Formal Methods in Epistemology: Epistemic Logic

KEVIN DORST
kmdorst@mit.edu

MIT IAP
January 24, 2017

Last time: probability as a tool for reasoning about uncertainty.

This time: **epistemic logic** as a tool for reasoning about knowledge.

Others' knowledge, and our own.

Knowledge of others...

1. Muddy children

Three children: Abby, Bianca, and Cora. Father: "You can play outside, but don't get dirty! If you do, you won't get dessert tonight."

They play; in fact all three get dirt on their foreheads. Each can see whether the *others* are dirty, but not whether they themselves are. (They won't tell each other—they want each others' cake!) Thus each knows that at least 2 of the children (the other two) are dirty.

Mother takes sympathy: if they reason well enough, they should get off the hook. First, she tells them: (1) "At least one of you is dirty." Then she repeatedly asks them: (2) "Do any of you know whether you are dirty?"

Result: if it's common ground that they will each reason perfectly, then the first two times Mother asks (2) they will each say "No," and the third time they will say, "Yes—I know I'm dirty."

More generally: if there are n children and k get dirty, then $k-1$ times they will all say 'No,' and on the kth questioning the dirty children will say 'Yes.'

Why? If 1 child is dirty, then on the first questioning he'll say 'Yes.' Suppose 2 are dirty. On the first questioning they say 'No'—since for all they know it's just the *other* that's dirty. But then they each realize that if they *weren't* muddy then the *other* child would've said 'Yes'; so on the second questioning they say 'Yes.' Suppose 3 are dirty. They say 'No' the first two times. But they each know that—by the $k = 2$ case—if they themselves weren't dirty, the other children would've said 'Yes'; so on the third questioning they say 'Yes.'

And so on...

But there's an easier way!

There are 8 possibilities for who's dirty. Each can tell whether the other's are dirty, but not themselves. Represent their knowledge by representing what they *don't* know—what possibilities they *leave open*:

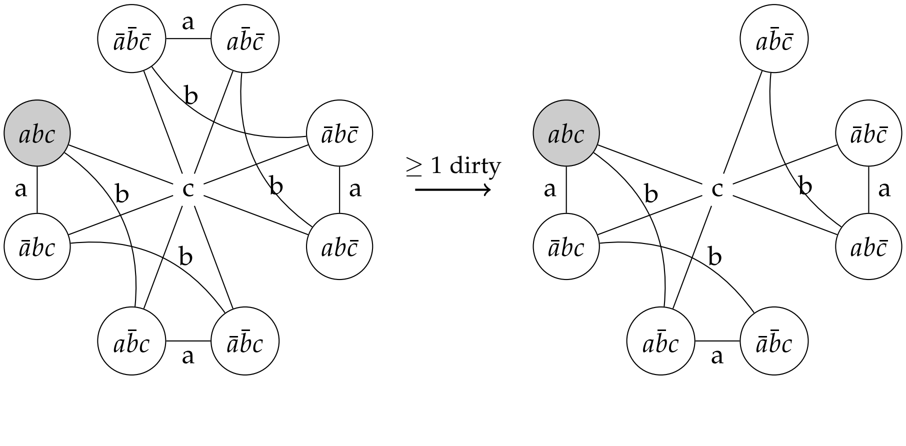
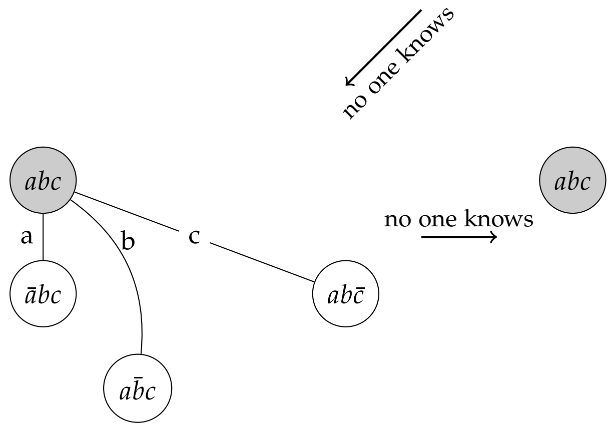


Diagram from Moss 2015



2. Epistemic frames

An *epistemic frame* is a tuple $\langle W, R_1, R_2, ..., R_n \rangle$ where W is a set of worlds and each R_i is a binary relation between worlds.

W a set of possibilities, like with probability spaces.

Hintikka's (1962) insight: what people know is *itself* a fact determined by the world. Uncertainty about what you or others know is uncertainty about what world you're in.

R_i captures this dependence for agent i. wR_ix iff x is consistent with i's knowledge at w (iff for all i knows in w, she might be in x).

Assume (for now): R_i is reflexive (wRw), symmetric ($wRx \Rightarrow xRw$), and transitive (wRx & $xRy \Rightarrow wRy$).

Logic (via set theory). If $p, q \subseteq W$, then:

p true at w iff $w \in p$.
$\neg p$ true at w iff $w \notin p$. Negation
$p \cap q$ true at w iff $w \in p$ and $w \in q$. Conjunction
$p \vee q$ true at w iff $w \in p$ or $w \in q$ (or both). Disjunction
$p \to q$ true at w iff either $w \notin p$ or $w \in q$. Material conditional; "p only if q"
$K_i p$ true at w iff for all w': if wRw' then $w' \in p$. i knows that p at w iff all worlds compatible with her knowledge at w are p-worlds.

