PHIL 222
Philosophical Foundations of Computer Science
Week 11, Tuesday

Nov. 5, 2024

Epistemology (2)
Multi-Agent Systems and AI:
The "Problem of Logical Omniscience"
(cont'd)

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

By the “why can’t β?” reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ldots, \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-(}B\text{-or-}C\text{)-then-}D \implies D$$

is provable

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ldots, \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-(}B\text{-or-}C\text{)-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1,\ \ldots,\ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B,\ \text{if-}(B\text{-or-}C)\text{-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ldots, \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-(}B\text{-or-}C\text{)-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-(}B\text{-or-}C\text{)-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

By the “why can't β?” reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ldots, \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-}(B\text{-or-}C)\text{-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Assm: β-knows-(A-and-B)
- Assm: β-knows-(if-(B-or-C)-then-D)

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-(}B\text{-or-}C\text{)-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Assm: β-knows-(A-and-B)
- Assm: β-knows-(if-(B-or-C)-then-D)
- β-knows-B

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-}(B\text{-or-}C)\text{-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Assm: β-knows-(A-and-B)
- Assm: β-knows-(if-(B-or-C)-then-D)
- β-knows-B
- β-knows-(B-or-C)

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1,\ \ldots,\ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B,\ \text{if-(}B\text{-or-}C\text{)-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Assm: β-knows-(A-and-B)
- Assm: β-knows-(if-(B-or-C)-then-D)
- β-knows-B
- β-knows-(B-or-C)

$\therefore$ β-knows-D

By the "why can't β?" reasoning, a more general rule can be justified:

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

Example of how this principle may be applied / justified:

$$A\text{-and-}B, \text{ if-}(B\text{-or-}C)\text{-then-}D \implies D$$

is provable because:

- Assm: A-and-B
- Assm: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Assm: β-knows-(A-and-B)
- Assm: β-knows-(if-(B-or-C)-then-D)
- β-knows-B
- β-knows-(B-or-C)

$\therefore$ β-knows-D

Thus we have proven this!

$$\beta\text{-knows-}(A\text{-and-}B), \ \beta\text{-knows-}(\text{if-}(B\text{-or-}C)\text{-then-}D) \implies \beta\text{-knows-}D.$$

The partition model,

ⓘ "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

The partition model,

i "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable,

The partition model,

(i) "α knows that φ" is true in a state w

$\Longleftrightarrow$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, which means that

(ii) "ψ" is true in every possibility where "φ_1", ..., "φ_n" are all true.

The partition model,

(i) "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, which means that

(ii) "ψ" is true in every possibility where "φ_1", ..., "φ_n" are all true.

Then, for any state w,

"β knows that φ_1", ..., "β knows that φ_n" are all true in w

$\Longrightarrow$

The partition model,

(i) "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, which means that

(ii) "ψ" is true in every possibility where "φ_1", ..., "φ_n" are all true.

Then, for any state w,

"β knows that φ_1", ..., "β knows that φ_n" are all true in w

$\Longrightarrow$ "φ_1", ..., "φ_n" are all true in all the states in the same β-cell as w

(by (i))

The partition model,

(i) "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, which means that

(ii) "ψ" is true in every possibility where "φ_1", ..., "φ_n" are all true.

Then, for any state w,

"β knows that φ_1", ..., "β knows that φ_n" are all true in w

$\Longrightarrow$ "φ_1", ..., "φ_n" are all true in all the states in the same β-cell as w (by (i))

$\Longrightarrow$ "ψ" is true in all the states in the same β-cell as w (by (ii))

The partition model,

ⓘ "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, which means that

ⓘⓘ "ψ" is true in every possibility where "φ_1", ..., "φ_n" are all true.

