The Modeling Power of Neural Network Models

Problem Statement

These days I've been trying to understand the bigger picture of our work on neural network models. I'll start by asking some suggestive questions, to point us in the right direction. I've been wondering:

Question 1. How do neural network models relate to other models for conditional & modal logic? What about the dynamics — how do policies like Hebbian learning relate to belief revision policies such as lexicographic & conservative upgrade?

Question 2. The FLaNN Group, specifically the work [6, 8, 7, 10] considers neural networks as automata, and asks the question "what functions can different neural networks encode?" Similarly, we consider neural networks as models for logic, and ask the question "what formulas can different neural networks model?" These questions are clearly related — but how, precisely?

Question 3. The FLaNN perspective (neural networks as automata) is one way to characterize the computational power of neural networks. Can we use neural network models to characterize the *modeling* power of neural networks? How does this all relate to the computational and descriptive complexity hierarchies?

Basic Setup and Definitions

My first goal is to compare neural network models against other models. To make the comparison fair, all models will share the basic multi-modal language $\mathcal{L}$, given by

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \{\Box_i\}_{i \in \mathbf{I}} \ \varphi$$

where **I** is some fixed set of indices. I have in mind that each $\Box_i$ represents a different modality per use-case (e.g. belief vs knowledge), although they could also be used to model a multi-agent setting (e.g. agent 1's belief vs agent 2's belief).

Here are some classes of models I'm interested in:

Relational (Kripke) Models. A relational model is $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V\rangle$, where

- W is some finite set of worlds (or states)
- Each $R_i \subseteq W \times W$ (the accessibility relations)
- V : Proposition $\rightarrow \mathcal{P}(W)$ (the valuation function)

Define **Rel** to be the class of all such models, and define $\mathbf{Rel_{S4}}$ to be the class of all such models where each R_i is additionally reflexive and transitive. The semantics for both classes is given by:

$$\begin{array}{lll} \mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\ \mathcal{M}, w \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\ \mathcal{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\ \mathcal{M}, w \Vdash \Box_i\varphi & \text{iff} & \text{for all } u \text{ with } wR_iu, \mathcal{M}, u \Vdash \varphi \end{array}$$

Plausibility Models. A plausibility model, first introduced in [3], is $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V \rangle$, i.e. the models themselves are just relational models. As before, I assume that W is finite, and as with $\mathbf{Rel_{S4}}$, R_i is reflexive and transitive. The key difference is that we interpret $\Box_i \varphi$ to hold in the best (or most plausible) states satisfying φ. Formally, let $\mathsf{best}_{R_i}(S) = \{w \in S \mid \text{For all } u \in S, \neg u R_i w\}$ (the R_i-minimal states over S). We additionally impose the following "smoothness condition" [3] on best_{R_i}:

Postulate. For all models $\mathcal{M}$, $i \in \mathbf{I}$, sets S, and all $w \in W$, if $w \in S$ then either $w \in \mathsf{best}_{R_i}(S)$, or there is some $v R_i w$ better than w that *is* the best, i.e. $v \in \mathsf{best}_{R_i}(S)$.

The new semantics for $\Box_i$ is

$$\mathcal{M}, w \Vdash \Box_i \varphi \quad \text{iff} \quad w \in \mathsf{best}_{R_i}(\llbracket \varphi \rrbracket)$$

where $\llbracket \varphi \rrbracket = \{u \mid \mathcal{M}, u \Vdash \varphi\}$. In practice, plausibility semantics coexist alongside relational semantics, so I allow some $\Box_i \varphi$ to be given relational semantics instead. Let **Plaus** be the class of all such models. Since we include relational operators, note that $\mathbf{Rel_{S4}} \subseteq \mathbf{Plaus}$.

Any plausibility operator $\Box_i$ picks out a corresponding conditional: $\Box_i \varphi \to \psi$ reads "the best φ are ψ," which in the KLM tradition is the semantics for the conditional $\varphi \Rightarrow \psi$.

Neighborhood Models. A neighborhood model is $\mathcal{M} = \langle W, \{f_i\}_{i \in \mathbf{I}}, V \rangle$, where W and V are as before and each $f_i : W \to \mathcal{P}(\mathcal{P}(W))$ is an accessibility *function*. The intuition is that f_i maps each state w to the "formulas" (sets of states) that hold at w. Define **Nbhd** to be the class of all neighborhood models.

Moreover, the *core* of f is $\cap f(x) = \bigcap_{X \in f(w)} X$. As with **Rel**, let $\mathbf{Nbhd_{S4}}$ be the class of all neighborhood models that are additionally reflexive ($\forall w, w \in \cap f(w)$) and transitive ($\forall w$, if $X \in f(w)$ then $\{v \mid X \in f(v)\} \in f(w)$).

The semantics for both classes is the same as the previous classes, except the $\Box_i$ case is now:

$$\mathcal{M}, w \Vdash \Box_i \varphi \quad \text{iff} \quad \llbracket \varphi \rrbracket \in f_i(w)$$

where again $\llbracket \varphi \rrbracket = \{u \mid \mathcal{M}, u \Vdash \varphi\}$.

Weighted Neural Network Models. In our work [2] so far, we've only defined the neural network state operators Prop and Reach. But in principle, we could define other closure operators, each reflecting a different kind of modality or conditional. I want to characterize what a neural network can *in principle model*, and for this we need a more general definition.

A neural network model is $\mathcal{N} = \langle N, \mathsf{bias}, \{E_i\}_{i \in \mathbf{I}}, \{W_i\}_{i \in \mathbf{I}}, \{A\}_{i \in \mathbf{I}}, V \rangle$, where

- N is a finite nonempty set (the set of neurons)
- bias is a fixed node (the bias node)
- Each $E_i \subseteq N \times N$ (the edge relations)
- $W_i : E_i \to \mathbb{Q}$ (the weights for each edge relation)
- $A_i : \mathbb{Q} \to \mathbb{Q}$ (the activation function for each edge relation)
- V : Proposition $\to \mathcal{P}(N)$ (the valuation function)

I assume that each A_i is a binary step function. Using the terminology of [8], this means the net is *saturated*. Later, I will assume one more constraint on the architecture of these nets.

A *state* is just a possible activation pattern of the net. Since our activation functions A_i are binary, either a neuron is active (1) or it is not (0). So states are just binary sets of neurons. Additionally, the bias node is the only node active in every state. (Since it is active in every state, we can assume that no edges go into bias.) Formally, we have

$$\mathsf{State} = \{S \mid S \subseteq N \text{ and } \mathsf{bias} \in S\}$$

I've generalized the definition from previous papers: These neural network models can have multiple kinds of edges (indexed by $i \in \mathbf{I}$) connecting the same nodes, along with their weights and a corresponding activation function for each i. Each choice E_i, W_i, A_i specifies a state transition function from state $S \in \mathcal{P}(W)$ to the next state, given by

$$F_i(S) = S \cup \{n \mid A_i(\sum_{m \in \text{preds}(n)} W_i(m,n) \cdot \chi_S(m)) = 1\}$$

where $\chi_S(m) = 1$ iff $m \in S$ is the indicator function. In other words, $F_i(S)$ is the current state S, along with the set of nodes that are activated by their predecessors in S. Notice that $F_i(S)$ is extensive — once activated, a node will stay activated in the next state.

