhat more precisely, it cannot common knowledge among agents that have a common prior that they have different posteriors) brought out the importance of epistemic reasoning to game theory. But this leads to an obvious puzzle: how do agents still manage to trade stocks (which, after all, involves agreeing to disagree about the expected value of a stock, which is impossible if agents cannot have different posteriors). (Of course, an agent that needs to sell a stock to finance his child's university education may be willing to trade even if he agrees that the stock that he is selling will go up relative to the rest of the market, but there is still an issue as to how agents only interested in making a profit are able to trade.) This seemed to me a great area in which to do research.

As a computer scientist, even more influential was what is perhaps my second-favorite puzzle, the *coordinated attack problem.* This problem was originally introduced by Gray [1978] as an abstraction of database issues, and can be described informally as follows (the description is taken from [Halpern and Moses 1990]:

> Two divisions of an army, each commanded by a general, are camped on two hilltops overlooking a valley. In the valley awaits the enemy. It is clear that if both divisions attack the enemy simultaneously they will win the battle, while if only one division attacks it will be defeated. As a result, neither general will attack unless he is absolutely sure that the other will attack with him. In particular, a general will not attack if he receives no messages. The commanding general of the first division wishes to coordinate a simultaneous attack (at

> some time the next day). The generals can communicate only by means of messengers. Normally, it takes a messenger one hour to get from one encampment to the other. However, it is possible that he will get lost in the dark or, worse yet, be captured by the enemy. Fortunately, on this particular night, everything goes smoothly. How long will it take them to coordinate an attack?

Suppose that the messenger sent by General A makes it to General B with a message saying "Let's attack at dawn". Will general B attack? Of course not, since General A does not know he got the message, and thus may not attack. So General B sends the messenger back with an acknowledgement. Suppose that the messenger makes it. Will General A attack? No, because now General B does not know that General A got the message, so General B thinks General A may think that he (B) didn't get the original message, and thus not attack. So A sends the messenger back with an acknowledgement. But of course, this is not enough either.

What is going on here is that each time an acknowledgement is received, the level of knowledge increases by one. However, the generals never achieve common knowledge of the message. As Yoram Moses and I showed [Halpern and Moses 1990], common knowledge is required for coordinated attack (and, more generally, for coordination of all types). Moreover, in a system where communication is not guaranteed (roughly speaking, one where there is always a possibility that a message sent might not arrive), agents can never attain common knowledge of a fact that wasn't common knowledge to start with. It easily follows that coordinated attack is impossible in systems where communication is not guaranteed.

The fact that reasoning about knowledge could provide such insights was to me very exciting.

2. *What do you see as being your main contributions to epistemology?*

   I started working on epistemic logic in 1983 or so, with Yoram Moses, who was then my Ph.D. student. (It's hard to believe that it was really 25 years ago!) I found much of the work in the philosophy literature in epistemology at the time quite frustrating. The focus seemed to be on finding the "right, true properties" of knowledge. The implicit picture seemed to be that we all, at heart, understood what "knowledge" was; it was the philosopher's job to explicate the notion clearly. So there would be debates about whether, for example, it was "really" the case that knowledge satisfied S4; was it really true that if you knew something, you knew that you knew it? Much of the debate was relatively informal. In some of the papers, it seemed that the definition of knowledge changed from paragraph to paragraph. (Perhaps fortunately, I've forgotten exactly which papers in this literature I found particularly annoying!) It seemed to me that there was no one "right, true" notion of knowledge. People use the word in many different ways. In any case, my interests were more pragmatic.[1] Although philosophers going back to Hintikka had argued that S5 was not the appropriate logic of knowledge (in particular, they argued that the negative introspection axiom—if you know something you know that you don't know it—was inappropriate), S5 seemed to me the most natural epistemic logic for distributed computing applications. Game theorists independently adopted S5 for their applications, with much the same motivation. While I was willing to back off from S5 for computational reasons (see below), using it gave a great deal of insight, so I was more than happy to adopt it.

