The Modeling Power of Neural Networks

Caleb Schultz Kisby
Department of Computer Science
Indiana University Bloomington

LIRa Seminar
September 5, 2024

Talk Outline

- Introduction
- Modal Logic and Its Models
  - Relational Models
  - Plausibility Models
  - Neighborhood Models
- Neural Networks as Models
  - Social Network Models
- The Modeling Power of Neural Nets
  - $\mathrm{Sat}(\mathbf{Rel}_{S4}) \subset \mathrm{Sat}(\mathbf{Net})$
  - $\mathrm{Sat}(\mathbf{Net}) = \mathrm{Sat}(\mathbf{Plaus})$
  - $\mathrm{Sat}(\mathbf{Net}) \subset \mathrm{Sat}(\mathbf{Nbhd}_{S4})$
- Dynamic Update and Neural Nets **(ongoing work)**

Talk Outline

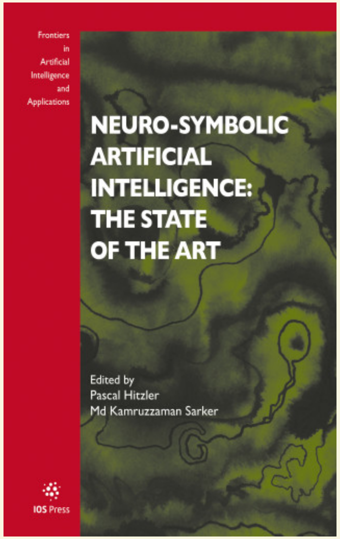
- Introduction
- Modal Logic and Its Models
  - Relational Models
  - Plausibility Models
  - Neighborhood Models
- Neural Networks as Models
  - Social Network Models
- The Modeling Power of Neural Nets
  - $\mathrm{Sat}(\mathbf{Rel}_{S4}) \subset \mathrm{Sat}(\mathbf{Net})$
  - $\mathrm{Sat}(\mathbf{Net}) = \mathrm{Sat}(\mathbf{Plaus})$
  - $\mathrm{Sat}(\mathbf{Net}) \subset \mathrm{Sat}(\mathbf{Nbhd}_{S4})$
- Dynamic Update and Neural Nets **(ongoing work)**

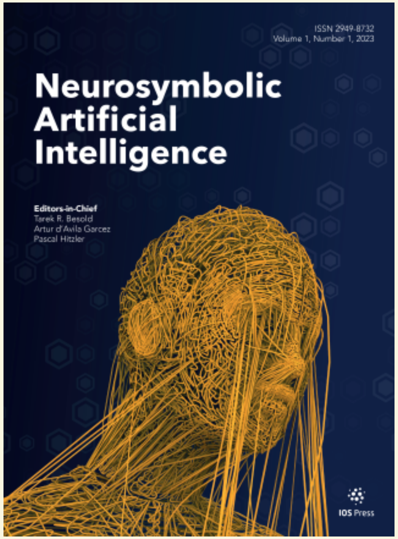
Talk Outline

- Introduction
- Modal Logic and Its Models
  - Relational Models
  - Plausibility Models
  - Neighborhood Models
- Neural Networks as Models
  - Social Network Models
- The Modeling Power of Neural Nets
  - $\mathrm{Sat}(\mathbf{Rel}_{S4}) \subset \mathrm{Sat}(\mathbf{Net})$
  - $\mathrm{Sat}(\mathbf{Net}) = \mathrm{Sat}(\mathbf{Plaus})$
  - $\mathrm{Sat}(\mathbf{Net}) \subset \mathrm{Sat}(\mathbf{Nbhd}_{S4})$
- Dynamic Update and Neural Nets **(ongoing work)**

The Neuro-Symbolic Question

How can we better understand and control neural networks using logic?





Neural Network Semantics

- In this talk, I will introduce you to one possible answer to this question:[1234]

We can treat neural networks as a class of models in modal logic simply by (1) adding a valuation of propositions, and (2) interpreting $\diamond\varphi$ as the spread of signal φ through the net.

- I will focus on understanding neural networks *in relation to* other models
  - What is the modeling power of neural networks?
  - Explicit constructions between nets and other models

.

1. Balkenius, C. and Gardenfors, P. Nonmonotonic inferences in neural networks. KR, 1991.
2. Leitgeb, H. Neural Network Models of Conditionals. In Introduction to Formal Philosophy, 2018.
3. Odense, S., and d'Avila Garcez, A. A Semantic Framework for Neural-Symbolic Computing. ArXiv, 2022.
4. Schultz Kisby, C., Blanco, S., and Moss, L. The Logic of Hebbian Learning. The International FLAIRS Conference Proceedings, 2022.

Why Modal Logic?

- All models I'll talk about are over the basic (multi)modal language $\mathcal{L}$:

$$\varphi \in p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box_i \varphi$$

- $i \in \boldsymbol{I}$, a fixed set of indices.
- Sometimes I will define semantics for $\Box_i \varphi$, other times I'll prefer $\Diamond_i \varphi \leftrightarrow \neg\Box_i \neg\varphi$. (We could have made $\Diamond_i \varphi$ primary, and defined $\Box_i \varphi \leftrightarrow \neg\Diamond_i \neg\varphi$, it doesn't matter.)

- But why interpret nets over **modal** logic? Three reasons:
  1. Modal logic is rich enough to express neural network *inference*.
     Intuition: We will interpret $\Diamond_i \varphi$ so that $\Box\varphi \rightarrow \psi$ reads "ψ is activated by input φ", alternatively "the net classifies φ as ψ"
  2. It's easier to lift modal logic to a dynamic setting (to study neural network *learning*)
  3. **An upsetting reason:** It's an open problem to find appropriate neural semantics for first-order logic!

