Caleb

Hi Alexandru and Sonja,

I would like your help with a problem I'm stuck on, if you don't mind giving it some thought. I ran into this issue in my work on neural network update, but the question I have for you are about epistemic/plausibility models, not neural network models. I'll try to state the question as cleanly as I can, and then if you're interested in knowing *why* I care about all this I'll be happy to explain face-to-face.

Problem Statement

Consider the dynamic epistemic language

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid \mathbf{K}\varphi \mid \mathbf{T}\varphi \mid [P]\varphi$$

$\Box$ is the general universal (for all worlds), though I'll more often use its dual $\Diamond$ (there is a world). $\mathbf{K}$ is knowledge. $\mathbf{T}$ is more interesting — $\mathbf{T}\varphi$ says that the current world is 'minimal' or 'most typical' over worlds satisfying φ. (As far as I can tell, this is not quite the same as the $[\mathsf{best}]$ operator, see Remark 10 in [**?**]). $[P]$ is some dynamic update given by $\mathcal{M} \to \mathcal{M}^{\star}_{P}$ (this is a free variable; the point will be to find the right update).

For the static part of the logic, choose your favorite semantics — plausibility models, evidence models, etc. For now, I'll take Johan's approach from [**?**], which I've been using as a desk reference for all this. Let's assume we have a single-agent plausibility model, with an extra accessibility relation R for knowledge: $\mathcal{M} = \langle W, R, \leq, V\rangle$. $\leq$ is uniform over all states; we do not have a different plausibility relation $\leq_s$ for each state. As usual, $x \leq y$ reads "the agent finds x at least as plausible as y."

Definition 1. The semantics are given by

$$\begin{array}{lll}
\mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\
\mathcal{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w \Vdash \Box\varphi & \text{iff} & \text{for all } u, \mathcal{M}, u \Vdash \varphi \\
\mathcal{M}, w \Vdash \mathbf{K}\varphi & \text{iff} & \text{for all } u \text{ with } wRu, \mathcal{M}, u \Vdash \varphi \\
\mathcal{M}, w \Vdash \mathbf{T}\varphi & \text{iff} & w \text{ is } \leq\text{-minimal over } \{u \mid \mathcal{M}, u \Vdash \varphi\} \\
\mathcal{M}, w \Vdash [P]\varphi & \text{iff} & \mathcal{M}^{\star}_{P}, w \models \varphi
\end{array}$$

I will use the shorthand $[\![\varphi]\!]_{\mathcal{M}} = \{u \mid \mathcal{M}, u \Vdash \varphi\}$, and drop $\mathcal{M}$ when it's understood from context.

I have three very simple neural network updates, and they can each be reduced to this language. I'd love to explain the details in the future, but for now these are the reduction laws I've found "in the wild." (I'll give the $\mathbf{T}$ case for each; for the first two assume $\mathbf{K}$ is not in the language, and for the third the update does not affect $\mathbf{K}$):

Axiom A. $[P]\mathbf{T}\varphi \leftrightarrow (\Diamond(P \wedge \langle\mathbf{T}\rangle\varphi) \wedge \mathbf{T}(P \wedge [P]\varphi)) \vee (\neg\Diamond(P \wedge \langle\mathbf{T}\rangle\varphi) \wedge \mathbf{T}[P]\varphi)$

Axiom B. $[P]\mathbf{T}\varphi \leftrightarrow (\Diamond(\langle\mathbf{T}\rangle P \wedge \langle\mathbf{T}\rangle\varphi) \wedge \mathbf{T}(\langle\mathbf{T}\rangle P \wedge [P]\varphi)) \vee (\neg\Diamond(\langle\mathbf{T}\rangle P \wedge \langle\mathbf{T}\rangle\varphi) \wedge \mathbf{T}[P]\varphi)$

Axiom C. $[\varphi]\mathbf{T}\psi \leftrightarrow \mathbf{T}([\varphi]\psi \wedge (\mathbf{T}\varphi \vee \mathbf{K}(\mathbf{T}\varphi \vee \mathbf{T}[\varphi]\psi)))$

I would like to understand what the neural network updates are doing "classically," i.e. for each neural update, what is an "equivalent" update over possible worlds / plausibility / evidence models? I've been stuck on this point since November (I probably should have reached out sooner). My question for you two is:

Question. Is there a dynamic model update (over your "classical" model of choice) that satisfies Axiom A? (and the same for Axioms B and C.)

Progress So Far

The crucial step of this proof is finding this $\leq$-minimal w in $[\![q]\!]_{\mathcal{M}^\star_p}$. Note that this step does not rely on the well-foundedness of $\leq$ — we can construct a similar model that is not well-founded if we like. But it *does* rely on the fact that $[p]q \leftrightarrow q$ is valid: re-ordering $\leq$ *cannot add or remove* elements from $[\![q]\!]$. In particular, the proof would break if our update could make $[\![q]\!]$ empty or make $[\![q]\!]$ include an infinite descending chain. (But I can't figure out an update that would do these in the right way…)

Corollary 1. No plausibility upgrade can make Axiom B valid.

The proof is a simple extension of the above proof, replacing p with $\langle \mathbf{T} \rangle p$. We can show the same for axiom C, by modifying the construction slightly.

Proposition 2. No plausibility upgrade can make Axiom C valid.

Proof. Consider the propositional instance of Axiom C:

$$[p]\mathbf{T}q \leftrightarrow \mathbf{T}(q \wedge (\mathbf{T}p \vee \mathbf{K}(\mathbf{T}p \vee \mathbf{T}q)))$$

Let $\mathcal{M} \to \mathcal{M}^\star$ be any plausibility upgrade. This time, let $\mathcal{M}$ be

[PICTURE]

Again, $\mathcal{M} \to \mathcal{M}^\star$ only modifies $\leq$, and in particular $[\![q]\!]_{\mathcal{M}^\star_p}$ is finite and nonempty. So there is some w that is $\leq$-minimal over $[\![q]\!]_{\mathcal{M}^\star_p}$. So $\mathcal{M}, w \Vdash [p]\mathbf{T}q$.

TODO □

All the best,