Higher-Order Probability: Word-Completion Tasks

Kevin Dorst
kevindorst@pitt.edu

March 2021

1 What you know, and what you should know

Hold off on probabilities for a moment. Start with two modalities, what an agent *knows* and what she *should know*, given her evidence and abilities. Write Kp for the claim that she knows p, and Sp for the claim that she *should* know p.

Model these as standard modal operators, using a bimodal Kripke frame $\langle W, \mathcal{K}, \mathcal{S}\rangle$ (Hintikka 1962; Kripke 1963). W is a (let's suppose finite) set of states or possible worlds. $\mathcal{K}$ and $\mathcal{S}$ are each binary relations on W. $\mathcal{K}_w := \{x \in W | w\mathcal{K}x\}$ is the set of worlds accessible from w under $\mathcal{K}$, and likewise $\mathcal{S}_w := \{x \in W : w\mathcal{S}x\}$. We'll use $\mathcal{K}$ to define the $K(\cdot)$ operator, and $\mathcal{S}$ to define the $S(\cdot)$ operator.

In particular, propositions (or events) are modeled as subsets of W; $p \subseteq W$ is true at w iff $w \in p$, $\neg p = W \setminus p$, $p \wedge q = p \cap w$, etc. For any proposition $p \subseteq W$, $Kp := \{w \in W : \mathcal{K}_w \subseteq p\}$ is the proposition that p is known; it's true at a world w iff all worlds that are $\mathcal{K}$-accessible from w are p-worlds, false otherwise. Likewise, $Sp := \{w \in W : \mathcal{S}_w \subseteq p\}$ is the proposition that p *should* be known; it's true at a world w iff all worlds that are $\mathcal{S}$-accessible from w are p-worlds.

Our agent will be presented with a word-completion task: a string of letters and some blanks. She will be given 5 seconds to look at it, and she will have to say how confident she is that the string is completable. For example she may see a string like this:

_EAR_T

In which case the answer is 'yes', because 'learnt' is a word. Or she may see a string like this:

P_G_ER

And the answer is 'no', because no English word completes that string.

Suppose she is presented with _EAR_T. Then, right after being presented with the string but before having any time to think about it, there are three relevant epistemic possibilities (worlds) for her. Maybe there's a word and she'll find it after 5 seconds (a). Maybe there's a world and she won't find it after 5 seconds (b). Or Maybe there's no word (and she won't find one after 5 seconds) (c). (We (now) know that c is not actual—there is a word, since 'learnt' is one. But she doesn't know that before having any time to think; so c is an epistemic possibility for her at this point; it's a possibility wherein fact 'learnt' isn't word, and therefore there's no completion to _EAR_T.)

Now consider her epistemic state after 5 seconds of looking at the string, right when it disappears. What *will* she know? That depends on which possibility she's in. After all, she'll know whether she found a word or not. Let's suppose she knows nothing more than this. Then her knowledge-state after looking at the string for 5 seconds can be represented with the following partial frame $\langle W, \mathcal{K} \rangle$, where black arrows represent $\mathcal{K}$-relations and therefore what's consistent with her knowledge in the various possibilities:

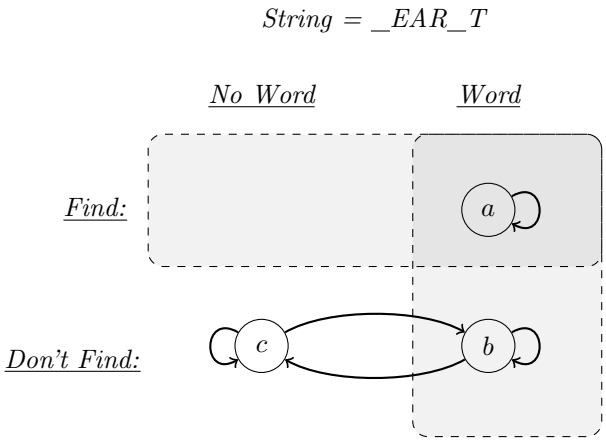

Figure 1: What she knows in the word-completion task.

Thus at a she knows that she's at a, while at c and b, all she knows is that she didn't find a word. Thus $K(\mathit{Word})$ is true at a and false at the other two possibilities. $K(\mathit{Find})$ is true at a, and $K(\neg \mathit{Find})$ is true at the other two.