Modal Logic and Its Models

- In first-order logic, we usually take for granted relational (graph) models
- In modal logic, there are many models for us to choose from!
  Here is our cast of characters for this talk:
  - Relational (Kripke) Models
  - Plausibility Models
  - Neighborhood Space Models
  - Social Network Models
  - **Neural Network Models**

Example: Mushroom Identification

You go out foraging for mushrooms. Morels are common in these woods, and as a general rule are safe to eat. But beware the false morels, which are an exception to this rule! False morels are mushrooms of the same family that can cause vomiting, diarrhea, dizziness, or even death! [5]

You happen upon a mushroom. You do not know what mushroom you are looking at, but you have certain possible beliefs about what it could be. How do you determine whether the mushroom is safe to eat?

Morchella conica
An edible morel

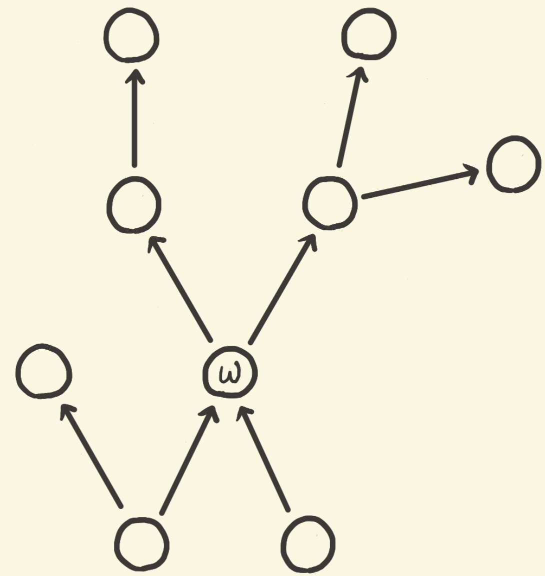
Gyromitra esculenta
A "false" morel

.
5. Leon Dufour. Atlas des Champignons Comestibles et Veneneux. 1891.

Relational Models ($\mathbf{Rel}_{S4}$)

- A relational model is $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle$, where
  - W is a finite set of **states**, $R_i \subseteq W \times W$, and
  - $V : \text{Proposition} \to \mathcal{P}(W)$ is the **valuation**
- I'll only consider reflexive and transitive R_i

  Let $\mathbf{Rel}_{S4}$ be the class of all such models.



Semantics:

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

Define: $\mathcal{M} \vDash \varphi$ iff $\mathcal{M}, w \Vdash \varphi$ for all $w \in W$

Define: $\vDash_{\mathbf{Rel}_{S4}} \varphi$ iff $\mathcal{M} \vDash \varphi$ for all $\mathcal{M} \in \mathbf{Rel}_{S4}$

Completeness: $\mathbf{Rel}_{S4}$ is complete with respect to $\vdash_{S4}$:

$$\text{If} \vdash \varphi \text{ then} \vdash \Box_i \varphi \qquad \Box_i(\varphi \to \psi) \to \Box_i \varphi \to \Box_i \psi \qquad \Box_i \varphi \to \varphi \qquad \Box_i \varphi \to \Box_i \Box_i \varphi$$

The (K) axiom does not allow for exceptions! $\mathbf{B}(\text{morel} \to \text{safe}) \to \mathbf{B}(\text{morel}) \to \mathbf{B}(\text{safe})$

Relational Models ($\mathbf{Rel}_{S4}$)

- A relational model is $\mathcal{M} = \langle W, \{R_i\}_{i\in \boldsymbol{I}}, V\rangle$, where
  - W is a finite set of **states**, $R_i \subseteq W \times W$, and
  - $V:\ \mathrm{Proposition} \to \mathcal{P}(W)$ is the **valuation**
- I'll only consider reflexive and transitive R_i

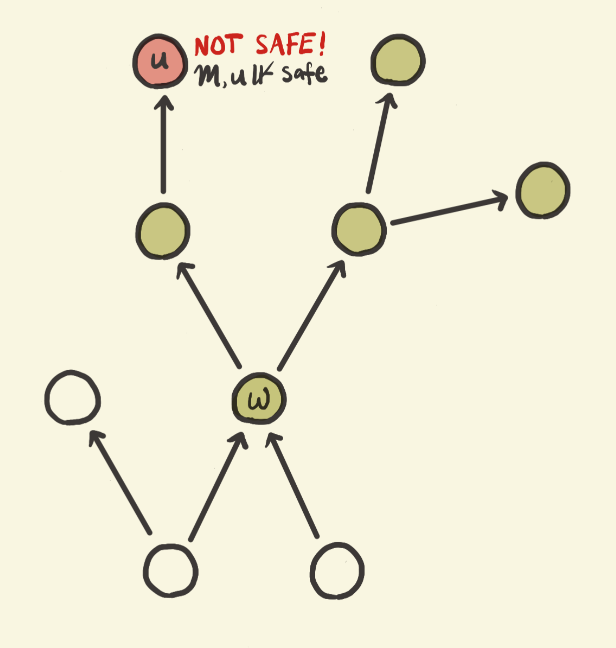
  Let $\mathbf{Rel}_{S4}$ be the class of all such models.