The philosophy literature also focused on the single-agent case—when debating the "right, true properties" of knowledge, there's no need to consider multiple agents. (Although, to be fair, Lewis's seminal work on convention [Lewis 1969] introduced the notion of common knowledge and pointed out its importance to characterizing conventions.) Distributed systems are composed of many agents, so I was interested in the describing the knowledge of not just of one agent, but of groups of agents, and how that could change as a result of communication. Particularly important was what one agents knew about other agents' knowledge.

I see my biggest contribution as the work I did with Yoram Moses [Halpern and Moses 1990] showing how knowledge could be used to help understand and analyze distributed systems, and in defining a

[1] While it's true that my primary interests were and continue to be somewhat more pragmatic, I must confess that I actually have a recent paper that considers definitions of knowledge in terms of (true) belief [Halpern, Samet, and Segev 2008].

concrete model of knowledge in distributed systems (the so-called *runs-and-systems model*, which is developed further in [Halpern and Fagin 1989] and [Fagin, Halpern, Moses, and Vardi 1995]). "Possible worlds" and "accessibility relations" had a concrete meaning in this framework, one that computer scientists interested in designing systems could understand, even if they were not so concerned with the philosophical issues. This work introduced the vocabulary of epistemic logic to distributed computing, introduced the vocabulary of epistemic logic to distributed computing, and emphasized the connection between common knowledge and important notions like agreement and coordination. This initial work was followed by a great deal of work that expanded on these themes, joint with Ron Fagin, Yoram Moses, Mark Tuttle, Moshe Vardi, Lenore Zuck, and others (largely summarized in [Fagin, Halpern, Moses, and Vardi 1995]).

Distributed computing concerns also motivated a second line of work that is enjoying newfound life in game theory. Hintikka [1962] had already observed that the standard possible-worlds semantics of epistemic logic suffers from the *logical omniscience problem*: agents knew all tautologies and that agents knew the logical consequences of their knowledge. This clearly is not a particularly accurate description of people's knowledge. Ron Fagin and I [1988] suggested that one way of dealing with the problem was in terms of *awareness*. The idea is that there is a distinction between an agent's *implicit knowledge* and her *explicit knowledge*. Implicit knowledge is defined in the usual way, as truth in all worlds; to have explicit knowledge of a fact ϕ, you must implicitly know ϕ and be aware of it. What does it mean to be aware of ϕ? That depends. It could mean that there are basic concepts in ϕ that the agent is not aware of (for example, an agent who has never heard of astrology might not be aware that the moon is in the seventh house); this is the interpretation that seems to be the one that game theorists are now focusing on. However, it might also mean that the agent cannot compute the truth of ϕ. In any case, there has been a great deal of work on awareness recently, largely published in economics journals (see, for example, [Dekel, Lipman, and Rustichini 1998; Feinberg 2004; Halpern 2001a; Halpern and Rêgo 2006a; Halpern and Rêgo 2006b; Heifetz, Meier, and Schipper 2006; Heifetz, Meier, and Schipper 2007; Modica and Rustichini 1994; Modica and Rustichini 1999], but this is just the tip of the iceberg).

3. *What do you think is the proper role of epistemology in relation to other areas of philosophy and other academic disciplines?*

   Since I'm not a philosopher by training or background, I can't comment on what the proper role of epistemology is in relation to other areas of philosophy. I feel somewhat more confident in commenting on the role of knowledge in computer science and economics. We make decisions on the basis of our knowledge and belief. Understanding the connection between knowledge, belief, and action plays a key role in both computer science and economics. This connection may involve counterfactual beliefs as well (see, for example, [Halpern 2001b; Halpern and Moses 2004]). I believe that epistemology can help elucidate these issues.