What about what she *should* know after 5 seconds of looking at the string? Again, that depends on what possibility she's in. But, we may suppose, it depends even more. After all, it's possible that what she *should* know outstrips what she *does* know. Sometimes she makes mistakes; sometimes she fails to use all her relevant evidence. In particular, her lexicon encodes the information that 'LEARNT' is a word, and therefore she is in a position to know—she *should* know—that the word is completable, if in fact it is. If 'learnt' is a word but she doesn't find it, then she knows less than the should—she

should know that although she didn't in fact find one, there is one. (If she properly used her lexical evidence, that's what she *would* conclude: "Ah, there is one, but I didn't find it in 5 seconds.") On the other hand, if 'learnt' is *not* a word (c is actual), then she's not in a position to know it's not. After all, her powers of word-identification are limited. The task is, intuitively, *semi-decidable* for her within 5 seconds. If there is a word, she should know that there is. But if there's not, she shouldn't (can't) know that there's not. Even if she properly uses all her lexical information, if c is actual, then she won't find one—but she can't know whether her failure to find one was because there isn't one (c is actual), or because there is one that she in fact missed (b is actual).

Thus we can represent what she *should* know with the following partial frame $\langle W, \mathcal{S} \rangle$, where red arrows indicate $\mathcal{S}$-relations and, therefore, what she should know in the relevant possibilities:

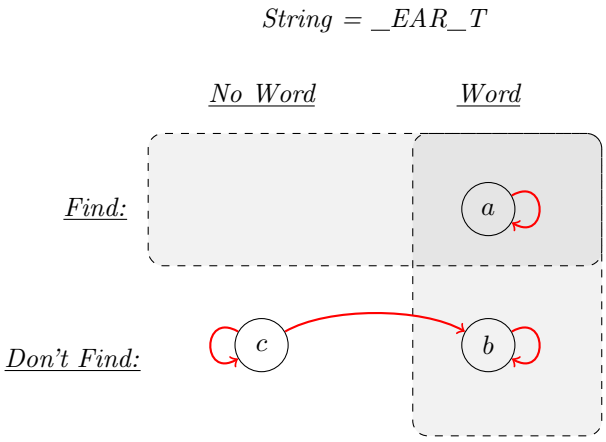

Figure 2: What she should know in the word-completion task.

Thus at a she should know she's at a: $S(\mathit{Find})$ and $S(\mathit{Word})$ are both true. At b, she should know that she didn't find one in the relevant time ($S(\neg \mathit{Find})$ is true) but that in fact there is one ($S(\mathit{Word})$ is true). At c, she should know she didn't find one ($S(\neg \mathit{Find})$ is true), but she shouldn't (can't) know whether there is one: $\neg S(\neg \mathit{Word})$ is true. Since every possibility is one where *if* there's a word, she should know it ($\mathit{Word} \subseteq S(\mathit{Word})$), it follows that at c there is a failure of **negative introspection** on the modality S. (As seen from the fact that $\mathcal{S}$ is not euclidean: $c\mathcal{S}b$, but $b \not\mathcal{S} c$.) Thus at world c, $\neg S(\mathit{Word})$ is true, but so is $\neg S \neg S(\mathit{Word})$: she should not know there's word, but she shouldn't know that she shouldn't know there's a word (for all she can know, perhaps she *should* know there is). She know she didn't find one—but she also knows that if there is one, she should find it (her failure to find one was an error); if not, she shouldn't.

We can then combine the two models to model both what she does know and what she should know after 5 seconds (see page 4).

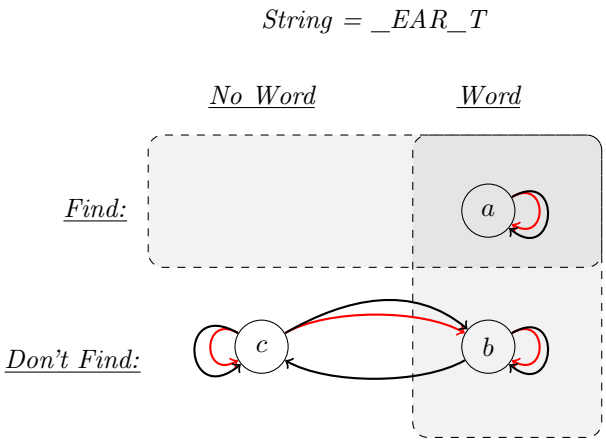

Figure 3: What she should know in the word-completion task.