Let **Net** be the class of all neural network models defined above, with the following additional constraint:

Postulate. I assume that for all $i \in \mathbf{I}$ and all states S, F_i applied to S, i.e.

$$S, F_i(S), F_i(F_i(S)), \ldots, F_i^k(S), \ldots$$

has a (finite) least fixed point, and moreover that it is *unique*. Let $\mathsf{Cl}_i : \mathcal{P}(N) \to \mathcal{P}(N)$ ("closure") be the function that produces that least fixed point. For concreteness, we can say that there is some $k \in \mathbb{N}$ for which

$$\mathsf{Cl}_i(S) = F_i^k(S)$$

This assumption implicitly constrains the allowed neural network architectures: We allow feed-forward nets, as well as certain controlled forms of recurrence. Characterizing nets that have a unique least fixed point is a big open problem.

I can now state the semantics for $\mathcal{N} \in$ **Net**:

$$\begin{array}{lll}
\mathcal{N}, n \Vdash p & \text{iff} & n \in V(p) \\
\mathcal{N}, n \Vdash \neg\varphi & \text{iff} & \mathcal{N}, n \nVdash \varphi \\
\mathcal{N}, n \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{N}, n \Vdash \varphi \text{ and } \mathcal{M}, n \Vdash \psi \\
\mathcal{N}, n \Vdash \Diamond_i \varphi & \text{iff} & n \in \mathsf{Cl}_i(\llbracket\varphi\rrbracket)
\end{array}$$

where $\llbracket\varphi\rrbracket = \{n \mid \mathcal{N}, n \Vdash \varphi\}$. A technical point is that our semantics differ from Hannes' [4] in how we handle propositions and connectives — his semantics are entirely neural, whereas the semantics above handle propositions and connectives classically. He battles with this issue of how to correctly interpret negation; I sidestep this issue by using neural networks for interpreting $\Diamond_i \varphi$ (where the "action" happens), but not for the propositional base.

Any network diffusion operator $\Diamond_i$ picks out a corresponding neural network inference: $\Diamond_i \varphi \leftarrow \psi$ says that on input φ the neural network "answers" with classification ψ. This is analogous to the way a plausibility operator picks out a conditional (I will say more about this later).

In the following examples, I'll walk through common constructions for neural networks. I mentioned before that each choice of E_i, W_i, A_i specifices a transition function $F_i(S)$. Different choices for $F_i(S)$ in turn give different interpretations for the closure function $\mathsf{Cl}_i(S)$. These examples should give you a feel for what sorts of state transitions we can represent with neural networks.

Example: The Graph-Reachability Construction. Say we want to build a neural network

$$\mathcal{N} = \langle N, \mathsf{bias}, \{E_i\}_{i \in \mathbf{I}}, \{W_i\}_{i \in \mathbf{I}}, \{A\}_{i \in \mathbf{I}}, V\rangle$$

Let's consider a particular i. Suppose the graph $\langle N, E_i\rangle$, bias, and evaluation V are given. Pick

$$W_i(m,n) = \begin{cases} 1 & \text{if } mE_i n \\ 0 & \text{otherwise} \end{cases}$$

Then pick $A_i(x) = 1$ iff $x > 0$. Recall that $n \in F_i(S)$ iff n is activated by its predecessors in S (by the weighted sum term, see the discussion earlier). In this case, $n \in F_i(S)$ whenever n is active ($n \in S$) or at least one E_i-predecessor m of n is in S. I call this the graph-reachability construction because the closure $\mathsf{Cl}_i(S)$ gives exactly the nodes graph-reachable from S:

Claim. $\mathsf{Cl}_i(S) = \{n \mid \exists$ an E_i-path from some $m \in S$ to $n\}$

Proof. First, the $(\subseteq)$ direction. Let $n \in \mathsf{Cl}_i(S) = F_i^k(S)$ for some $k \in \mathbb{N}$. We proceed by induction on k.

Base Step. $n \in F_i^0(S) = S$. So there is a trivial E_i-path (length $= 0$) from $n \in S$ to itself.

Inductive Step. Let $k \geq 0$. We have $n \in F_i^k(S) = F_i(F_i^{k-1}(S))$. By construction of F_i, we have two cases: Either $n \in F_i^{k-1}(S)$ or at least one E_i-predecessor x of n is in $F_i^{k-1}(S)$. In the first case, our inductive hypothesis gives a path from some $m \in S$ to n. In the second case, our inductive hypothesis gives a path from some $m \in S$ to x. But since xE_in, we can extend this path to be from m to n.

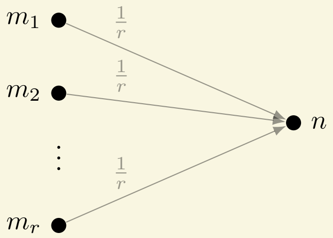
As for the $(\supseteq)$ direction, suppose there is an E_i-path from some $m \in S$ to n. We proceed by induction on the length of that path.

Base Step. The path is trivial, i.e. has length 0. So $n \in S$. But $S = F_i^0(S) \subseteq \mathsf{Cl}_i(S)$, and so $n \in \mathsf{Cl}_i(S)$.

Inductive Step. Say the path is of length $l \geq 0$. Let x be some immediate E_i-predecessor of n. By the inductive hypothesis, $x \in \mathsf{Cl}_i(S)$, and so $x \in F_i^k(S)$ for some natural k. But since x is an E_i-predecessor of n, by construction of F_i, $n \in F_i(F_i^k(S)) = F_i^{k+1}(S)$. Since $\mathsf{Cl}_i(S)$ is a closure, it includes $F_i^{k+1}(S)$. So $n \in \mathsf{Cl}_i(S)$.

□

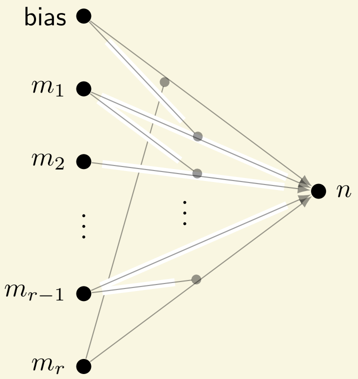
Example: The Social Majority Construction. As before, we want to build a neural network $\mathcal{N}$ where the graph $\langle N, E_i\rangle$, bias, and evaluation V are given. This time, pick $W_i(m,n) = \frac{1}{|\text{preds}(n)|}$, and then pick $A_i(x) = 1$ iff $x \geq \frac{1}{2}$. Visually, for each node n and its predecessors $m_1, \ldots, m_r$ we have



This gives us $n \in F_i(S)$ if n is already active ($n \in S$) or if the majority (more than half) of E_i-predecessors are in S. In this case, the closure Cl_i can be interpreted as the diffusion of an opinion or attitude through a social network. This is one of the choices that [1] consider for modelling influence in social networks.

Example: The Not-Every Construction. Here is an interesting construction that Hannes uses in [4] to prove completeness for weighted neural network models. He does this by way of inhibition nets, i.e. nets with inhibitory edges that block excitatory edges.

As before, we want to build a neural network $\mathcal{N}$ where the graph $\langle N, E_i\rangle$, bias, and evaluation V are given. Here's the *inhibition net* construction: First, create an edge from bias to every n that is not E_i-minimal (in other words, if n has any predecessors at all, then bias is one of them). Then for each node n and its predecessors $\mathsf{bias} = m_0, m_1, \ldots, m_r$, connect inhibition edges as follows.