Semantics:

$\mathcal{M}, w \Vdash p$	iff	$w \in V(p)$
$\mathcal{M}, w \Vdash \neg\varphi$	iff	$\mathcal{M}, w \nVdash \varphi$
$\mathcal{M}, w \Vdash \varphi \wedge \psi$	iff	$\mathcal{M}, w \Vdash \varphi$ and $\mathcal{M}, w \Vdash \psi$
$\mathcal{M}, w \Vdash \Box_i \varphi$	iff	For all u with $wR_iu, \mathcal{M}, u \Vdash \varphi$

Define: $\mathcal{M} \vDash \varphi$ iff $\mathcal{M}, w \Vdash \varphi$ for all $w \in W$

Define: $\vDash_{\mathbf{Rel}_{S4}} \varphi$ iff $\mathcal{M} \vDash \varphi$ for all $\mathcal{M} \in \mathbf{Rel}_{S4}$



Completeness: $\mathbf{Rel}_{S4}$ is complete with respect to $\vdash_{S4}$:

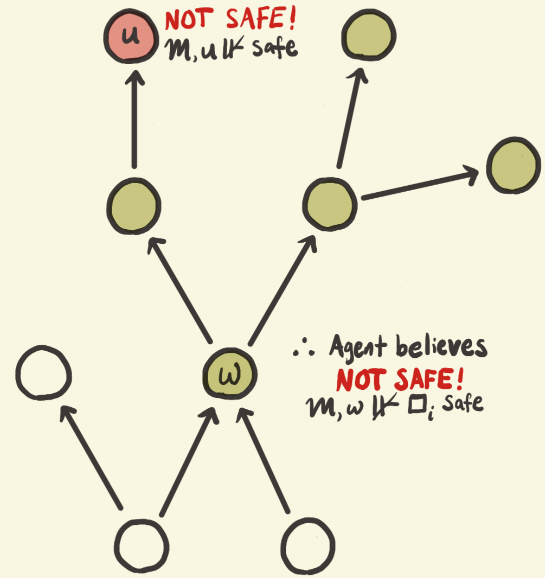
$$\text{If} \vdash \varphi \text{ then} \vdash \Box_i\varphi \qquad \Box_i(\varphi \to \psi) \to \Box_i\varphi \to \Box_i\psi \qquad \Box_i\varphi \to \varphi \qquad \Box_i\varphi \to \Box_i\Box_i\varphi$$

The (K) axiom does not allow for exceptions! $\mathbf{B}(\text{morel} \to \text{safe}) \to \mathbf{B}(\text{morel}) \to \mathbf{B}(\text{safe})$

Relational Models ($\mathbf{Rel}_{S4}$)

- A relational model is $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle$, where
  - W is a finite set of **states**, $R_i \subseteq W \times W$, and
  - $V : \text{Proposition} \to \mathcal{P}(W)$ is the **valuation**
- I'll only consider reflexive and transitive R_i

  Let $\mathbf{Rel}_{S4}$ be the class of all such models.



Semantics:

$\mathcal{M}, w \Vdash p$	iff	$w \in V(p)$
$\mathcal{M}, w \Vdash \neg\varphi$	iff	$\mathcal{M}, w \nVdash \varphi$
$\mathcal{M}, w \Vdash \varphi \wedge \psi$	iff	$\mathcal{M}, w \Vdash \varphi$ and $\mathcal{M}, w \Vdash \psi$
$\mathcal{M}, w \Vdash \Box_i \varphi$	iff	For all u with wR_iu, $\mathcal{M}, u \Vdash \varphi$

Define: $\mathcal{M} \vDash \varphi$ iff $\mathcal{M}, w \Vdash \varphi$ for all $w \in W$

Define: $\vDash_{\mathbf{Rel}_{S4}} \varphi$ iff $\mathcal{M} \vDash \varphi$ for all $\mathcal{M} \in \mathbf{Rel}_{S4}$

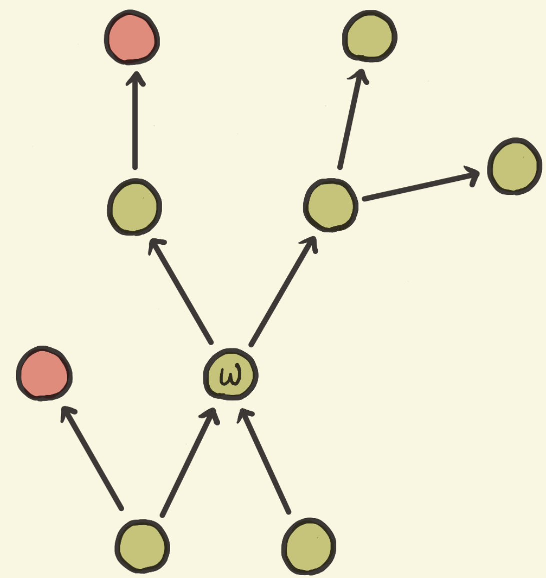
Completeness: $\mathbf{Rel}_{S4}$ is complete with respect to $\vdash_{S4}$:

$$\text{If } \vdash \varphi \text{ then } \vdash \Box_i \varphi \qquad \Box_i(\varphi \to \psi) \to \Box_i \varphi \to \Box_i \psi \qquad \Box_i \varphi \to \varphi \qquad \Box_i \varphi \to \Box_i \Box_i \varphi$$

The (K) axiom does not allow for exceptions! $\mathbf{B}(\text{morel} \to \text{safe}) \to \mathbf{B}(\text{morel}) \to \mathbf{B}(\text{safe})$

Plausibility Models (Plaus)

- A plausibility model is $\mathcal{M} = \langle W, \{R_i\}_{i\in \boldsymbol{I}}, V\rangle$
  - R_i is reflexive and transitive
- Let **Plaus** be the class of all such models
- Let $\text{best}_{R_i}(S) = \{w \mid w \text{ is } R_i\text{-minimal in } S\}$



Postulate on best$_i$: For all S, w, either $w \in \text{best}_{R_i}(S)$ or there is some better $v \in \text{best}_{R_i}(S)$.