Thus we can say that, e.g. at c, she knows that she doesn't know there's a word—$K(\neg K(\mathit{Word}))$—but for all she knows, she *should* know there's a world—$\neg K(\neg S(\mathit{Word}))$. Thus she doesn't know whether what she knows lines up with what she should know: $\neg K(K(\mathit{Word}) \leftrightarrow S(\mathit{Word}))$.

Thus we can represent her (higher-order) epistemic states easy enough: what she knows about what she knows, and what she should know about what she should know (and, indeed, what she knows about what she should know, and what she should know about what she knows) can all be represented with a bimodal Kripke frame.

2 Adding probabilities

If you're okay with what I've said so far, adding probabilities (and, therefore, higher-order probabilities) is simple. Let '$C(p)$' be her actual credence in p, whatever it is; let '$P(p)$' be the credence in p she *should* have, whatever it is. These are definite descriptions for numbers (i.e. they are random variables), so the values they give vary across worlds. At some worlds, $C(\mathit{Word}) = 1$; at others, perhaps $C(\mathit{Word}) = \frac{1}{3}$.

Model these as probabilistic modal operators, using a bimodal *probabilistic* Kripke frame $\langle W, \mathcal{C}, \mathcal{P}\rangle$ (compare Gaifman 1988; Samet 2000). W is again a (let's suppose finite) set of worlds or states. $\mathcal{C}$ and $\mathcal{P}$ can now be thought of as *graded* accessibility relations, which assign a non-negative weight to each pair of worlds $\langle x, y\rangle$, $\mathcal{C}_x(y)$ and $\mathcal{P}_x(y)$. The weights from a given world must sum to 1: for any $x \in W$: $\sum_{y\in W} \mathcal{C}_x(y) = 1$ and $\sum_{y\in W} \mathcal{P}_x(y) = 1$. Equivalently, $\mathcal{C}$ and $\mathcal{P}$ can be thought of as functions from worlds x to probability functions $\mathcal{C}_x$ and $\mathcal{P}_x$ defined over the subsets of W: for any $x \in W$ and $q \subseteq W$: $\mathcal{C}_x(q) = \sum_{y\in q} \mathcal{C}_x(y)$, and likewise $\mathcal{P}_x(q) = \sum_{y\in q} \mathcal{P}_x(y)$. Whereas '$C(q)$' above was a definite description for a number (i.e. was a random variable), '$\mathcal{C}_x(q)$' is a a rigid

designator for a number. (That is, $C(q)$ varies across words; $\mathcal{C}_x(q)$ does not.) We can use $\mathcal{C}$ and $\mathcal{P}$ to define the $C(\cdot) = t$ and $P(\cdot) = t$ propositional operators.

In particular, for any world x, think of $\mathcal{C}_x$ as our agent's actual credence function at world x. Likewise, think of $\mathcal{P}_x$ as the credence function our agent *should* have at x—the one they would have, were they to properly use their evidence. Since our agent has different opinions and evidence at difference worlds, we may well have $\mathcal{C}_x \neq \mathcal{C}_y$ for $x \neq y$, and likewise $\mathcal{P}_x \neq \mathcal{P}_y$.

This then allows us to define propositions about the agent's actual credences, as well as the credences they should have, as propositions (sets of worlds) in the frame. In particular, for any proposition $q \subseteq W$ and $t \in [0, 1]$, $[C(q) = t] := \{w \in W : \mathcal{C}_w(q) = t\}$ is the proposition (set of worlds) at which $\mathcal{C}$ assigns credence t to q, i.e. the agent's actual degree of confidence in q equals t. Likewise, $[P(q) = t] := \{w \in W : \mathcal{P}_w(q) = t\}$ is the proposition (set of worlds) at which $\mathcal{P}$ assigns t to q, i.e. the agent *should* assign degree of confidence t to q.