That is, each node m_i is inhibited by m_{i-1} ($\textsf{bias} = m_0$ inhibited by m_r). This has the following effect: if all m_i activate, they each inhibit each other, and so n does not activate. If only *some* m_i activate, then there is some m_i that is uninhibited, and so n activates. And finally, since bias is always active we cannot have *no* m_i active. In other words, $n \in F_i(S)$ iff n is already active ($n \in S$), or *at least one, but not all* predecessors m_i are in S.

We can simulate this effect with weighted neural networks. Create an edge from bias to every n that is not E_i-minimal. Then pick $W_i(m, n) = \frac{1}{|\text{preds}(n)|+1}$ (the extra $+1$ accounts for the bias). Finally, pick $A_i(x) = 1$ iff $x < 1$. Take a moment to check that $n \in F_i(S)$ iff $n \in S$, or at least one, but not all predecessors m are in S.

Consider the best function from before, but over E_i. That is, $\textsf{best}_{E_i}(S)$ is the set of E_i-minimal nodes in S. It turns out that, *if* E_i is transitive (which is true in our use case), the closure function $\textsf{Cl}_i$ for the not-every construction is precisely the *dual* of $\textsf{best}_{E_i}$:

Claim. (Leitgeb, [4]) Suppose E_i is transitive. Then for all S, $\textsf{Cl}_i(S) = (\textsf{best}_{E_i}(S^{\complement}))^{\complement}$

Proof. First, the ($\subseteq$) direction. Let $n \in \textsf{Cl}_i(S) = F_i^k(S)$ for some $k \in \mathbb{N}$. We proceed by induction on k.

Base Step. $n \in F_i^0(S) = S$. By $\textsf{best}_{E_i}$-inclusion, $\textsf{best}_{E_i}(S^{\complement}) \subseteq S^{\complement}$. Flipping this containment gives us $S \subseteq \textsf{best}_{E_i}(S^{\complement})^{\complement}$. So $n \in \textsf{best}_{E_i}(S^{\complement})^{\complement}$.

Inductive Step. Let $k \geq 0$. We have $n \in F_i^k(S) = F_i(F_i^{k-1}(S))$. By construction of F_i, we have two cases: n is already active ($n \in F_i^{k-1}(S)$), or some predecessor is $m \in F_i^{k-1}(S)$, and not all predecessors are. In the first case, our inductive hypothesis says $n \in \textsf{best}_{E_i}(S^{\complement})^{\complement}$. In the latter case, the inductive hypothesis gives $m \in \textsf{best}_{E_i}(S^{\complement})^{\complement}$. From here we split by cases: either $m \in S$ or $m \notin S$.

Case: $m \in S$**.** In this case, we trivially have $n \notin \textsf{best}_{E_i}(S^{\complement})$, since n is not even in $S^{\complement}$.

Case: $m \notin S$**.** Since $m \in S^{\complement}$ but $m \notin \textsf{best}_{E_i}(S^{\complement})$, by the smoothness condition there must be some better x with xE_im that *is* the best, i.e. $x \in \textsf{best}_{E_i}(S^{\complement})$. By best-inclusion, $x \in S^{\complement}$. Well, xE_imE_in, and since E_i is transitive we have xE_in. And so we have an $x \in S^{\complement}$ that is E_i-better than n. So $n \notin \textsf{best}_{E_i}(S^{\complement})$, which is what we wanted to show.

As for the ($\supseteq$) direction, suppose contrapositively that $n \notin \textsf{Cl}_i(S)$. Note that $n \notin S$, by Cl-inclusion. First, I claim that every predecessor m of n is in S. Suppose not, i.e. suppose that some predecessor $m \notin S$. Note that we always have $\textsf{bias} \in S$. By construction, we also have $\textsf{bias}E_in$. So n has one predecessor, m, not in S, and another predecessor, bias, in S. In other words, some but not all predecessors of n are in S. By construction of F_i, $n \in F_i(S)$. But $F_i(S) \subseteq \textsf{Cl}_i(S)$, so this contradicts $n \notin \textsf{Cl}_i(S)$.

So every predecessor m of n is in S. But this implies that any $m \notin S$ cannot be an E_i-predecessor of n. In other words, $\forall m \in S^{\complement}, \neg mE_in$. Since $n \in S^{\complement}$ from before, we have $n \in \textsf{best}_{E_i}(S)$ by the definition of best. This concludes this direction of the proof. □

Measuring Modeling Power.

To compare the modeling power of neural networks with other logic models, I need to pick a measure of complexity. There are three questions about model complexity we can answer this way:

Satisfiability. "What formulas can a class of models in-principle model (satisfy)?"

Validity. "What formulas *must* a class of models satisfy (make valid)?"

Definability. "What properties (or formal languages) can a class of models *define*?"

I will focus on comparing *satisfiability*, in part because for neural networks it is a useful proxy for the question "what functions can a neural network architecture in-principle compute?" I'm going to state definability for the sake of completeness, and in the future I want to relate it to my work (to connect with descriptive complexity of neural networks). But at the moment, I don't have anything interesting to say about definability. (Maybe I'll add something here when I do.)

Definition 1. Let $\mathcal{M}$ be any model whatsoever with universe W and satisfaction relation $\Vdash$. $\mathcal{M} \models \varphi$ if for all $w \in W$, $\mathcal{M}, w \Vdash \varphi$.

Definition 2. Let $\mathscr{C}_1, \mathscr{C}_2$ be two classes of models. A *class embedding* $f : \mathscr{C}_1 \to \mathscr{C}_2$ is an injective function such that for all models $\mathcal{M} \in \mathscr{C}_1$, and formulas $\varphi \in \mathcal{L}$, $\mathcal{M} \models \varphi$ iff $f(\mathcal{M}) \models \varphi$. If there is an embedding from $\mathscr{C}_1$ into $\mathscr{C}_2$, we write $\mathscr{C}_1 \hookrightarrow \mathscr{C}_2$.

Satisfiability and validity are the measures of complexity preferred in model theory. We define them as follows. Note that the two are duals of one another — intuitively, $\mathrm{Sat}(\mathscr{C})$ puts more "powerful" models on top, whereas $\mathrm{Th}(\mathscr{C})$ puts them on the bottom.

Definition 3. Let $\mathscr{C}$ be a class of models, $\mathcal{M}$ be a model in $\mathscr{C}$, and $\mathcal{L}$ be a language. The *formulas satisfied by* $\mathcal{M}$ over $\mathcal{L}$ is given by $\mathrm{Sat}(\mathcal{M}) = \{\varphi \in \mathcal{L} \mid \mathcal{M} \models \varphi\}$. The *formulas satisfiable by class* $\mathscr{C}$ over $\mathcal{L}$ is given by

$$\mathrm{Sat}(\mathscr{C}) = \{\varphi \in \mathcal{L} \mid \text{there is some } \mathcal{M} \in \mathscr{C} \text{ such that } \mathcal{M} \models \varphi\}$$

Definition 4. Let $\mathscr{C}$ be a class of models, and $\mathcal{L}$ be a language. The *theory* of $\mathscr{C}$ over $\mathcal{L}$ is given by

$$\mathrm{Th}(\mathscr{C}) = \{\varphi \in \mathcal{L} \mid \text{for all } \mathcal{M} \in \mathscr{C} \text{ we have } \mathcal{M} \models \varphi\}$$

Proposition 1. Suppose $\mathscr{C}_1 \hookrightarrow \mathscr{C}_2$. We have $\mathrm{Sat}(\mathscr{C}_1) \subseteq \mathrm{Sat}(\mathscr{C}_2)$, and similarly, $\mathrm{Th}(\mathscr{C}_2) \subseteq \mathrm{Th}(\mathscr{C}_1)$.