Semantics: $\mathcal{M}, w \Vdash \Box_i \varphi$ iff $w \in \text{best}_{R_i}(\llbracket\varphi\rrbracket)$
where $\llbracket\varphi\rrbracket = \{u \mid \mathcal{M}, u \Vdash \varphi\}$

Completeness:[6] **Plaus** is complete with respect to cumulative logic $\vdash_{\mathbf{C}}$:

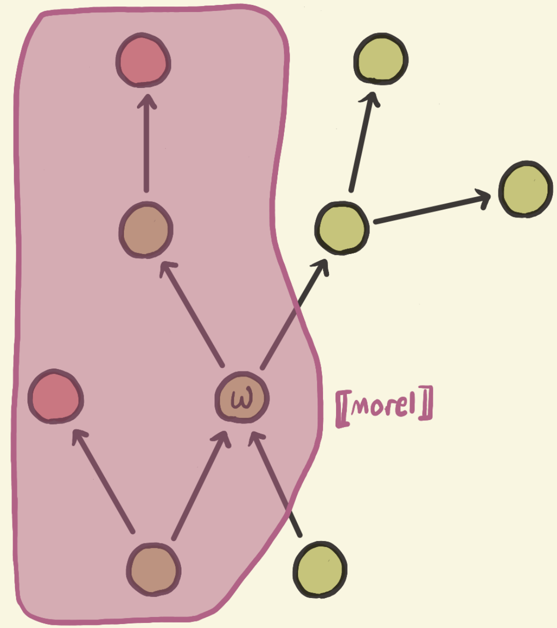
$$\text{If } \vdash \varphi \leftrightarrow \psi \text{ then} \vdash \Box_i\varphi \leftrightarrow \Box_i\psi \qquad (\Box_i\varphi \to \psi) \wedge (\psi \to \varphi) \to (\Box_i\varphi \leftrightarrow \Box_i\psi) \qquad \Box_i\varphi \to \varphi \qquad \Box_i\varphi \to \Box_i\Box_i\varphi$$

Exceptions can now be handled by conditional belief: $\mathbf{B}^{\varphi}\psi := \Box_i\varphi \to \psi$

6. Kraus, Lehmann, and Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. Artif. Intell. 1990

Plausibility Models (Plaus)

- A plausibility model is $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle$
  - R_i is reflexive and transitive
- Let $\mathbf{Plaus}$ be the class of all such models
- Let $\text{best}_{R_i}(S) = \{ w \mid w \text{ is } R_i\text{-minimal in } S \}$



Postulate on best$_i$: For all S, w, either $w \in \text{best}_{R_i}(S)$ or there is some better $v \in \text{best}_{R_i}(S)$.

Semantics: $\mathcal{M}, w \Vdash \Box_i \varphi$ iff $w \in \text{best}_{R_i}([\![\varphi]\!])$ where $[\![\varphi]\!] = \{ u \mid \mathcal{M}, u \Vdash \varphi \}$

Completeness:[7] $\mathbf{Plaus}$ is complete with respect to cumulative logic $\vdash_{\mathbf{C}}$:

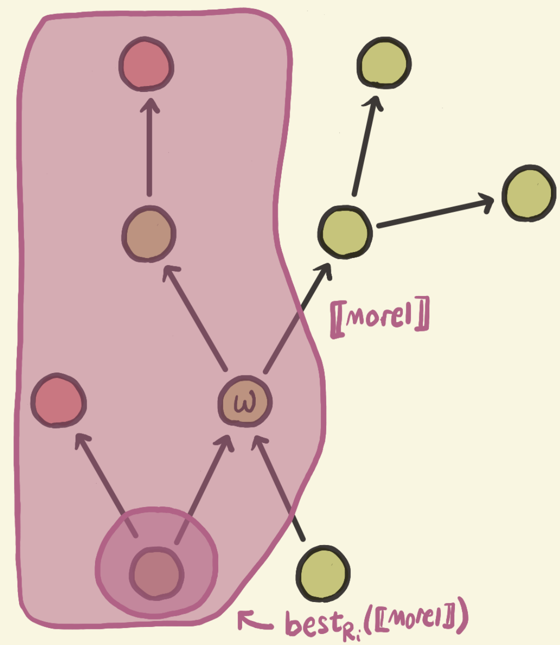
$$\text{If } \vdash \varphi \leftrightarrow \psi \text{ then } \vdash \Box_i \varphi \leftrightarrow \Box_i \psi \qquad (\Box_i \varphi \to \psi) \land (\psi \to \varphi) \to (\Box_i \varphi \leftrightarrow \Box_i \psi) \qquad \Box_i \varphi \to \varphi \qquad \Box_i \varphi \to \Box_i \Box_i \varphi$$

Exceptions can now be handled by conditional belief: $\mathbf{B}^{\varphi} \psi := \Box_i \varphi \to \psi$

7. Kraus, Lehmann, and Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. Artif. Intell. 1990

Plausibility Models (Plaus)

- A plausibility model is $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle$
  - R_i is reflexive and transitive
- Let $\mathbf{Plaus}$ be the class of all such models
- Let $\text{best}_{R_i}(S) = \{ w \mid w \text{ is } R_i\text{-minimal in } S \}$



Postulate on best$_i$: For all S, w, either $w \in \text{best}_{R_i}(S)$ or there is some better $v \in \text{best}_{R_i}(S)$.

Semantics: $\mathcal{M}, w \Vdash \Box_i \varphi$ iff $w \in \text{best}_{R_i}(\llbracket \varphi \rrbracket)$ where $\llbracket \varphi \rrbracket = \{ u \mid \mathcal{M}, u \Vdash \varphi \}$

Completeness:[8] $\mathbf{Plaus}$ is complete with respect to cumulative logic $\vdash_{\mathbf{C}}$:

$$\text{If } \vdash \varphi \leftrightarrow \psi \text{ then } \vdash \Box_i \varphi \leftrightarrow \Box_i \psi \qquad (\Box_i \varphi \to \psi) \wedge (\psi \to \varphi) \to (\Box_i \varphi \leftrightarrow \Box_i \psi) \qquad \Box_i \varphi \to \varphi \qquad \Box_i \varphi \to \Box_i \Box_i \varphi$$