Now let's apply this to our word-completion model. A natural model of the agent's credences in this situation says this: her prior credences in the various possibilities (the ones she had right after she saw '_EAR_T' but before she had time to think) assigned some numbers to the various possibilities—say $\frac{1}{2}$ probability for the No-Word-No-Find possibility (c), $\frac{1}{4}$ for the Word-No-Find possibility (b), and $\frac{1}{4}$ for the Word-Find possibility (a).[1] Then her *actual* credences at w, $\mathcal{C}_w$ are recovered by conditioning this prior on what she knows according to the model above, i.e. on whether or not she found a word. (Formally, for all $w \in W : \mathcal{C}_w = \mathcal{P}^0_w(\cdot|\mathcal{K}_w)$, where $\mathcal{P}^0_w$ is her prior and $\mathcal{K}_w = \{x \in W | w\mathcal{K}x\}$, as above, is the strongest proposition she knows at w.) And the credences she *should* have at w, $\mathcal{P}_w$, are recovered by conditioning this prior on what she *should* know according to the model above. (Formally, for all $w \in W : \mathcal{P}_w = \mathcal{P}^0_w(\cdot|\mathcal{S}_w)$, where $\mathcal{P}^0_w$ is her prior and $\mathcal{S}_w = \{x \in W : w\mathcal{S}x\}$, as above, is the strongest proposition she *should* know at w.)

If so, her posterior actual credences in the various possibilities are represented by the frame in Figure 4.

Thus $[C(\textit{Word}) = 1] = \{a\}$, i.e. a is the only world at which she's certain there's a word. Meanwhile $[C(\textit{Word}) = \frac{1}{3}] = \{b, c\}$, since at both b and c she assigns $\frac{1}{3}$ credence to possibilities in which there's a word (i.e. to b). Now, since we've identified $[C(\textit{Word}) = \frac{1}{3}]$ as a set of worlds in the frame, and her credences are always distributed over such worlds, it follows that she assigns credences to facts about what her credences are. In particular at world c she is certain that she is $\frac{1}{3}$ confident that there's a word: $\mathcal{C}_c([C(\textit{Word}) = \frac{1}{3}]) = \mathcal{C}_c(\{b, c\}) = 1$. Likewise at b. Meanwhile, at a she is certain that she assigns credence 1 to there being a world: $\mathcal{C}_a([C(\textit{Word}) = 1]) = \mathcal{C}_a(\{a\}) = 1$. Thus

[1] For simplicity, I won't represent these priors formally in the model; but we of course could by adding a constant function from worlds to probability functions $\mathcal{P}^0$ such that for all $w \in W$: $\mathcal{P}^0_w(a) = \mathcal{P}^0_w(b) = \frac{1}{4}$ and $\mathcal{P}^0_w(c) = \frac{1}{2}$.

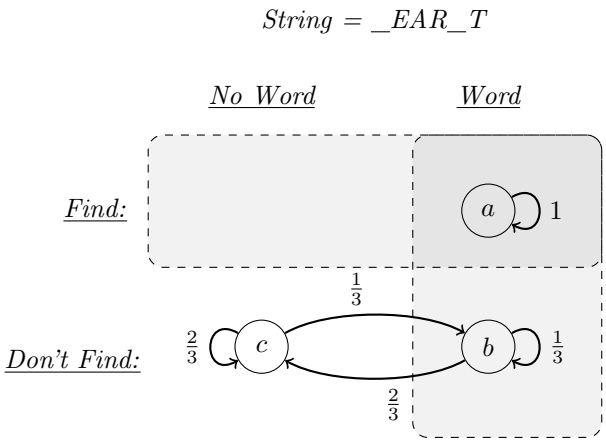

Figure 4: What her actual credences are in the word-completion task.

the proposition that she's certain that she assigns credence $\frac{1}{3}$ to there being a world is true at worlds b and c:

$$\begin{aligned}
[C([C(\textit{Word}) = \tfrac{1}{3}]) = 1] &= \\
\{w \in W : \mathcal{C}_w([C(\textit{Word}) = \tfrac{1}{3}]) = 1\} &= \\
\{w \in W : \mathcal{C}_w(\{b, c\}) = 1\} &= \\
\{b, c\}.
\end{aligned}$$

Likewise, for even higher-order claims as well: the proposition that she assigns credence 1 to the proposition that she assigns credence 0 to the proposition that she assigns credence 1 to there being a word is true at b and c in this frame:

$$\begin{aligned}
[C([C([C(\textit{Word}) = 1]) = 0]) = 1] &= \\
[C([C([C(\{a, b\}) = 1]) = 0]) = 1] &= \\
[C([C(\{a\}) = 0]) = 1] &= \\
[C(\{b, c\}) = 1] &= \\
\{b, c\}
\end{aligned}$$

So identifying claims about credences with sets of worlds allows us to "unravel" any iterated higher-order probability claim to just be a claim about credences assigned to a set of worlds. This is as it should be. When you have credences, you distribute them over possibilities. And in those possibilities, there are facts about what credences you have. Therefore, when you have credences about what credences you have, those are simply credences about which possibilities you're in.