4. *What do you consider to be the most neglected topics and/or contributions in contemporary epistemology?*

   I'm not sure what the most neglected topics or contributions in epistemology are (I've probably neglected them along with everyone else!), so let me change the question slightly to one that I think is more interesting: What are the most pressing issues in epistemology today? To me, perhaps the most pressing issue is that of getting a good model of knowledge that takes resource-bounded reasoners, and particularly the computation involved in computing what you know, into account. This is hardly a new topic, but I think the problem is far from solved. A "good" model is one what is easy to work with, mathematically well founded, and leads to new insights. In particular, I would hope that such a model would give insights into cryptographic notions like *zero knowledge proofs* [Goldwasser, Micali, and Rackoff 1989]), and be used to define solutions concepts that take computation into account in a serious way. I believe that the notion of awareness should help here, as will perhaps the notion of *algorithmic knowledge* [Fagin, Halpern, Moses, and Vardi 1995, Chapter 10]. However, much more remains to be done.

   In addition, I would like to see more work using epistemic logic on analyzing and synthesizing programs from knowledge-based specifications. The *synthesis* problem in computer science is that of deriving a program, given a specification that the program should satisfy. In many cases, what the system is supposed to do is best expressed in terms of knowledge. This is particularly true for specifications involving security; for example, we may not want an adversary to *know* certain secret information. (See [Halpern and O'Neill 2002; Halpern and O'Neill] for some discussion of how security specifications can be expressed in terms of knowledge.) There has been some work on automatically synthesizing programs from the specifications that they must satisfy (see [Bickford, Constable, Halpern, and Petride 2005; Engelhardt, Meyden, and Moses 1998; Engelhardt, Meyden, and Moses 2001]), but I think that much more can and should be done.

The first topic I mentioned, finding a good computational notion of knowledge, should have some important philosophical implications. Among other things, it would give a notion of knowledge that does not suffer from the logical omniscience problem. To the extent that it can be used to help clarify important issues in distributed computing and game theory, it will also capture important features of human epistemic reasoning. The second topic is perhaps of more pragmatic rather than philosophical interest, but it would show the applicability of epistemic reasoning to an important class of problems.

5. *What do you think the future of epistemology will (or should) hold?*

   I believe that the connections between epistemology, computer science, and game theory will continue to broaden and deepen. Issues of knowledge, belief, awareness (or lack of it), computation, and counterfactuals will play an increasingly important role in the future. Philosophers could have a significant impact here and, indeed, some already have. For example, Bob Stalnaker's recent work on game theory has been quite influential (see, for example, [Stalnaker 1996]); his early work on counterfactuals [Stalnaker 1968] has turned out to be quite relevant to game theory as well. However, having an influence will require understanding, not just the philosophical issues, but the concerns of computer scientists and economists.

References

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics 4*(6), 1236–1239.

Barwise, J. (1981). Scenes and other situations. *Journal of Philosophy 78*(7), 369–397.

Bickford, M., R. L. Constable, J. Y. Halpern, and S. Petride (2005). Knowledge-based synthesis of distributed systems using event structures. In *Proc. 11th Int. Conf. on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 2004)*, Lecture Notes in Computer Science, vol. 3452, pp. 449–465. Springer-Verlag.

Dekel, E., B. Lipman, and A. Rustichini (1998). Standard state-space models preclude unawareness. *Econometrica 66*, 159–173.

Engelhardt, K., R. van der Meyden, and Y. Moses (1998). A program refinement framework supporting reasoning about knowledge and time. In J. Tiuryn (Ed.), *Proc. Foundations of Software Science and Computation Structures (FOSSACS 2000)*, pp. 114–129. Berlin/New York: Springer-Verlag.

Engelhardt, K., R. van der Meyden, and Y. Moses (2001). A refinement theory that supports reasoning about knowledge and time for synchronous agents. In *Proc. International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, pp. 125–141. Berlin/New York: Springer-Verlag.

Fagin, R. and J. Y. Halpern (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence 34*, 39–76.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995). *Reasoning About Knowledge*. Cambridge, Mass.: MIT Press. A slightly revised paperback version was published in 2003.

