I know you know, but do you know I know you know? And related questions...

Michael Lieberman
Kalamazoo College

Math/CS/Physics Colloquium

February 5th, 2014

It's an old (incredibly old) cliche that logic is the study of the structure of valid human reasoning, and that mathematical logic is meant to capture the thought processes of an ideal mathematician, cutting away all the bells and whistles.

In either view, logic is directed toward purely internal, introspective activity.

Today we discuss one of many *logics of rational agency*, logics that look outward into the world, and try to provide a framework in which to understand actual human interaction...

A subdivision of philosophy concerned with knowledge, principally:

- What knowledge is: what does it mean to know a given statement p? Is knowledge of X possible, and how does one answer the skeptic?
- How knowledge is acquired: how is new information integrated into/reconciled with one's existing knowledge and beliefs? In a system of interacting agents, how does knowledge propogate?

The former, internal and static, has been a favored area of study for ages. The latter is the domain of epistemic logic...

"As we know,
There are known knowns.
There are things we know we know.
We also know
There are known unknowns.
That is to say
We know there are some things
We do not know.
But there are also unknown unknowns,
The ones we don't know
We don't know. "

—Donald Rumsfeld, February 2003
Pieces of Intelligence: The Existential Poetry of Donald Rumsfeld

Of course, we all have our off days. Concerning the WMD:

"We know where they are. They're in the area around Tikrit and Baghdad and east, west, south and north somewhat."

—Donald Rumsfeld, March 2003

What does it mean to say that "A knows p?"

Old answer: "A has a justified true belief that p."

That is to say:

- A believes p,
- has justification for this belief,
- and p is true.

There's a cottage industry of counterexamples (*Gettier problems*), refinements of this answer, and further counterexamples. All very silly: we take the old answer as a reasonable approximation.

For many reasons, we try to capture this and other properties of individual human knowledge through a *logical calculus*. First, a set of propositions, denoted by lower case letters: p, q, r, ...

Also, a few basic logical symbols:

Symbol	*Name*	*Usage*	*Meaning*
$\neg$	Negation	$\neg p$	not p
$\wedge$	Conjunction	$p \wedge q$	p and q
$\vee$	Disjunction	$p \vee q$	p or q
$\rightarrow$	Conditional	$p \rightarrow q$	p implies q
$\leftrightarrow$	Biconditional	$p \leftrightarrow q$	p iff q

What's missing? Knowledge.

We introduce knowledge into the mix with a new operator, K. This operator can be applied to any proposition p, giving $\mathrm{K}p$, whose intended interpretation is "the agent knows p."

We introduce new axioms to make sure the logic supports this interpretation:

Truth	$\mathrm{K}p \to p$
Closure	$\mathrm{K}(p \to q) \wedge \mathrm{K}p \to \mathrm{K}q$
Positive Introspection	$\mathrm{K}p \to \mathrm{K}\mathrm{K}p$
Negative Introspection	$\neg\mathrm{K}p \to \mathrm{K}\neg\mathrm{K}p$

This is idealized, obviously—part of the motivation is that this gives a famous form of *modal logic*.

Having modeled the situation in this way, we can use what we know of modal logic—proof theory, semantics, etc.—to analyze questions arising in philosophy, linguistics, and cognitive psychology, among other places.

We also have a concise way of expressing certain fine distinctions:

- p is a known known: $\mathrm{K}\mathrm{K}p$
- p is a known unknown: $\mathrm{K}\neg\mathrm{K}p$
- p is an unknown unknown: $\neg\mathrm{K}\neg\mathrm{K}p$

Things get more interesting—and difficult—with more agents.

Say there are finitely many agents, A_1, A_2,. . . , A_n. We introduce an operator K_i for each agent, and require that each K_i satisfy the axioms for individual knowledge.

More interesting: agents interact, and know things about what others know—our language must expand to include, e.g.

- A_1 knows a secret p, and knows A_2 is unaware of her knowledge:

$$\mathsf{K}_1(p \wedge \neg \mathsf{K}_2 \mathsf{K}_1 p)$$

.

- A_1 and A_2 know p, and are sure A_3 is out of the loop:

$$\mathsf{K}_1 p \wedge \mathsf{K}_2 p \wedge \mathsf{K}_1 \neg \mathsf{K}_3 p \wedge \mathsf{K}_2 \neg \mathsf{K}_3 p$$

.

Suppose I've just entered France through the Channel Tunnel, and am driving a car with British plates down a narrow road. A car approaches, also with British plates. If I keep right, per French traffic law, can I be certain no collision will result?

Suppose I've just entered France through the Channel Tunnel, and am driving a car with British plates down a narrow road. A car approaches, also with British plates. If I keep right, per French traffic law, can I be certain no collision will result?

Let r be "Here people drive on the right." Let K_1 represent my knowledge, and K_2 the knowledge of the other driver.

Suppose I've just entered France through the Channel Tunnel, and am driving a car with British plates down a narrow road. A car approaches, also with British plates. If I keep right, per French traffic law, can I be certain no collision will result?

Let r be "Here people drive on the right." Let K_1 represent my knowledge, and K_2 the knowledge of the other driver.

At first glance, what we might call *shared knowledge* seems like it might provide enough certainty:

$$Sr := K_1 r \wedge K_2 r$$

Suppose I've just entered France through the Channel Tunnel, and am driving a car with British plates down a narrow road. A car approaches, also with British plates. If I keep right, per French traffic law, can I be certain no collision will result?