But so far, such higher-order credences are, in a sense, trivial: since in the above

frame actual credences are fully introspective, we have that for any q, t: $[C(q) = t] \leftrightarrow [C([C(q) = t]) = 1]$ is true at every world.

Now turn from what our agent's credences *are*, C, to what they *should* be, P. Here we will get failures of introspection and therefore nontrivial higher-order probabilities. As said above, the credences she *should* have are obtained by conditioning her prior distribution on what she *should* know. As seen in Figure 2, at a she should know she's at a; at b she should know she's at b, and at c she should know that she's either at b or c. Thus Figure 5 is what her credences should be (given what she should know) in the various possibilities:

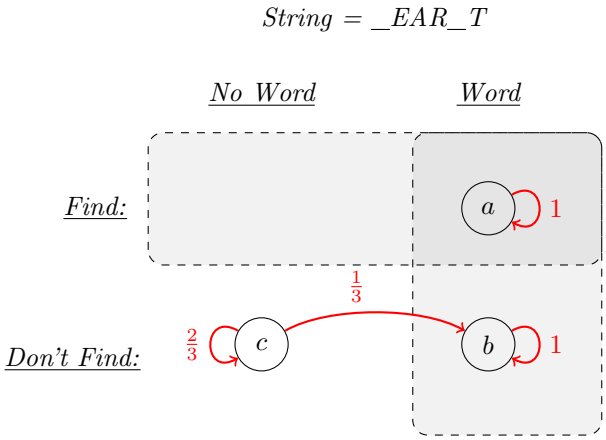

Figure 5: What her credences *should* be in the word-completion task.

Thus $[P(\textit{Word}) = 1] = \{a, b\}$, i.e. at both worlds a and b she should be certain (should know) that there's a word. (If she were to properly use her lexical evidence, she would become certain there's a word.) Similarly, at both a and b she should be certain that she should be certain of this: $[P([P(\textit{Word}) = 1]) = 1] = [P(\{a, b\}) = 1] = \{a, b\}$. Meanwhile, at c she should be $\frac{1}{3}$ confident that there's a word: $[P(\textit{Word}) = \frac{1}{3}] = \{c\}$. Crucially, at c the credences she should have, P, are not introspective. (Recall that Negative Introspection failed for what she should know, S, in the above model of this scenario. Likewise, now, for introspection of how confident she should be.) In particular, $[P([P(\textit{Word}) = \frac{1}{3}]) = \frac{2}{3}] = [P(\{c\}) = \frac{2}{3}] = \{c\}$, so at c she should assign $\frac{2}{3}$ credence to the claim that she should assign $\frac{1}{3}$ credence to *Word*. Meanwhile, $[P([P(\textit{Word}) = 1]) = \frac{1}{3}] = [P(\{a, b\}) = \frac{1}{3}] = \{c\}$, so at c she should also assign $\frac{1}{3}$ credence to the claim that she should assign credence 1 to *Word*.

This is as it should be: (she knows that) if there is a word, she should know (and thus be sure) there is, whereas if there's no word, she can't know (and thus should be unsure) whether there is. Thus at possibility c, she should be unsure what credence she should have. Combining this frame with the above model of her *actual* credences,

since she does know what her actual credences are, it follows that she should be unsure whether her actual credences equal the credences she should have.