Proof. **TODO** □

Definability is the measure preferred in descriptive complexity. It's usually considered a measure of complexity of a *language*, while the models are kept fixed. But we can also consider it a measure of complexity of models, with the language kept fixed.

Definition 5. Let $\mathscr{C}$ be a class of models. The *properties (problems) definable by* $\mathscr{C}$ *over* $\mathcal{L}$ is given by

$$\mathrm{Def}(\mathscr{C}) = \{P \subseteq \mathscr{C} \mid \text{there exists } \varphi \in \mathcal{L} \text{ such that for all } \mathcal{M} \in \mathscr{C}, \mathcal{M} \in P \text{ iff } \mathcal{M} \models \varphi\}$$

Example. Here's an example that's a sort of tutorial for comparing the modeling power of different model classes using $\mathrm{Sat}(\mathscr{C})$. Let's compare relational models **Rel** with neighborhood models **Nbhd** over the static language. Our goal is to show that neighborhood models are strictly more powerful than relational models: $\mathrm{Sat}(\textbf{Rel}) \subset \mathrm{Sat}(\textbf{Nbhd})$.

First, here are some formulas that are valid in every relational model:

- $\Box_i(\varphi \land \psi) \to (\Box_i\varphi \land \Box_i\psi)$

- $(\Box_i \varphi \wedge \Box_i \psi) \to \Box_i(\varphi \wedge \psi)$
- $\Box_i \top$

This means that *no* $\mathcal{M} \in$ **Rel** can satisfy their negations (for details, see [9], page 8). But neighborhood models can: Let $\mathcal{M} = \langle W, f, V\rangle$ be a neighborhood model with $W = \{a, b, c\}$, propositions $\{p, q\}$ with $V(p) = \{a, b\}, V(q) = \{b, c\}$, and f given by

$$\begin{tikzcd}
\{a,c\} & \{c\} & \{b\} & \emptyset & \{a\} & \{b,c\} & \{b\} \\
 & a \arrow[ul] \arrow[u] \arrow[ur] & & b \arrow[ul] \arrow[u] \arrow[ur] & & c \arrow[ul] \arrow[u] \arrow[ur] &
\end{tikzcd}$$

For all $w \in W$ we have $[\![p]\!] \cap [\![q]\!] = \{b\} \in f(w)$, but also $[\![p]\!] = \{a, b\} \notin f(w)$. This means that for all w, $\mathcal{M}, w \nVdash \Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)$ — in other words, $\mathcal{M} \models \neg(\Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi))$. Consequently, this formula is in Sat(**Nbhd**), but *not* in Sat(**Rel**).

Moreover, we can embed relational models into neighborhood models. That is, **Rel** $\hookrightarrow$ **Nbhd**, which gives us Sat(**Rel**) $\subseteq$ Sat(**Nbhd**). Let $\varphi \in \mathcal{L}$ and suppose $\mathcal{M} \models \varphi$ for some $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V\rangle \in$ **Rel**. Eric Pacuit in [9], page 47 shows how to construct an equivalent neighborhood model: Let $f(\mathcal{M}) = \langle W, \{f_i\}_{i \in \mathbf{I}}, V\rangle$, where each $f_i(w) = \{X \mid \{v \mid wR_i v\} \subseteq X\}$. All we need to show is $\mathcal{M}' \models \varphi$, but we are able to prove something even stronger:

Claim. For all φ and all w, $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}', w \Vdash \varphi$.

Proof. By induction on φ. The key case is $\Box_i \varphi$.

($\to$): Suppose $\mathcal{M}, w \Vdash \Box_i\varphi$, i.e. for all u with wR_iu, $\mathcal{M}, u \Vdash \varphi$. It follows that $\{u \mid wR_iu\} \subseteq [\![\varphi]\!]$. By construction of $\mathcal{M}'$, this means $[\![\varphi]\!] \in f_i(w)$, and so $\mathcal{M}', w \Vdash \Box_i\varphi$.

($\leftarrow$): Now suppose $\mathcal{M}', w \Vdash \Box_i\varphi$, i.e. $[\![\varphi]\!] \in f_i(w)$. By construction of $\mathcal{M}'$, we have $\{u \mid wR_iu\} \subseteq [\![\varphi]\!]$. So for all u with wR_iu, $\mathcal{M}, u \Vdash \varphi$. This gives us $\mathcal{M}, w \Vdash \Box_i\varphi$. □

Known Results (The Static Case)

Recall the first part of Question 1: “how do neural network models relate to other models for conditional & modal logic?” The static case is already understood. I’ll collect the known results in this section.

We can see how the modeling power of neural network models compares against known models in logic by arranging their satisfiable sets in a “modeling hierarchy.” To make the comparison with neural networks fair, I will only consider the reflexive and transitive variants $\mathbf{Rel_{S4}}$, $\mathbf{Nbhd_{S4}}$ of relational and neighborhood models. Over the static language (no dynamic operators), we have:

$$\text{Sat}(\mathbf{Rel_{S4}}) \subset \text{Sat}(\mathbf{Plaus}) = \text{Sat}(\mathbf{Net}) \subset \text{Sat}(\mathbf{Nbhd_{S4}})$$

Both **Plaus**-inclusions are folklore. Rather than showing these, I will instead prove $\text{Sat}(\mathbf{Rel_{S4}}) \subset \text{Sat}(\mathbf{Net})$ and $\text{Sat}(\mathbf{Net}) \subseteq \text{Sat}(\mathbf{Nbhd_{S4}})$, putting the emphasis on how neural networks compare to the other models. The inclusion $\text{Sat}(\mathbf{Plaus}) = \text{Sat}(\mathbf{Net})$ is already known, but the backwards direction has never been proven with an explicit model construction. So although the results in this section are already known, three of the four proofs I give here are totally new (indicated by ☞).

Proposition 2. **(☞)** $\text{Sat}(\mathbf{Rel_{S4}}) \subset \text{Sat}(\mathbf{Net})$

Proof. First, let’s show inclusion. Let $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V\rangle \in \mathbf{Rel_{S4}}$ be a model satisfying φ. We construct the neural network model $\mathcal{N} = \langle N, \mathsf{bias}, \{E_i\}_{i \in \mathbf{I}}, \{W_i\}_{i \in \mathbf{I}}, \{A_i\}_{i \in \mathbf{I}}, V\rangle$ as follows. First, let $N = W$

and keep the propositional evaluation V the same. We first *reverse* R_i, and then do the graph-reachability construction. In other words, let uE_iv iff vR_iu,

$$W_i(m,n) = \begin{cases} 1 & \text{if } mE_in \\ 0 & \text{otherwise} \end{cases}$$

Then pick $A_i(x) = 1$ iff $x > 0$. Recall that this choice gives us

$$\mathsf{Cl}_i(S) = \{w \mid \exists \text{ an } E_i\text{-path from some } u \in S \text{ to } w\}$$