Exceptions can now be handled by conditional belief: $\mathbf{B}^{\varphi} \psi := \Box_i \varphi \to \psi$

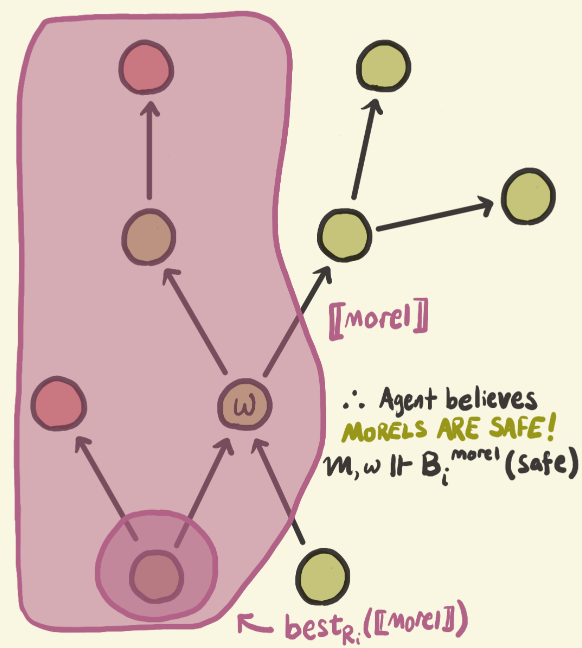
8. Kraus, Lehmann, and Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. Artif. Intell. 1990

Plausibility Models (Plaus)

- A plausibility model is $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle$
  - R_i is reflexive and transitive
- Let **Plaus** be the class of all such models
- Let $\text{best}_{R_i}(S) = \{w \mid w \text{ is } R_i\text{-minimal in } S\}$

Postulate on best$_i$: For all S, w, either $w \in \text{best}_{R_i}(S)$ or there is some better $v \in \text{best}_{R_i}(S)$.

Semantics: $\mathcal{M}, w \Vdash \Box_i \varphi$ iff $w \in \text{best}_{R_i}(\llbracket \varphi \rrbracket)$ where $\llbracket \varphi \rrbracket = \{u \mid \mathcal{M}, u \Vdash \varphi\}$



Completeness:[9] **Plaus** is complete with respect to cumulative logic $\vdash_{\mathbf{C}}$:

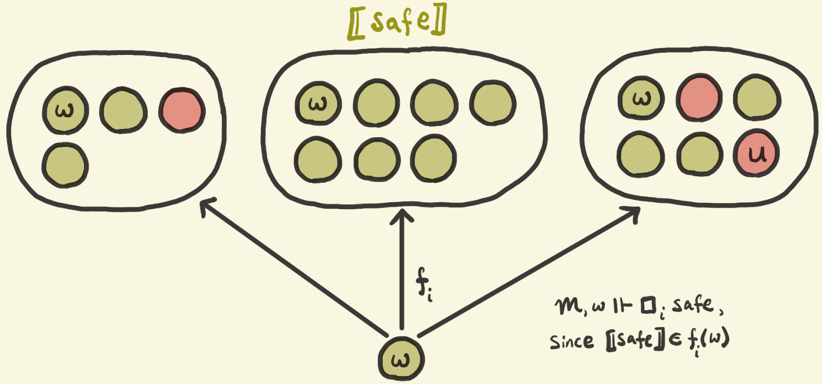
$$\text{If } \vdash \varphi \leftrightarrow \psi \text{ then } \vdash \Box_i \varphi \leftrightarrow \Box_i \psi \qquad (\Box_i \varphi \to \psi) \wedge (\psi \to \varphi) \to (\Box_i \varphi \leftrightarrow \Box_i \psi) \qquad \Box_i \varphi \to \varphi \qquad \Box_i \varphi \to \Box_i \Box_i \varphi$$

Exceptions can now be handled by conditional belief: $\mathbf{B}^{\varphi} \psi := \Box_i \varphi \to \psi$

9. Kraus, Lehmann, and Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. Artif. Intell. 1990

Neighborhood Models ($\mathbf{Nbhd}_{S4}$)

- A neighborhood model is $\mathcal{M} = \langle W, \{f_i\}_{i \in \boldsymbol{I}}, V \rangle$
  - $f_i : W \to \mathcal{P}(\mathcal{P}(W))$ (**neighborhood function**)
- f_i is transitive: $\forall w, X$, if $X \in f_i(w)$ then $\{u \mid X \in f_i(u)\} \in f_i(w)$
- f_i is reflexive: $\forall w, w \in \bigcap_{X \in f_i(w)} X$
- $\mathbf{Nbhd}$ is the class of all such models



Semantics:

$$\mathcal{M}, w \Vdash \Box_i \varphi \quad \text{iff} \quad [\![\varphi]\!] \in f_i(w)$$

where $[\![\varphi]\!] = \{u \mid \mathcal{M}, u \Vdash \varphi\}$

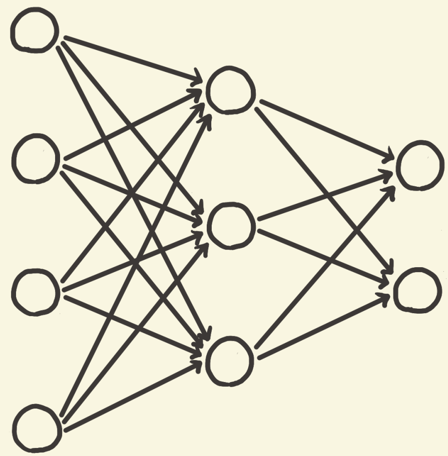
Completeness:[10] $\mathbf{Nbhd}_{S4}$ is complete with respect to:

$$\text{If } \vdash \varphi \leftrightarrow \psi \text{ then } \vdash \Box_i \varphi \leftrightarrow \Box_i \psi \qquad \Box_i \varphi \to \varphi \qquad \Box_i \varphi \to \Box_i \Box_i \varphi$$

Intuition: We can freely pick and choose which $[\![\varphi]\!]$ are accessible from a state w

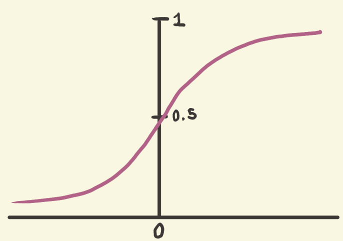
10. Pacuit, E. Neighborhood semantics for modal logic. Springer, 2017.