2.1 Why Reflection/Martingale principles fail

If you're unsure of the value of a random variable, you can take it's expectation. Since at c our agent should be unsure of the value of $P(\mathit{Word})$, i.e. how confident she should be that there's a word, she can take it's expectation. For any random variable X ($X : W \to \mathbb{R}$), let $\mathbb{E}_w(X) := \sum_{t\in\mathbb{R}} \mathcal{P}_w(X = t) \cdot t$ be $\mathcal{P}_w$'s expectation of X, i.e. the expectation for X that our agent *should* have at world w.[2]

What should our agent's expectation of $P(\mathit{Word})$ be at c? It's easy to calculate that

$$\begin{aligned}\mathbb{E}_c(P(\mathit{Word})) &= \mathcal{P}_c(P(\mathit{Word}) = \tfrac{1}{3}) \cdot \frac{1}{3} + \mathcal{P}_c(P(\mathit{Word}) = 1) \cdot 1 \\ &= \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot 1 \\ &= \frac{5}{9}\end{aligned}$$

Since $\mathcal{P}_c(\mathit{Word}) = \frac{1}{3}$, it follows that $\mathcal{P}_c(\mathit{Word}) = \frac{!}{3} \neq \frac{5}{9} = \mathbb{E}_c(P(\mathit{Word}))$. That is, the credence our agent should have in *Word* does not equal the expectation she should have for the credence she should have in *Word*.[3]

This is because once we have higher-order uncertainty, "Reflection" or Martingale-principles fail. Such principles say things like: $\mathcal{P}_c(q|P(q) = t) = t$; or, using '$\pi$' as a rigid designator for credence functions, $\mathcal{P}_c(\cdot|P = \pi) = \pi$. Such principles always fail when there is higher-order uncertainty (Samet 2000; Dorst 2019). This is easy to see conceptually. Suppose π has higher-order uncertainty, and so in particular $\pi(P = \pi) < 1$. Then conditional on P being π, you should of course be certain that P is π: $\mathcal{P}_c(P = \pi|P = \pi) = 1$. But since by hypothesis π is not certain of that, $\pi(P = \pi) < 1$. Thus $\mathcal{P}_c(\cdot|P = \pi) \neq \pi$ (Elga 2013).

Here's a concrete example. At world c, our agent should leave open that the credence she should have that there's a word is either $\frac{1}{3}$ (if there's no word), or 1 (if there is a word). How confident should she be that there's a word, conditional on the the claim that she should be $\frac{1}{3}$ confident that there's a word? Reflection/martingale principles would say $\frac{1}{3}$. But that's wrong. If there is a word, she should have credence 1 there is. If there's not, she should have credence $\frac{1}{3}$ there is. So conditional on the claim that she should be $\frac{1}{3}$ confident there's a word, she should be certain there's *not* a word. Formally, $[P(\mathit{Word}) = \frac{1}{3}] = \{c\} \subseteq \neg\mathit{Word}$, in our frame, therefore

[2] Remember W is finite, so the summation is kosher.

[3] Note that, of course $\mathbb{E}_b(P(\mathit{Word})) = 1$. Thus since $\mathcal{P}_c$ is unsure whether c or b is actual, it is likewise unsure whether the rational expectation of $P(\mathit{Word})$ is $\frac{5}{9}$ or 1. This is as it should be: whenever probabilities aren't introspective, so too expectations won't be introspective.

$\mathcal{P}_c(\mathit{Word}|P(\mathit{Word}) = \frac{1}{3}) = \mathcal{P}_c(\mathit{Word}\ |\{c\}) = \mathcal{P}_c(\{a, b\}\ |\{c\}) = 0$. The reason is this: if she should be $\frac{1}{3}$ confident that there's a word, that's only because she *can't be sure* she should be $\frac{1}{3}$ confident there's a word (she has higher-order uncertainty). Thus *learning* that she should be $\frac{1}{3}$ confident that there's a word gives her new information—information she didn't have before—and therefore can change how confident she should be.

References

Dorst, Kevin, 2019. 'Higher-Order Uncertainty'. In Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 35–61. Oxford University Press.

Elga, Adam, 2013. 'The puzzle of the unmarked clock and the new rational reflection principle'. *Philosophical Studies*, 164(1):127–139.

Gaifman, Haim, 1988. 'A Theory of Higher Order Probabilities'. In Brian Skyrms and William L Harper, eds., *Causation, Chance, and Credence*, volume 1, 191–219. Kluwer.

Hintikka, Jaako, 1962. *Knowledge and Belief*. Cornell University Press.

Kripke, Saul A, 1963. 'Semantical analysis of modal logic i normal modal propositional calculi'. *Mathematical Logic Quarterly*, 9(56):67–96.

Samet, Dov, 2000. 'Quantified Beliefs and Believed Quantities'. *Journal of Economic Theory*, 95(2):169–185.