From here, we need to show that for all $\varphi, w, \mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$. We do this by induction on φ. The key inductive case is $\Diamond_i\varphi$.

$$\begin{array}{lll} \mathcal{M}, w \Vdash \Diamond_i\varphi & \text{iff} \quad \exists \text{ some } u \text{ with } wR_iu \text{ and } \mathcal{M}, u \Vdash \varphi & \text{(By relational semantics)} \\ & \text{iff} \quad \exists \text{ some } u \text{ with } uE_iw \text{ and } \mathcal{M}, u \Vdash \varphi & \text{(Since we reversed } R_i\text{)} \\ & \text{iff} \quad \exists \text{ some } u \text{ with } uE_iw \text{ and } \mathcal{N}, u \Vdash \varphi & \text{(Inductive hypothesis)} \\ & \text{iff} \quad \exists \text{ an } E_i\text{-path from some } u \in [\![\varphi]\!]_{\mathcal{N}} \text{ to } w & \text{(Defn of } [\![\varphi]\!] \text{ and since} \\ & & \quad E_i \text{ is refl and trans)} \\ & \text{iff} \quad w \in \mathsf{Cl}_i([\![\varphi]\!]_{\mathcal{N}}) & \text{(By construction)} \\ & \text{iff} \quad \mathcal{N}, w \Vdash \Diamond_i\varphi & \end{array}$$

We conclude that for our particular φ, $\mathcal{M} \models \varphi$ implies $\mathcal{N} \models \varphi$.

As for strictness, nonmonotonicity $\neg(\Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)) \in \mathrm{Sat}(\textbf{Net})$, yet it is not satisfiable by $\mathcal{M} \in \textbf{Rel}_{\textbf{S4}}$. □

Proposition 3. **(☞)** $\mathrm{Sat}(\textbf{Plaus}) \subset \mathrm{Sat}(\textbf{Nbhd}_{\textbf{S4}})$

Proof. **TODO** □

LEGACY:

Proposition 4. $\mathrm{Sat}(\textbf{Plaus}) \subset \mathrm{Sat}(\textbf{Nbhd}_{\textbf{S4}})$

Proof. First, let's show inclusion. Let $\mathcal{M} = \langle W, \{R_i\}_{i\in\mathbf{I}}, V\rangle$ be a plausibility model satisfying φ. We construct the neighborhood model $\mathcal{M}' = \langle W, \{f_i\}_{i\in\mathbf{I}}, V\rangle$, where f_i is given as follows. If $\Box_i$ is given relational semantics, then $f_i(w) = \{X \mid \{v \mid wR_iv\} \subseteq X\}$ as in our example above. But if $\Box_i$ has plausibility semantics,

$$f_i(w) = \{X \in \mathcal{P}(W) \mid w \in \mathsf{best}_{R_i}(X)\}$$

First, we need to check that this is in fact in $\textbf{Nbhd}_{\textbf{S4}}$, i.e. these choices of f_i are reflexive and transitive.

f_i **is reflexive (relational case).** We want to show that for all w, $w \in \cap f_i(w)$. First, since R_i is reflexive, for all w, wR_iw. So $w \in \{u \mid wR_iu\}$. But this means $\{u \mid wR_iu\}$ implies $w \in X$. Applying our choice of f_i, we have $X \in f(w)$ implies $w \in X$. This immediately gives $w \in \bigcap_{X\in f_i(w)} X = \cap f_i(w)$.

f_i **is transitive (relational case).** We want to show that for all w and X, if $X \in f_i(w)$ then $\{u \mid X \in f_i(u)\} \in f_i(w)$. Suppose that $X \in f_i(w)$. By definition, $\{u \mid wR_iu\} \subseteq X$. Now let u, v be arbitrary, and suppose wR_iu, uR_iv. Since R_i is transitive, wR_iv. But then $v \in \{u \mid wR_iu\} \subseteq X$, so $v \in X$. Since u and v were chosen arbitrarily, in general we have

$$\{u \mid wR_iu\} \subseteq \{u \mid \{v \mid uR_iv\} \subseteq X\}$$

But by choice of f_i, this is exactly $\{u \mid X \in f(u)\} \in f(w)$.

f_i **is reflexive (plausibility case).** We want to show that for all w, $w \in \cap f_i(w)$. First, for all sets X we have $\mathsf{best}_{R_i}(X) \subseteq X$. In other words, $w \in \mathsf{best}_{R_i}(X)$ implies $w \in X$. So $X \in f_i(w)$ implies $w \in X$. But this immediately gives $w \in \bigcap_{X \in f_i(w)} X = \cap f_i(w)$.

f_i **is transitive (plausibility case).** We want to show that for all w and X, if $X \in f_i(w)$ then $\{u \mid X \in f_i(u)\} \in f_i(w)$. Suppose that $X \in f_i(w)$. By definition, $w \in \mathsf{best}_{R_i}(X)$. But by idempotence of best, $w \in \mathsf{best}_{R_i}(\mathsf{best}_{R_i}(X))$. Applying our choice of f_i to the inner best, we get $w \in \mathsf{best}_{R_i}(\{u \mid X \in f(u)\})$. Applying the definition once more, we have $\{u \mid X \in f(u)\} \in f(w)$.

Next, we will show that for all φ, w, $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{M}', w \Vdash \varphi$. We do this by induction on φ. The key inductive case is $\Box_i \varphi$. The relational case is handled by the example in the previous section. For plausibility operators $\Box_i$, we have

$$\begin{array}{lll} \mathcal{M}, w \Vdash \Box_i \varphi & \text{iff} & w \in \mathsf{best}_{R_i}(\llbracket \varphi \rrbracket) \\ & \text{iff} & \llbracket \varphi \rrbracket \in f_i(w) \\ & \text{iff} & \mathcal{M}', w \Vdash \Box_i \varphi \end{array}$$

We conclude that for our particular φ, $\mathcal{M} \models \varphi$ implies $\mathcal{M}' \models \varphi$.

Strictness is easy: nonreflexivity $\neg(\Box\varphi \to \varphi) \in \text{Sat}(\textbf{Nbhd})$, but $\neg(\Box\varphi \to \varphi) \notin \text{Sat}(\textbf{Plaus})$ (since both relational *and* plausibility operators are always reflexive). □

Let's move on to the neural network inclusions. I claim that $\text{Sat}(\textbf{Net}) = \text{Sat}(\textbf{Plaus})$: neural network and plausibility models are equally powerful (up to inference in the static language). The $(\to)$ direction follows from Hannes' completeness result [4, 5] — he uses the *not-every construction* from before to build a neural net from a plausibility model. The $(\leftarrow)$ direction is essentially the soundness result [4, 5]. But explicitly building a plausibility model from a neural net has not been done, as far as I know. So the proof of $(\leftarrow)$ is my own.