Then, for any state w,

"β knows that φ_1", ..., "β knows that φ_n" are all true in w

$\Longrightarrow$ "φ_1", ..., "φ_n" are all true in all the states in the same β-cell as w (by ⓘ)

$\Longrightarrow$ "ψ" is true in all the states in the same β-cell as w (by ⓘⓘ)

$\Longrightarrow$ "β knows that ψ" is true in w. (by ⓘ)

The partition model,

(i) "α knows that φ" is true in a state w

$\iff$ "φ" is true in all the states in the same α-cell as w,

also confirms (or "validates") the inference rule.

Proof. Suppose $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, which means that

(ii) "ψ" is true in every possibility where "φ_1", ..., "φ_n" are all true.

Then, for any state w,

"β knows that φ_1", ..., "β knows that φ_n" are all true in w

$\Longrightarrow$ "φ_1", ..., "φ_n" are all true in all the states in the same β-cell as w (by (i))

$\Longrightarrow$ "ψ" is true in all the states in the same β-cell as w (by (ii))

$\Longrightarrow$ "β knows that ψ" is true in w. (by (i))

Thus, "β knows that ψ" is true in every possibility where "β knows that φ_1", ..., "β knows that φ_n" are all true. $\square$

But the inference rule

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

means **logical omniscience**:

- β knows everything logically provable.

But the inference rule

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1,\ \ldots,\ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

means **logical omniscience**:

- β knows everything logically provable.
- All the theorems of arithmetic are (by definition) provable from axioms of arithmetic. Therefore β knows all the theorems, as long as she knows all the axioms.

But the inference rule

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1,\ \ldots,\ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

means **logical omniscience**:

- β knows everything logically provable.
- All the theorems of arithmetic are (by definition) provable from axioms of arithmetic. Therefore β knows all the theorems, as long as she knows all the axioms.
- So, as long as β knows the axioms of arithmetic, she knows that $f(\bar{n}) = \overline{m}$ for all computable functions f and all inputs n.

But the inference rule

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

means **logical omniscience**:

- β knows everything logically provable.
- All the theorems of arithmetic are (by definition) provable from axioms of arithmetic. Therefore β knows all the theorems, as long as she knows all the axioms.
- So, as long as β knows the axioms of arithmetic, she knows that $f(\bar{n}) = \overline{m}$ for all computable functions f and all inputs n.
- E.g., β knows the optimal solution in the travelling salesman problem (which is at least as hard as SAT!).

But the inference rule

- If $\varphi_1, \ldots, \varphi_n \Rightarrow \psi$ is provable, then so is

$$\beta \text{ knows that } \varphi_1, \ \ldots, \ \beta \text{ knows that } \varphi_n \implies \beta \text{ knows that } \psi.$$

means **logical omniscience**:

- β knows everything logically provable.
- All the theorems of arithmetic are (by definition) provable from axioms of arithmetic. Therefore β knows all the theorems, as long as she knows all the axioms.
- So, as long as β knows the axioms of arithmetic, she knows that $f(\bar{n}) = \overline{m}$ for all computable functions f and all inputs n.
- E.g., β knows the optimal solution in the travelling salesman problem (which is at least as hard as SAT!).
- Something even worse may hold in the partition model: β knows all the mathematical truths (e.g. which Turing machine halts on which input), since they are true in all the states.

We usually do not think human agents have such inference power.

The first of famous examples may be the case of geometry we saw in Plato's *Meno* (Socrates in black, Meno in red, a boy in green):

> "Tell me now, boy, you know that a square figure is like this?" "I do." "A square then is a figure in which all these four sides are equal?" "Yes indeed." [. . .] "How many feet is twice two feet? Work it out and tell me." "Four, Socrates." "Now we could have another figure twice the size of this one, with the four sides equal like this one." "Yes." "How many feet will that be?" "Eight." "Come now, try to tell me how long each side of this will be. The side of this is two feet. What about each side of the one which is its double?" "Obviously, Socrates, it will be twice the length." "You see, Meno, that I am not teaching the boy anything, but all I do is question him. And now he thinks he knows the length of the line on which an eight-foot figure is based. Do you agree?" "I do." "And does he know?" "Certainly not." [82b–e]

The logic of “knows that” we reviewed admits logical omniscience, and therefore does not seem to be a realistic model of either

- human knowledge,
- what an AI agent knows, or
- what a computer can compute.