Let A (B, C) be the set of worlds where Abby (Bianca, Cora) is dirty. Then ≥ 1 *dirty* = $A \vee B \vee C$. And *no one knows whether they're dirty* is:

$$(\neg K_a A \wedge \neg K_a \neg A) \wedge (\neg K_b B \wedge \neg K_b \neg B) \wedge (\neg K_c C \wedge \neg K_c \neg C) \qquad (*)$$

As can be checked, after it is announced that at least one child is dirty, repeatedly announcing (*) (deleting worlds where it's false) leads the model to "unravel" as above.

3. Common knowledge

In Muddy Children, the publicity of the announcements are crucial. Say p is *mutual knowledge* if everyone knows p:

Mp true at w iff $K_1 p \wedge K_2 p \wedge ... \wedge K_n p$.

At world *abc* where everyone's dirty, they *already* have mutual knowledge that *someone dirty*. If Mother told them each in private, nothing would change. In

All worlds w consistent with anyone's knowledge in *abc* are such that $w \in$*someone dirty*.

fact, they already have mutual knowledge *that* they have mutual knowledge that *someone dirty*. If Mother told them in private *that* she had told the other two, nothing would change.

All worlds w consistent with anyone's knowledge in *abc* are such that all worlds x consistent with anyone's knowledge in w are such that $x \in$*someone dirty*.

But at *abc* they don't have mutual knowledge that they have mutual knowledge that they have mutual knowledge that *someone dirty*; i.e. Mp and MMp but $\neg MMMp$.

To get the Muddy Children reasoning going in full generality (for any number n of children), they all must have *common* knowledge of the announcement: everyone knows, and everyone knows that everyone knows, and everyone knows that everyone knows that everyone knows, and so on.

Cp true at w iff Mp and MMp and $MMMp$ and...

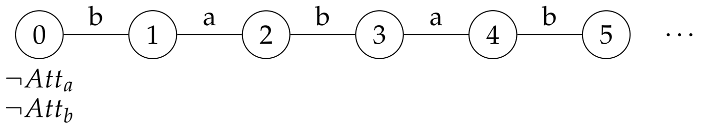
4. Common knowledge and coordination

Abby and Bianca generals encamped on either side of the Enemy. If one of them attacks alone, it will be catastrophic—so it's obvious that neither will attack if they don't know the other will. Thus:

(**) $Att_i \to K_i(Att_j)$

It's night; they can only communicate by sending messengers through the Enemy camp. Each messenger has some chance of getting caught. Abby will decide whether to send the first messenger, and each general will send a confirmation messenger after they receive a message.

Result: no matter how many messengers they send, they will never attack. Why? Because they can never obtain common knowledge that they will attack, yet they have common knowledge of (**). Let worlds be number of messengers sent:



Lessons:

Interpersonal reasoning is subtle!

More knowledge/reasoning power can make it *harder* to coordinate.

If we sometimes have common knowledge of obvious background conditions, need publicity for coordination.

Maybe we *never* have common knowledge?

Lederman 2016a, 2016b

Knowledge of self...

5. Introspection

Consider a single-agent epistemic frame $\langle W, R\rangle$. Above we assumed that R was an equivalence relation:

Reflexive: wRw
Transitive: wRx & $xRy \Rightarrow wRy$
Symmetric: $wRx \Rightarrow xRw$

It turns out that such assumptions correspond nicely to formal properties of the knowledge operator K.

Reflexivity yields factivity: $Kp \to p$.
Transitivity yields **positive introspection**: $Kp \to KKp$.
Given these two, symmetry yields **negative introspection**: $\neg Kp \to K\neg Kp$.

Knowing requires truth.

Knowing requires knowing that you know.

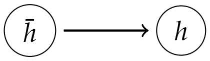
If you fail you know, you know that you do.

We can then use cases to assess the plausibility of these axioms.

6. Skeptical scenarios

Billy is a regular guy, with regular limbs and regular perception. He knows he has hands.

Brain-Billy has every reason to think he's a regular guy, with regular limbs and regular perception. But in fact he's a brain in a vat. He has no hands, so he can't know that he *does* have hands. And yet, he's in no position to know this. He has every reason to think he's like Billy—every reason to think that he knows he has hands. His epistemic predicament:



Negative introspection fails at $\bar{h}$. $\neg Kh$ but $\neg K\neg Kh$. Why? Symmetry fails. $\bar{h}Rh$, but $h\not R\bar{h}$.

Upshot: if the case is right, negative introspection fails and our R_i needn't be symmetric.

7. Margins for error

Williamson 2000: Ch. 5

Mr. Magoo stairs out the window at a towering oak, trying to estimate its height. In fact it's 600 inches tall, but he doesn't know that—his eyesight's not nearly that good, and he knows as much. He knows that he can't reliably tell the difference between a tree that's n inches tall and one that's $n+1$ inches tall. So he knows that if the tree *isn't* (at least) $n+1$ inches tall, then he can't know that it's n inches tall. (If he believed it, he'd only be guessing; that belief could easily have been false.) Equivalently, he knows that if he knows the tree is n inches tall, then it's $n+1$ inches tall:

Margin: $K(Kp_n \to p_{n+1})$

Suppose, for reductio, positive introspection holds. Clearly Magoo knows that tree is at least 1 inch tall: Kp_1. By positive introspection, he knows that he knows this: KKp_1. By Margin, he knows that he can know this only if the tree is in fact 2 inches tall: $K(Kp_1 \rightarrow p_2$. Hence he can infer that it's 2 inches tall: Kp_2. Again, by positive introspection we iterate up: KKp_2. By Margin we slide over: Kp_3. With a few hundred repetitions, we infer the conclusion that Mr. Magoo knows the tree is 601 inches tall: Kp_{601}. But that conclusion is false! The tree *isn't* 601 inches tall, so Mr. Magoo can't know that it is.

Upshot: Either Margin or positive introspection fails.