Proposition 5. $\text{Sat}(\textbf{Plaus}) \subseteq \text{Sat}(\textbf{Net})$

Proof. Let $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V \rangle \in$ **Plaus** be a plausibility model satisfying φ. We construct the neural network model $\mathcal{N} = \langle N, \mathsf{bias}, \{E_i\}_{i \in \mathbf{I}}, \{W_i\}_{i \in \mathbf{I}}, \{A_i\}_{i \in \mathbf{I}}, V \rangle$ as follows. First, let $N{=}W$, let bias be a fresh node, and keep the propositional evaluation V the same. If $\Diamond_i$ is given relational semantics, then we first *reverse* R_i, and then do the graph-reachability construction. In other words, let uE_iv iff vR_iu,

$$W_i(u, v) = \begin{cases} 1 & \text{if } uE_iv \\ 0 & \text{otherwise} \end{cases}$$

Then pick $A_i(x) = 1$ iff $x > 0$. Recall that this choice gives us

$$\mathsf{Cl}_i(S) = \{w \mid \exists \text{ an } E_i\text{-path from some } u \in S \text{ to } w\}$$

If instead $\Diamond_i$ is given plausibility semantics, we simply do the "not-every" construction. In other words, let E_i be R_i (we do not reverse it), and create an edge from bias to every n that is not E_i-minimal. Then pick $W_i(m, n) = \frac{1}{|\text{preds}(n)|+1}$. Finally, pick $A_i(x) = 1$ iff $x < 1$. Recall that this choice gives us (since $E_i = R_i$ and R_i is transitive):

$$\mathsf{Cl}_i(S) = \mathsf{best}_{R_i}(S^{\complement})^{\complement}$$

From here, we need to show that For all φ, w, $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$. We do this by induction on φ. The key inductive case is $\Diamond_i \varphi$. We have the two cases:

$\Diamond_i$ is given relational semantics.

$$\begin{array}{rlll}
\mathcal{M}, w \Vdash \Diamond_i \varphi & \text{iff} & \exists \text{ some } u \text{ with } wR_iu \text{ and } \mathcal{M}, u \Vdash \varphi & \text{(By relational semantics)} \\
& \text{iff} & \exists \text{ some } u \text{ with } uE_iw \text{ and } \mathcal{M}, u \Vdash \varphi & \text{(Since we reversed } R_i\text{)} \\
& \text{iff} & \exists \text{ some } u \text{ with } uE_iw \text{ and } \mathcal{N}, u \Vdash \varphi & \text{(Inductive hypothesis)} \\
& \text{iff} & \exists \text{ an } E_i\text{-path from some } u \in [\![\varphi]\!]_{\mathcal{N}} \text{ to } w & \text{(Defn of } [\![\varphi]\!] \text{ and since} \\
& & & \quad E_i \text{ is refl and trans)} \\
& \text{iff} & w \in \mathsf{Cl}_i([\![\varphi]\!]_{\mathcal{N}}) & \text{(By construction)} \\
& \text{iff} & \mathcal{N}, w \Vdash \Diamond_i \varphi &
\end{array}$$

$\Diamond_i$ is given plausibility semantics.

$$\begin{array}{rlll}
\mathcal{M}, w \Vdash \Diamond_i \varphi & \text{iff} & \mathcal{M}, w \Vdash \neg \Box_i \neg \varphi & \\
& \text{iff} & w \in \mathsf{best}_{R_i}(\{u \mid \mathcal{M}, u \Vdash \varphi\}^{\mathsf{C}})^{\mathsf{C}} & \text{(By plausibility semantics)} \\
& \text{iff} & w \in \mathsf{best}_{R_i}(\{u \mid \mathcal{N}, u \Vdash \varphi\}^{\mathsf{C}})^{\mathsf{C}} & \text{(Inductive hypothesis)} \\
& \text{iff} & w \in \mathsf{best}_{R_i}([\![\varphi]\!]_{\mathcal{N}}^{\mathsf{C}})^{\mathsf{C}} & \text{(Defn of } [\![\varphi]\!]\text{)} \\
& \text{iff} & w \in \mathsf{Cl}_i([\![\varphi]\!]_{\mathcal{N}}) & \text{(By construction)} \\
& \text{iff} & \mathcal{N}, w \Vdash \Diamond_i \varphi &
\end{array}$$

We conclude that for our particular φ, $\mathcal{M} \models \varphi$ implies $\mathcal{N} \models \varphi$. □

Proposition 6. **(☞)** $\mathrm{Sat}(\mathbf{Net}) \subseteq \mathrm{Sat}(\mathbf{Plaus})$

Proof. Let $\mathcal{N} = \langle N, \mathsf{bias}, \{E_i\}_{i \in \mathbf{I}}, \{W_i\}_{i \in \mathbf{I}}, \{A_i\}_{i \in \mathbf{I}}, V \rangle$ be a neural network model satisfying φ. For each i, we have a closure operator Cl_i that is specified by the particular choice of E_i, W_i, A_i. We construct the plausibility model $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V \rangle$, where $W = N$, V is kept the same, and

$$uR_iv \text{ iff } u \in \bigcap_{v \in \mathsf{Cl}_i(X^{\mathsf{C}})} X$$

It helps to think of this as a kind of topological "core" construction. As with the "not-every" construction from before, the best_{R_i} we get from this construction is precisely the dual of the neural network closure Cl_i:

Claim. For all S, $\mathsf{best}_{R_i}(S) = (\mathsf{Cl}_i(S^{\mathsf{C}}))^{\mathsf{C}}$

Proof. ($\rightarrow$) Suppose $w \in \mathsf{best}_{R_i}(S)$. By definition, $w \in S$ and for all $u \in S$, $\neg uR_iw$. By construction of R_i, $u \notin \cap_{w \in \mathsf{Cl}_i(X^{\mathsf{C}})} X$. Since $w \in S$, we have in particular $w \notin \cap_{w \in \mathsf{Cl}_i(X^{\mathsf{C}})} X$. **Wait, did I mess up the negation-intersection???** We want to show that $w \notin \mathsf{Cl}_i(S^{\mathsf{C}})$; suppose for contradiction that $w \in \mathsf{Cl}_i(S^{\mathsf{C}})$. But then it follows from our choice of R_i that $w \notin S$, which is a contradiction. This concludes this direction of the proof.

($\leftarrow$) **TODO** □

For convenience, I'll deal with $\Box_i$ modalities. All $\Box_i$ operators are given plausibility semantics. From here, I want to show:

Claim. For all $\varphi, w, \mathcal{N}, w \Vdash \varphi$ iff $\mathcal{M}, w \Vdash \varphi$.

Proof. By induction on φ. The key inductive case is $\Box_i \varphi$, which each have plausibility semantics:

$$\begin{array}{rlll}
\mathcal{N}, w \Vdash \Box_i \varphi & \text{iff} & w \in (\mathsf{Cl}_i([\![\varphi]\!]_{\mathcal{N}}^{\mathsf{C}}))^{\mathsf{C}} & \text{(By neural network semantics)} \\
& \text{iff} & w \in (\mathsf{Cl}_i(\{u \mid \mathcal{N}, u \Vdash \varphi\}^{\mathsf{C}}))^{\mathsf{C}} & \text{(Defn of } [\![\varphi]\!]\text{)} \\
& \text{iff} & w \in (\mathsf{Cl}_i(\{u \mid \mathcal{M}, u \Vdash \varphi\}^{\mathsf{C}}))^{\mathsf{C}} & \text{(Inductive hypothesis)} \\
& \text{iff} & w \in \mathsf{best}_{R_i}(\{u \mid \mathcal{M}, u \Vdash \varphi\}) & \text{(By construction)} \\
& \text{iff} & \mathcal{M}, w \Vdash \Box_i \varphi & \text{(By plausibility semantics)}
\end{array}$$

□

We conclude that for our particular φ, $\mathcal{M} \models \varphi$ implies $\mathcal{N} \models \varphi$. □

Progress So Far (The Dynamic Case)

Dynamic Models. I would also like to compare neural network *update* against other model updates — what kind of updates are neural networks capable of modelling, and how powerful are they in this sense? We can "dynamify" each of the models above using the dynamic epistemic logic trick. First, we extend our language $\mathcal{L}$ to $\mathcal{L}^\star$, which includes dynamic modal operators:

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \{\Box_i\}_{i \in \mathbf{I}} \ \varphi \mid \{[\varphi]_j\}_{j \in \mathbf{J}} \ \psi$$

where **J** is a new set of indices. As before, the idea is that each $[\varphi]_i$ represents a different update per use-case, although taking $\mathbf{I} = \mathbf{J}$ these could also be used to model different updates per agent.