The logic of “knows that” we reviewed admits logical omniscience, and therefore does not seem to be a realistic model of either

- human knowledge,
- what an AI agent knows, or
- what a computer can compute.

This is reflected in the following remark by Herlihy et al.:

> Here, too, there are many problems that are not computable, but these computability failures reflect the difficulty of making decisions in the face of ambiguity and have little to do with the inherent computational power of individual participants. [p. 11]

The logic of “knows that” we reviewed admits logical omniscience, and therefore does not seem to be a realistic model of either

- human knowledge,
- what an AI agent knows, or
- what a computer can compute.

This is reflected in the following remark by Herlihy et al.:

> Here, too, there are many problems that are not computable, but these computability failures reflect the difficulty of making decisions in the face of ambiguity and have little to do with the inherent computational power of individual participants. [p. 11]

Thus, logical omniscience is out of the scope of the theory of distributed computing. — But does it mean that it is not problematic in distributed computing?

Stalnaker (1999), “The Problem of Logical Omniscience, II”:

> As distributed systems theorists have emphasized, their conception of knowledge is an **externalist** one in the sense that the content of a knowledge claim is characterized from the point of view of the theorist, and not of the knower. The language of the epistemic logic [logic of “knows that”] talks about what processors know, but it is not intended to model the knower’s way of expressing or representing what it knows. The content clause in a knowledge attribution in this language is the attributor’s way of expressing the information about the system that, according to the attribution, is reflected in the local state of the knower. [p. 258, emphasis KK]

One might think that the uncompromising externalism that I have been illustrating is one of the features that distinguishes the simplified, perhaps metaphorical, conception of knowledge used in distributed systems theory from a realistic conception [. . .] and that perhaps this feature explains why, in the distributed systems sense of "know," even simple processors with no computational capacities at all know all the logical consequences of their knowledge, while in the sense of "know" we ordinarily use, even the most brilliant logician does not.

One might think that the uncompromising externalism that I have been illustrating is one of the features that distinguishes the simplified, perhaps metaphorical, conception of knowledge used in distributed systems theory from a realistic conception [. . .] and that perhaps this feature explains why, in the distributed systems sense of "know," even simple processors with no computational capacities at all know all the logical consequences of their knowledge, while in the sense of "know" we ordinarily use, even the most brilliant logician does not. But I don't think this is the root of the problem, since I think our ordinary concept of knowledge is an externalist one in the same sense. Real knowledge must be available in some sense, but the knower need not be able to say what he knows; much of what we know is manifested in action, but not in speech, and so it cannot be required by our concept of knowledge that we describe knowledge the way the knower would express it. Nor is it [. . .] [pp. 259f., emphasis KK]

Nor is it reasonable to take a knowledge attribution to make a claim about the form in which the knowledge is stored. The form in which information is stored — the structure of the local states of knowers in virtue of which they are correctly described as knowing things — will have an effect on the availability of the knowledge, but it is the notion of availability itself, and not what may influence it, that we need to get clear about to understand the problem of logical omniscience.

Nor is it reasonable to take a knowledge attribution to make a claim about the form in which the knowledge is stored. The form in which information is stored — the structure of the local states of knowers in virtue of which they are correctly described as knowing things — will have an effect on the availability of the knowledge, but it is the notion of availability itself, and not what may influence it, that we need to get clear about to understand the problem of logical omniscience. There is a problem of logical omniscience for the concept of knowledge defined in distributed systems theory, in fact a problem with some urgency. I think it is essentially the same problem as the one that infects our ordinary concepts of knowledge and belief, and that getting clear about how the problem arises in this simpler setting will help us understand how it arises for us.