Brief Intro to Neural Networks

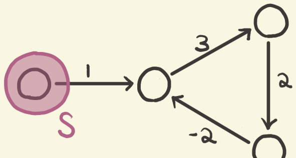


SAFE

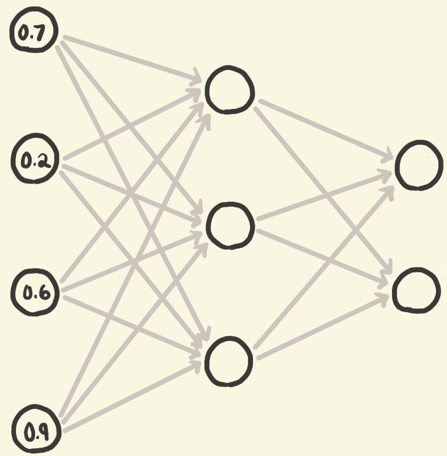
Binary (or "Saturated")



Terminating

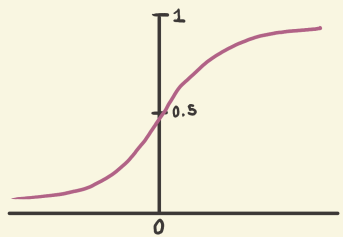


Brief Intro to Neural Networks

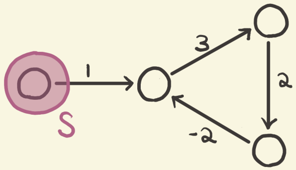


SAFE

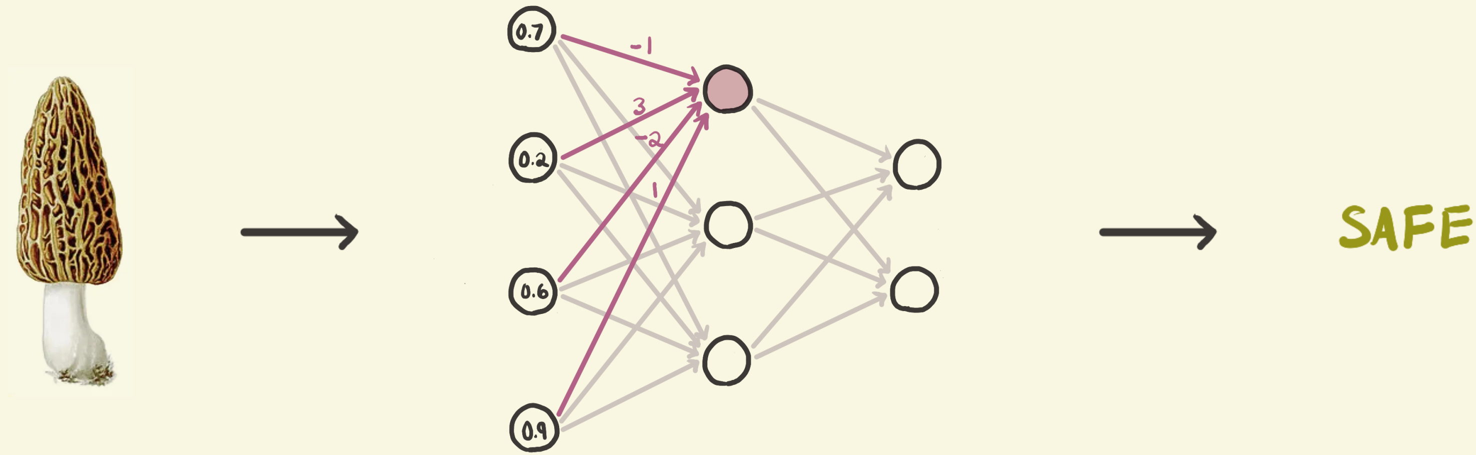
Binary (or “Saturated”)