I will define the semantics for dynamic models by example. First, consider **Rel**. For any $\mathcal{M} \in \mathbf{Rel}$ and $S \in \mathcal{P}(W)$, we can define a variety of dynamic update functions $\{\mathsf{Update}_j\}_{j \in \mathbf{J}}$, where each $\mathsf{Update}_j : \mathbf{Rel} \times \mathcal{P}(W) \to \mathbf{Rel}$. A classical example is public announcement, where $\mathsf{Update}(\langle W, R, V \rangle, S) = \langle W \cap S, R, V \rangle$. Note that we're using sets S as input rather than formulas, although the choice of one over the other is just my preference.

Also note that these updates don't depend on the current world w. I've read a good bit of [11] (a great book on dynamic epistemic logics), and while Johan includes dependence on w I haven't yet run across an update that uses this extra information. I'll drop it for now, but if we need it later I can add it back in.

The semantics for $\mathcal{M} \in \mathbf{Rel}$ over the dynamic language is exactly as before, with an additional case for $[\varphi]_j \psi$. We just interpret $[\varphi]_j \psi$ as "ψ holds after updating by φ":

$$\mathcal{M}, w \Vdash [\varphi]_j \psi \quad \text{iff} \quad \mathcal{M}^\star_{[\![\varphi]\!]}, w \Vdash \psi$$

We make the same move for each of the other model classes. For all classes $\mathscr{C}$, define $\mathscr{C}^\star$ to be the class of all such models with the new semantics over the dynamic language (e.g. $\mathbf{Rel}^\star$).

The dynamic case is much more interesting. For a class of models $\mathscr{C}$, $\mathrm{Sat}(\mathscr{C}^\star)$ captures those dynamic updates that $\mathscr{C}$ is capable of modelling. This gets at the second part of Question 1: What updates are possible over neural networks? What updates are possible over plausibility models? What is the relationship between neural network updates and plausibility model updates?

I'm not aware of any work that compares different model classes' capacity to model different dynamic updates — this is unexplored turf. I'll take a first stab by **conjecturing** the following:

$$\begin{tikzcd}
 & & \mathrm{Sat}(\mathbf{Nbhd}^\star_{\mathbf{S4}}) \arrow[drr, phantom, "\supset" sloped] & & \\
\mathrm{Sat}(\mathbf{Plaus}^\star) \arrow[urr, phantom, "\subset" sloped] \arrow[drr, phantom, "\supset" sloped] \arrow[rr, dash] & & \textcircled{?} \arrow[rr, dash] & & \mathrm{Sat}(\mathbf{Net}^\star) \\
 & & \mathrm{Sat}(\mathbf{Rel}^\star_{\mathbf{S4}}) \arrow[urr, phantom, "\subset" sloped] & &
\end{tikzcd}$$

I expect that the outermost inclusions are going to be easy(ish). The crucial inclusion is $\textcircled{?}$, which I'm not willing to conjecture on yet. There are exactly four possibilities, and fortunately for us all of them are exciting:

$\mathrm{Sat}(\mathbf{Plaus}^\star) = \mathrm{Sat}(\mathbf{Net}^\star)$. If the two are equal, this work creates a bridge between plausibility models and neural networks. It would suggest that two groups of mostly independent researchers arrived at the same representation for learning. And in this case, it would be nice to explicitly see (1) what neural network updates radical and conservative upgrade correspond to, and (2) what "classical" plausibility updates correspond to network policies such as Hebbian learning or backpropagation.

$\mathrm{Sat}(\mathbf{Plaus}^\star) \subset \mathrm{Sat}(\mathbf{Net}^\star)$**.** In this case, we discover that neural networks are strictly more powerful than plausibility models: for every plausibility upgrade there is a neural network equivalent, but not the other way around. What kind of neural network policies cannot be represented as plausibility upgrade? Is this what makes Hebbian learning or backpropagation special? This possibility offers a theoretical justification for why we use neural networks in machine learning rather than plausibility models, and perhaps why classical models historically have not had the same kind of success as neural networks.

$\mathrm{Sat}(\mathbf{Plaus}^\star) \supset \mathrm{Sat}(\mathbf{Net}^\star)$**.** This flips the previous scenario, and perhaps goes counter to common wisdom. Instead we discover that plausibility models can represent any kind of update neural networks can. What kind of plausiblity updates can't be represented with a neural network? This case offers a new direction for machine learning researchers to look for new learning algorithms, and suggests that they should adopt a different model.

$\mathrm{Sat}(\mathbf{Plaus}^\star)$ **and** $\mathrm{Sat}(\mathbf{Net}^\star)$ **are incomparable.** This is the case that I personally believe is most likely. As before, we can ask: what specific update policies aren't representable with the other model? But we see that neither plausibility models nor neural networks are as rich as we would like them to be. This puts us in the difficult — but exciting! — position of finding a model class $\mathscr{C}$ that can model both types of update. Of course, $\mathbf{Nbhd}^\star_{\mathbf{S4}}$ will work, but it would be nice to find a tighter bound.

Okay, enough waxing philosophical. Let's get to work. A first observation is that if an inclusion holds in the dynamic case, it holds in the static case as well. This is because the dynamic language $\mathcal{L}^\star$ extends $\mathcal{L}$, and static formulas have the exact same semantics in the dynamic setting.

Lemma 7. For all classes of models $\mathscr{C}_1, \mathscr{C}_2$, if $\mathrm{Sat}(\mathscr{C}_1^\star) \subseteq \mathrm{Sat}(\mathscr{C}_2^\star)$ then $\mathrm{Sat}(\mathscr{C}_1) \subseteq \mathrm{Sat}(\mathscr{C}_2)$.

Proof. Suppose $\mathrm{Sat}(\mathscr{C}_1^\star) \subseteq \mathrm{Sat}(\mathscr{C}_2^\star)$, and let $\varphi \in \mathcal{L}$ be a static formula satisfied by some (static) model $\mathcal{M} \in \mathscr{C}_1$. **TODO** □

A major consequence is that we inherit inequalities from the static case. This means that, for the outer inclusions, we only need to show $\subseteq$ (we get $\neq$ for free).

Corollary 8. For all classes of models $\mathscr{C}_1, \mathscr{C}_2$, if $\mathrm{Sat}(\mathscr{C}_1) \neq \mathrm{Sat}(\mathscr{C}_2)$ then $\mathrm{Sat}(\mathscr{C}_1^\star) \neq \mathrm{Sat}(\mathscr{C}_2^\star)$.