[p. 260, emphasis KK]

To sum up Stalnaker's point:

- Computer scientists may distinguish "externalist" knowledge from ordinary knowledge, and say that agents can be, or must be, logically omniscient from the external perspective.

To sum up Stalnaker's point:

- Computer scientists may distinguish "externalist" knowledge from ordinary knowledge, and say that agents can be, or must be, logically omniscient from the external perspective.
- But our ordinary criterion of knowledge often concerns whether the agent has it **available / accessible** from the external perspective.

To sum up Stalnaker's point:

- Computer scientists may distinguish "externalist" knowledge from ordinary knowledge, and say that agents can be, or must be, logically omniscient from the external perspective.
- But our ordinary criterion of knowledge often concerns whether the agent has it **available / accessible** from the external perspective.

A smart agent probably can act on the instruction ⓐ but not ⓑ:

ⓐ If you know your hat color, announce it.

ⓑ If you know the optimal route for the travelling salesman, take the route.

To sum up Stalnaker's point:

- Computer scientists may distinguish "externalist" knowledge from ordinary knowledge, and say that agents can be, or must be, logically omniscient from the external perspective.
- But our ordinary criterion of knowledge often concerns whether the agent has it **available / accessible** from the external perspective.

A smart agent probably can act on the instruction (a) but not (b):

(a) If you know your hat color, announce it.

(b) If you know the optimal route for the travelling salesman, take the route.

The agent may "know" the optimal route in the logically omniscient sense, but not in the availability sense — and it is the latter sense of "know" that matters to the instruction.

To sum up Stalnaker's point:

- Computer scientists may distinguish "externalist" knowledge from ordinary knowledge, and say that agents can be, or must be, logically omniscient from the external perspective.
- But our ordinary criterion of knowledge often concerns whether the agent has it **available / accessible** from the external perspective.

A smart agent probably can act on the instruction ⓐ but not ⓑ:

ⓐ If you know your hat color, announce it.

ⓑ If you know the optimal route for the travelling salesman, take the route.

The agent may "know" the optimal route in the logically omniscient sense, but not in the availability sense — and it is the latter sense of "know" that matters to the instruction.

Thus, the question of what knowledge is available can be essential when, e.g., designing protocols for distributed computing.

The same has been observed in expected utility theory — a theory that is also essential to distributed computing (cf. Vlassis, Ch. 2).

Savage 1967:

> The analysis should be careful not to prove too much; for some departures from theory are inevitable, and some even laudable. For example, a person required to risk money on a remote digit of π would, in order to comply fully with the theory, have to compute that digit, though this would really be wasteful if the cost of computation were more than the prize involved. For the postulates of the theory imply that you should behave in accordance with the logical implication of all that you know. Is it possible to improve the theory in this respect, making allowances within it for the cost of thinking, or would that entail paradox, as I am inclined to believe but unable to demonstrate?

(Quoted from
`https://plato.stanford.edu/entries/bounded-rationality/`)

We saw an example of a logically not omniscient agent in Plato's *Meno* (Socrates in black, Meno in red, a boy in green):

> "Come now, try to tell me how long each side of this will be. The side of this is two feet. What about each side of the one which is its double?" "Obviously, Socrates, it will be twice the length." "You see, Meno, that I am not teaching the boy anything, but all I do is question him. And now he thinks he knows the length of the line on which an eight-foot figure is based. Do you agree?" "I do." "And does he know?" "Certainly not." [82d–e]

We saw an example of a logically not omniscient agent in Plato's *Meno* (Socrates in black, Meno in red, a boy in green):

> "Come now, try to tell me how long each side of this will be. The side of this is two feet. What about each side of the one which is its double?" "Obviously, Socrates, it will be twice the length." "You see, Meno, that I am not teaching the boy anything, but all I do is question him. And now he thinks he knows the length of the line on which an eight-foot figure is based. Do you agree?" "I do." "And does he know?" "Certainly not." [82d–e]

Indeed, Plato gives two themes that computer scientists incorporate in their approaches to logical omniscience based on their distinction between **explicit** and **implicit knowledge**. Let's observe these themes, the computer scientists' approaches — and Stalnaker's criticisms.