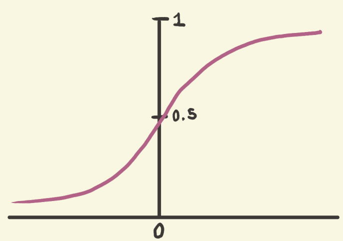
Terminating



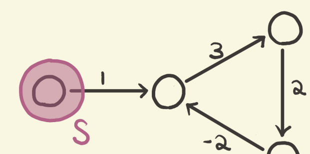
Brief Intro to Neural Networks



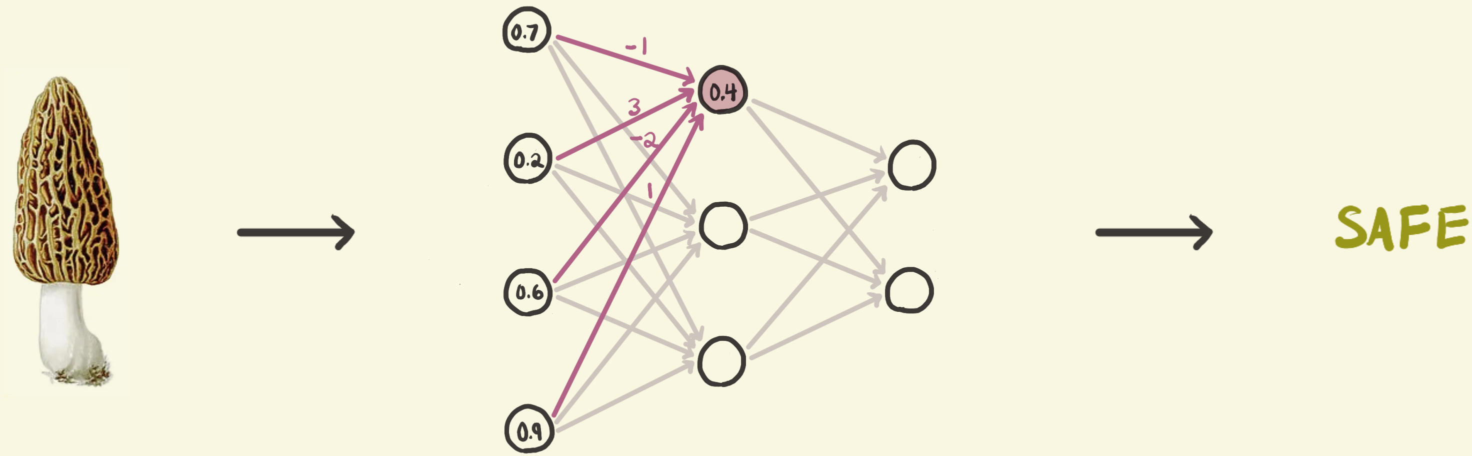
Binary (or "Saturated")



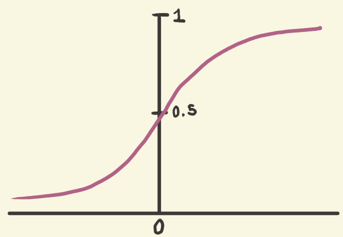
Terminating



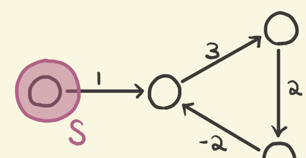
Brief Intro to Neural Networks



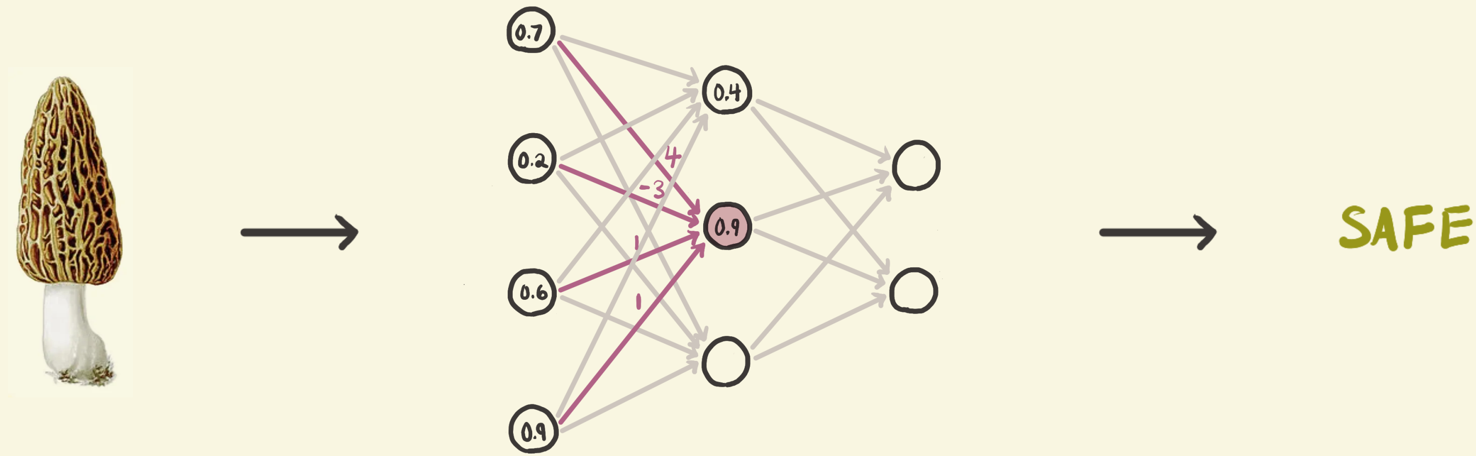
Binary (or "Saturated")