Let r be "Here people drive on the right." Let K_1 represent my knowledge, and K_2 the knowledge of the other driver.

At first glance, what we might call *shared knowledge* seems like it might provide enough certainty:

$$\mathsf{S}r := \mathsf{K}_1 r \wedge \mathsf{K}_2 r$$

But this is woefully inadequate!

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$...

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$... but maybe $\neg K_1 K_2 r$.

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$... but maybe $\neg K_1 K_2 r$.
- To the contrary, say $K_1 K_2 r$...

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$... but maybe $\neg K_1 K_2 r$.
- To the contrary, say $K_1 K_2 r$... but maybe $\neg K_2 K_1 K_2 r$.

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$... but maybe $\neg K_1 K_2 r$.
- To the contrary, say $K_1 K_2 r$... but maybe $\neg K_2 K_1 K_2 r$.
- To the contrary, say $K_2 K_1 K_2 r$...

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$... but maybe $\neg K_1 K_2 r$.
- To the contrary, say $K_1 K_2 r$... but maybe $\neg K_2 K_1 K_2 r$.
- To the contrary, say $K_2 K_1 K_2 r$... but maybe $\neg K_1 K_2 K_1 K_2 r$.

Say we start with Sr, so we each know we're meant to keep right.

- So $K_1 r$ and $K_2 r$... but maybe $\neg K_1 K_2 r$.
- To the contrary, say $K_1 K_2 r$... but maybe $\neg K_2 K_1 K_2 r$.
- To the contrary, say $K_2 K_1 K_2 r$... but maybe $\neg K_1 K_2 K_1 K_2 r$.
- And so on...

Each of these is clearly inadequate: whatever formula like this we write, the situation it describes is still one in which I will be uncertain whether to keep right or left. What's needed is an infinite alternation:

$$\ldots \mathrm{K}_2\mathrm{K}_1\mathrm{K}_2\mathrm{K}_1\mathrm{K}_2\mathrm{K}_1\mathrm{K}_2 r$$

or, for the other driver,

$$\ldots \mathrm{K}_1\mathrm{K}_2\mathrm{K}_1\mathrm{K}_2\mathrm{K}_1\mathrm{K}_2\mathrm{K}_1 r$$

This is an approximation of the idea of *common knowledge*—in the two-agent situation, it expresses that everybody knows, everybody knows everybody knows, and so on.

No one likes infinite formulas, though. Does this repeated alternation terminate in something sensible? A first, and very subtle, application of serious logic.

A fixed point argument guarantees that we can define a new operator C that is equivalent to the conjunction ("and") of both putatively infinite formulas.

Note

In the many-agent situation, with knowers represented by K_1, $K_2, \ldots K_n$, *a similar argument gives a common knowledge operator,* C.

This is a very powerful idea, with applications in linguistics and the philosophy of language, but also in game theory and elsewhere:

- Robert Aumann: economics and game theory.
- David Lewis: philosophical analysis of social and linguistic convention.
- Herbert Clark: conventionalist account of language.
- Steven Pinker: the purpose of innuendo.

Our knowledge changes over time as we are presented with new information, so we must allow for this possibility: dynamic epistemic logic.

Formally, this introduces a family of *update operators* $[p]$ ("after public announcement p...") in the language, and axiomatizes the way these updates influence the agents. Example:

- $[p]Cp$

This is important, even in the informal realm.

Three children are playing outside after a heavy rain, and just as they enter the house their father says to them "at least one of you has mud on your forehead." Each child can see the others, but not him or herself. Their father then asks them repeatedly "Do you know whether you have mud on your forehead?" What happens?

Three children are playing outside after a heavy rain, and just as they enter the house their father says to them "at least one of you has mud on your forehead." Each child can see the others, but not him or herself. Their father then asks them repeatedly "Do you know whether you have mud on your forehead?" What happens?

What if there are n children?

Three children are playing outside after a heavy rain, and just as they enter the house their father says to them "at least one of you has mud on your forehead." Each child can see the others, but not him or herself. Their father then asks them repeatedly "Do you know whether you have mud on your forehead?" What happens?

What if there are n children?

This is an example of the power of direct language, public announcements, and common knowledge. It is also an example of a paradoxical update: "I don't know" as a knowledge-producing statement.

On Friday, a teacher informs his students that there will be a surprise exam the following week—so surprising that even the night before, they won't know it's coming.

A cruel and unusual, but not paradoxical, state of affairs. But if we really commit to the update, a paradox does arise...

This is known as the Suprise Exam Paradox, or, sometimes, as the Surprise Execution Paradox. One of many possible solutions comes out of Baltag's work on belief revision in dynamic epistemic logic.

The basic idea: all updates are not created equal. Depending on the source and context, we may not allow new information to cut away possible worlds entirely, but rather to modify their plausibility. Shades of gray:

- Update: Infallable, wholly trusted source. If announces p, we rule out all $\neg p$ worlds.
- Radical Upgrade: Source generally trustworthy. If announces p, all p worlds become more plausible than $\neg p$ worlds.
- Conservative Upgrade: Source barely trusted. If announces p, only the most plausible p world is promoted above the $\neg p$ worlds.

This gives a way out of the surprise exam paradox (and a useful life lesson): don't treat the teacher as an infallible source. If the remark has the effect of one of the softer upgrades, there's no problem...

This actually points us in the direction of doxastic logic—the logic of belief—and exciting applications in the study of belief revision and learning theory.

But that's a topic for another day.