“Watch him now recollecting things in order, as one must recollect. Tell me, boy, [. . .] Well, let us draw from it four equal lines, and surely that is what you say is the eight-foot square?” “Certainly.” [. . .] [82e–83b]

“Does not this line from one corner to the other cut each of these figures in two? “Yes.” “So these are four equal lines which enclose this figure? “They are.” “Consider now: how large is the figure? “I do not understand.” “Within these four figures, each line cuts off half of each, does it not? “Yes.” “How many of this size are there in this figure? “Four.” “How many in this? “Two.” “What is the relation of four to two? “Double.” “How many feet in this? “Eight.” “Based on what line? “This one.” “That is, on the line that stretches from corner to corner of the four-foot figure? “Yes.” “Clever men call this the diagonal, so that if diagonal is its name, you say that the double figure would be that based on the diagonal? “Most certainly, Socrates.” [85a–85b]

"What do you think, Meno? Has he, in his answers, expressed any opinion that was not his own?" "No, they were all his own." "And yet, as we said a short time ago, he did not know?" "That is true." "So these opinions were in him, were they not?" "Yes." "So the man who does not know has within himself true opinions about the things that he does not know?" "So it appears." "These opinions have now just been stirred up like a dream, but if he were repeatedly asked these same questions in various ways, you know that in the end his knowledge about these things would be as accurate as anyone's." "It is likely." "And he will know it without having been taught but only questioned, and find the knowledge within himself?" "Yes." "And is not finding knowledge within oneself recollection?" "Certainly." [85b–d]

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: *A*-and-*B*
- Hyp: if-(*B*-or-*C*)-then-*D*

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Hyp: β-knows-(A-and-B)
- Hyp: β-knows-(if-(B-or-C)-then-D)

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Hyp: β-knows-(A-and-B)
- Hyp: β-knows-(if-(B-or-C)-then-D)
- β-knows-B

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Hyp: β-knows-(A-and-B)
- Hyp: β-knows-(if-(B-or-C)-then-D)
- β-knows-B
- β-knows-(B-or-C)

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Hyp: β-knows-(A-and-B)
- Hyp: β-knows-(if-(B-or-C)-then-D)
- β-knows-B
- β-knows-(B-or-C)

$\therefore$ β-knows-D

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B • B-or-C

$\therefore$ D

- Hyp: β-knows-(A-and-B)
- Hyp: β-knows-(if-(B-or-C)-then-D)
- β-knows-B • β-knows-(B-or-C)

$\therefore$ β-knows-D

"So the man who does not know has within himself true opinions about the things that he does not know?"

Plato claims that every person is logically omniscient in the sense of "having within himself", while ordinary knowing is recollecting — computer scientists may see them as implicit & explicit knowledge.

Two themes: ❶ Socrates's argument that the boy has had the geometric knowledge all along is similar to this proof:

- Hyp: A-and-B
- Hyp: if-(B-or-C)-then-D
- B
- B-or-C

$\therefore$ D

- Hyp: β-knows-(A-and-B)
- Hyp: β-knows-(if-(B-or-C)-then-D)
- β-knows-B
- β-knows-(B-or-C)

$\therefore$ β-knows-D

"So the man who does not know has within himself true opinions about the things that he does not know?"

Plato claims that every person is logically omniscient in the sense of "having within himself", while ordinary knowing is recollecting — computer scientists may see them as implicit & explicit knowledge.

❷ "These opinions have now just been stirred up like a dream, but if he were repeatedly asked these same questions in various ways, you know that in the end his knowledge about these things would be as accurate as anyone's."

Plato argues that questions help turn implicit knowledge explicit.

Computer scientists formalize, e.g.

1. Awareness (on top of the logic of “knows that”)

Computer scientists formalize, e.g.