Proof. Suppose $\mathrm{Sat}(\mathscr{C}_1) \neq \mathrm{Sat}(\mathscr{C}_2)$. Then either $\mathrm{Sat}(\mathscr{C}_1) \not\subseteq \mathrm{Sat}(\mathscr{C}_2)$, or $\mathrm{Sat}(\mathscr{C}_2) \not\subseteq \mathrm{Sat}(\mathscr{C}_1)$. In the first case, the previous lemma gives us $\mathrm{Sat}(\mathscr{C}_1^\star) \not\subseteq \mathrm{Sat}(\mathscr{C}_2^\star)$, and in the second case we have $\mathrm{Sat}(\mathscr{C}_2^\star) \not\subseteq \mathrm{Sat}(\mathscr{C}_1^\star)$. Regardless of which case we're in, we see that $\mathrm{Sat}(\mathscr{C}_1^\star) \neq \mathrm{Sat}(\mathscr{C}_2^\star)$. □

Now let's tackle the outer inclusions:

Proposition 9. $\mathrm{Sat}(\mathbf{Rel}^\star_{\mathbf{S4}}) \subset \mathrm{Sat}(\mathbf{Plaus}^\star)$

Proof. **TODO** □

Proposition 10. $\mathrm{Sat}(\mathbf{Plaus}^\star) \subset \mathrm{Sat}(\mathbf{Nbhd}^\star_{\mathbf{S4}})$

Proof. **TODO** □

Proposition 11. $\mathrm{Sat}(\mathbf{Rel}^\star_{\mathbf{S4}}) \subset \mathrm{Sat}(\mathbf{Net}^\star)$

Proof. **TODO** □

Proposition 12. $\mathrm{Sat}(\textbf{Net}^\star) \subset \mathrm{Sat}(\textbf{Nbhd}^\star_{\textbf{S4}})$

Proof. **TODO** □

As for the center relationship, for now I'll just pose the question.

Question 4. What is the relationship between $\mathrm{Sat}(\textbf{Plaus}^\star)$ and $\mathrm{Sat}(\textbf{Net}^\star)$?

My best guess is that they're incomparable. I have a proof strategy that *might* work — see **4-22-24-satisfying-hebbian-reduction-axioms.tex** for the idea and beginnings of this approach. But to avoid getting stuck on this proof, I think it's really important that I try to get a good intuition for which case we're in. This means I should think about how to simulate randomly generated update policies on a computer, generate billions of explicit examples, and check how probable these cases are.

Todo List

□ Prove that $\mathrm{Sat}(\textbf{Plaus}) \subseteq \mathrm{Sat}(\textbf{Net})$ by construction.

□ Prove that $\mathrm{Sat}(\textbf{Net}) \subseteq \mathrm{Sat}(\textbf{Plaus})$ by construction.

□ Prove Lemma 6.

□ Try to prove Propositions 7–10 (keeping in mind I may be wrong about any of them)

□ Write out my best attempt to prove that $\mathrm{Sat}(\textbf{Plaus}^\star)$ and $\mathrm{Sat}(\textbf{Net}^\star)$ are incomparable.

□ Write a program to generate possible model updates & check inclusions (this is going to be hard! But I expect that the proof itself will be harder.)

Text for slides:

- $\mathrm{Sat}(\textbf{Rel}_{\textbf{S4}}) \subset \mathrm{Sat}(\textbf{Net})$
- $\mathrm{Sat}(\textbf{SocialNet}) \subseteq \mathrm{Sat}(\textbf{Net})$
- $\mathrm{Sat}(\textbf{Net}) \subset \mathrm{Sat}(\textbf{Nbhd}_{\textbf{S4}})$
- $\mathrm{Sat}(\textbf{Net}) = \mathrm{Sat}(\textbf{Plaus})$
- $\mathrm{Sat}(\textbf{Rel}_{\textbf{S4}}) \subset \mathrm{Sat}(\textbf{Net})$
- $\mathrm{Sat}(\textbf{SocialNet}) \subseteq \mathrm{Sat}(\textbf{Net})$
- $\mathrm{Sat}(\textbf{Net}) \subset \mathrm{Sat}(\textbf{Nbhd}_{\textbf{S4}})$
- $\mathrm{Sat}(\textbf{Net}) = \mathrm{Sat}(\textbf{Plaus})$
- $\varphi \in \mathrm{Sat}(\textbf{Rel}_{\textbf{S4}})$
- $\in \mathrm{Sat}(\textbf{Net})$
- $\notin \mathrm{Sat}(\textbf{Rel}_{\textbf{S4}})$
- $\mathrm{Sat}(\textbf{SocialNet}) \subseteq \mathrm{Sat}(\textbf{Net})$
- $\mathsf{Cl}_i(S) = \{w \mid \exists \text{an } E\text{-path from } w \text{ to some } u \in S\}$
- uE_iw iff uR_iw

1. To show inclusion: Given $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbf{I}}, V\rangle \in \textbf{Rel}_{\textbf{S4}}$, construct

$$\mathcal{N} = \langle N, \mathsf{bias}, \{E_i\}_{i\in\mathbf{I}}, \{W_i\}_{i\in\mathbf{I}}, \{A_i\}_{i\in\mathbf{I}}, V\rangle$$

as follows:

- $N = W$ and keep V the same (bias doesn't matter)
- **Edges:** E_i is just R_i
- **Weights:**

$$\begin{tikzcd}
 & & \mathrm{Sat}(\textbf{Nbhd}_{\textbf{S4}}) \arrow[drr, phantom, sloped, "\supset"] & & \\
\mathrm{Sat}(\textbf{Plaus}) \arrow[urr, phantom, sloped, "\subset"] \arrow[drr, phantom, sloped, "\supset"] \arrow[rr, dash, dotted] & & = \arrow[rr, dash, dotted] & & \mathrm{Sat}(\textbf{Net}) \\
 & & \mathrm{Sat}(\textbf{Rel}_{\textbf{S4}}) \arrow[urr, phantom, sloped, "\subset"] & &
\end{tikzcd}$$

References

[1] Alexandru Baltag et al. "Dynamic epistemic logics of diffusion and prediction in social networks". In: *Studia Logica* 107 (2019), pp. 489–531.

[2] Caleb Schultz Kisby, Saúl A Blanco, and Lawrence S Moss. "What Do Hebbian Learners Learn? Reduction Axioms for Iterated Hebbian Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 13. 2024, pp. 14894–14901.

[3] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. "Nonmonotonic reasoning, preferential models and cumulative logics". In: *Artificial intelligence* 44.1-2 (1990), pp. 167–207.

[4] Hannes Leitgeb. "Nonmonotonic reasoning by inhibition nets". In: *Artificial Intelligence* 128.1-2 (2001), pp. 161–201.

[5] Hannes Leitgeb. "Nonmonotonic reasoning by inhibition nets II". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11.supp02 (2003), pp. 105–135.

[6] William Merrill. "Sequential neural networks as automata". In: *arXiv preprint arXiv:1906.01615* (2019).

[7] William Merrill and Ashish Sabharwal. "The expresssive power of transformers with chain of thought". In: *arXiv preprint arXiv:2310.07923* (2023).

[8] William Merrill et al. "A formal hierarchy of RNN architectures". In: *arXiv preprint arXiv:2004.08500* (2020).

[9] Eric Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017.

[10] Lena Strobl et al. "What Formal Languages Can Transformers Express? A Survey". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 543–561.

[11] Johan Van Benthem. *Logical dynamics of information and interaction*. Cambridge University Press, 2011.