Feinberg, Y. (2004). Subjective reasoning—games with unawareness. Technical Report Resarch Paper Series #1875, Stanford Graduate School of Business.

Gamow, G. and M. Stern (1958). *Puzzle Math*. New York: Viking Press.

Gardner, M. (1984). *Puzzles From Other Worlds*. New York: Viking Press.

Goldwasser, S., S. Micali, and C. Rackoff (1989). The knowledge complexity of interactive proof systems. *SIAM Journal on Computing 18*(1), 186–208.

Gray, J. (1978). Notes on database operating systems. In R. Bayer, R. M. Graham, and G. Seegmuller (Eds.), *Operating Systems: An Advanced*

Course, Lecture Notes in Computer Science, Volume 66. Berlin/New York: Springer-Verlag. Also appears as IBM Research Report RJ 2188, 1978.

Halpern, J. Y. (2001a). Alternative semantics for unawareness. *Games and Economic Behavior 37*, 321–339.

Halpern, J. Y. (2001b). Substantive rationality and backward induction. *Games and Economic Behavior 37*, 425–435.

Halpern, J. Y. and R. Fagin (1989). Modelling knowledge and action in distributed systems. *Distributed Computing 3*(4), 159–179. A preliminary version appeared in *Proc. 4th ACM Symposium on Principles of Distributed Computing*, 1985, with the title "A formal model of knowledge, action, and communication in distributed systems: preliminary report".

Halpern, J. Y. and Y. Moses (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM 37*(3), 549–587. A preliminary version appeared in *Proc. 3rd ACM Symposium on Principles of Distributed Computing*, 1984.

Halpern, J. Y. and Y. Moses (2004). Using counterfactuals in knowledge-based programming. *Distributed Computing 17*(2), 91–106.

Halpern, J. Y. and K. O'Neill. Anonymity and information hiding in multiagent systems. *Journal and Computer Security 13*(3), 483–514.

Halpern, J. Y. and K. O'Neill (2002). Secrecy in multiagent systems. In *Proc. 15th IEEE Computer Security Foundations Workshop*, pp. 32–46. To appear, *ACM Transactions on Information and System Security.*

Halpern, J. Y. and L. C. Rêgo (2006a). Extensive games with possibly unaware players. In *Proc. Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 744–751. Full version available at arxiv.org/abs/0704.2014.

Halpern, J. Y. and L. C. Rêgo (2006b). Reasoning about knowledge of unawareness. In *Principles of Knowledge Representation and Reasoning: Proc. Tenth International Conference (KR '06)*, pp. 6–13. Full version available at arxiv.org/cs.LO/0603020.

Halpern, J. Y., D. Samet, and E. Segev (2008). On definability in multimodal logic i. defining knowledge in terms of belief. Unpublished manuscript, available at www.cs.cornell.edu/home/halpern/papers.

Heifetz, A., M. Meier, and B. Schipper (2006). Interactive unawareness. *Journal of Economic Theory 130*, 78–94.

Heifetz, A., M. Meier, and B. Schipper (2007). Unawareness, beliefs and games. In *Theoretical Aspects of Rationality and Knowledge: Proc. Eleventh Conference (TARK 2007)*, pp. 183–192.

Hintikka, J. (1962). *Knowledge and Belief.* Ithaca, N.Y.: Cornell University Press.

Lewis, D. (1969). *Convention, A Philosophical Study.* Cambridge, Mass.: Harvard University Press.

Littlewood, J. E. (1953). *A Mathematician's Miscellany.* London: Methuen and Co.

Modica, S. and A. Rustichini (1994). Awareness and partitional information structures. *Theory and Decision 37*, 107–124.

Modica, S. and A. Rustichini (1999). Unawareness and partitional information structures. *Games and Economic Behavior 27*(2), 265–298.

Stalnaker, R. C. (1968). A semantic analysis of conditional logic. In N. Rescher (Ed.), *Studies in Logical Theory*, pp. 98–112. Oxford University Press.

Stalnaker, R. C. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy 12*, 133–163.