1. Awareness (on top of the logic of "knows that")

ARTIFICIAL INTELLIGENCE 39

Belief, Awareness, and Limited Reasoning*

Ronald Fagin and Joseph Y. Halpern
IBM Almaden Research Center, San Jose, CA 95120, U.S.A.

Recommended by Daniel G. Bobrow

ABSTRACT

Several new logics for belief and knowledge are introduced and studied, all of which have the property that agents are not logically omniscient. In particular, in these logics, the set of beliefs of an agent does not necessarily contain all valid formulas. Thus, these logics are more suitable than traditional logics for modelling beliefs of humans (or machines) with limited reasoning capabilities. Our first logic is essentially an extension of Levesque's logic of implicit and explicit belief, where we extend to allow multiple agents and higher-level belief (i.e., beliefs about beliefs). Our second logic deals explicitly with "awareness," where, roughly speaking, it is necessary to be aware of a concept before one can have beliefs about it. Our third logic gives a model of "local reasoning," where an agent is viewed as a "society of minds," each with its own cluster of beliefs, which may contradict each other.

The animal knows, of course. But it certainly does not know that it knows.

Teilhard de Chardin

1. Introduction

There has long been interest in both philosophy and AI in finding natural semantics for logics of knowledge and belief. The standard approach has been the so-called *possible-worlds* model. The intuitive idea, which goes back to

Fagin–Halpern 1988:

"α-is-aware-of-φ",

Fagin–Halpern 1988:

"α-is-aware-of-φ", which can be interpreted variously:

- "α is aware of φ",
- "α is able to figure out the truth of φ",
- "α is able to compute the truth of φ within time T".

Fagin–Halpern 1988:

"α-is-aware-of-φ", which can be interpreted variously:

- "α is aware of φ",
- "α is able to figure out the truth of φ",
- "α is able to compute the truth of φ within time T".

We can then entertain various inference rules: e.g.,

- "α-is-aware-of-φ" $\iff$ "α-is-aware-of-(not-φ)",
- "α-is-aware-of-(φ-and-ψ)" $\implies$ "α-is-aware-of-φ",
- "α-is-aware-of-(φ-and-ψ)" $\iff$ "α-is-aware-of-(ψ-and-φ)".

Fagin–Halpern 1988:

"α-is-aware-of-φ", which can be interpreted variously:

- "α is aware of φ",
- "α is able to figure out the truth of φ",
- "α is able to compute the truth of φ within time T".

We can then entertain various inference rules: e.g.,

- "α-is-aware-of-φ" $\iff$ "α-is-aware-of-(not-φ)",
- "α-is-aware-of-(φ-and-ψ)" $\implies$ "α-is-aware-of-φ",
- "α-is-aware-of-(φ-and-ψ)" $\iff$ "α-is-aware-of-(ψ-and-φ)".

Then define "α explicitly knows that φ"

$$\iff (\alpha\text{-knows-that-}\varphi)\text{-and-}(\alpha\text{-is-aware-of-}\varphi).$$

Fagin–Halpern 1988:

"α-is-aware-of-φ", which can be interpreted variously:

- "α is aware of φ",
- "α is able to figure out the truth of φ",
- "α is able to compute the truth of φ within time T".

We can then entertain various inference rules: e.g.,

- "α-is-aware-of-φ" $\iff$ "α-is-aware-of-(not-φ)",
- "α-is-aware-of-(φ-and-ψ)" $\implies$ "α-is-aware-of-φ",
- "α-is-aware-of-(φ-and-ψ)" $\iff$ "α-is-aware-of-(ψ-and-φ)".

Then define "α explicitly knows that φ"

$$\iff (\alpha\text{-knows-that-}\varphi)\text{-and-}(\alpha\text{-is-aware-of-}\varphi).$$

So, given any logical truth φ, α knows that φ, but maybe only implicitly and not explicitly, since α may not be aware of φ.