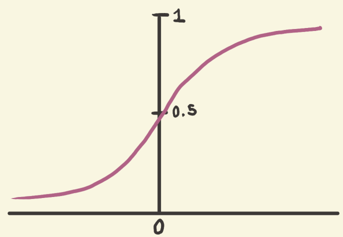
Terminating



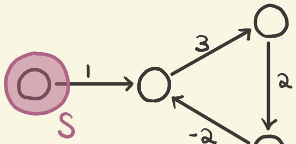
Brief Intro to Neural Networks



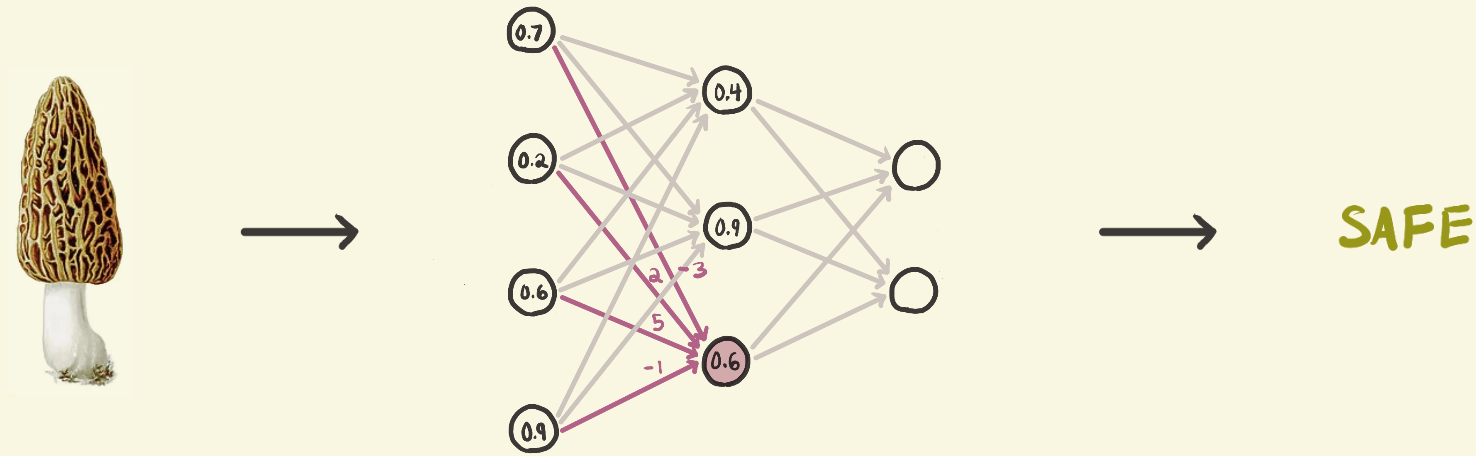
Binary (or “Saturated”)



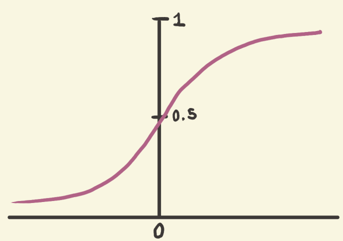
Terminating



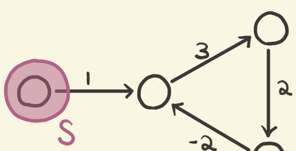
Brief Intro to Neural Networks



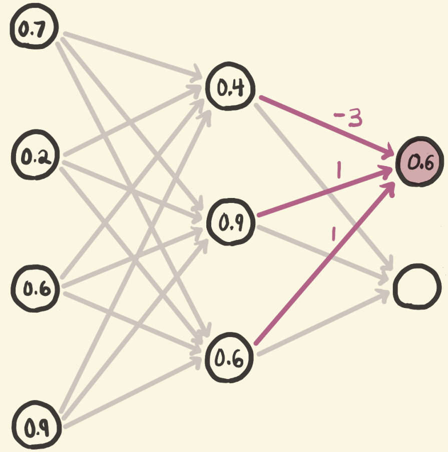
Binary (or “Saturated”)



Terminating

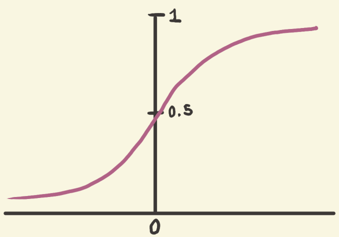


Brief Intro to Neural Networks

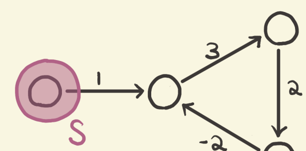


SAFE

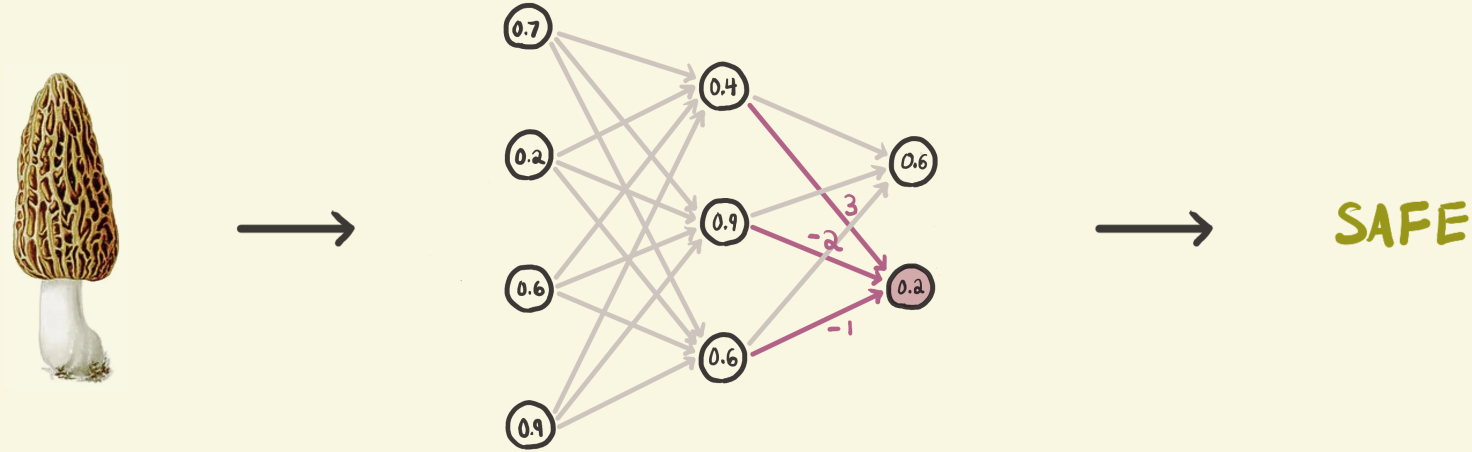
Binary (or "Saturated")



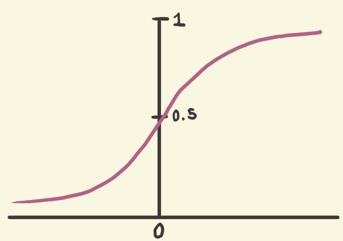
Terminating



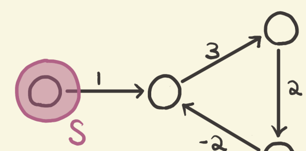
Brief Intro to Neural Networks



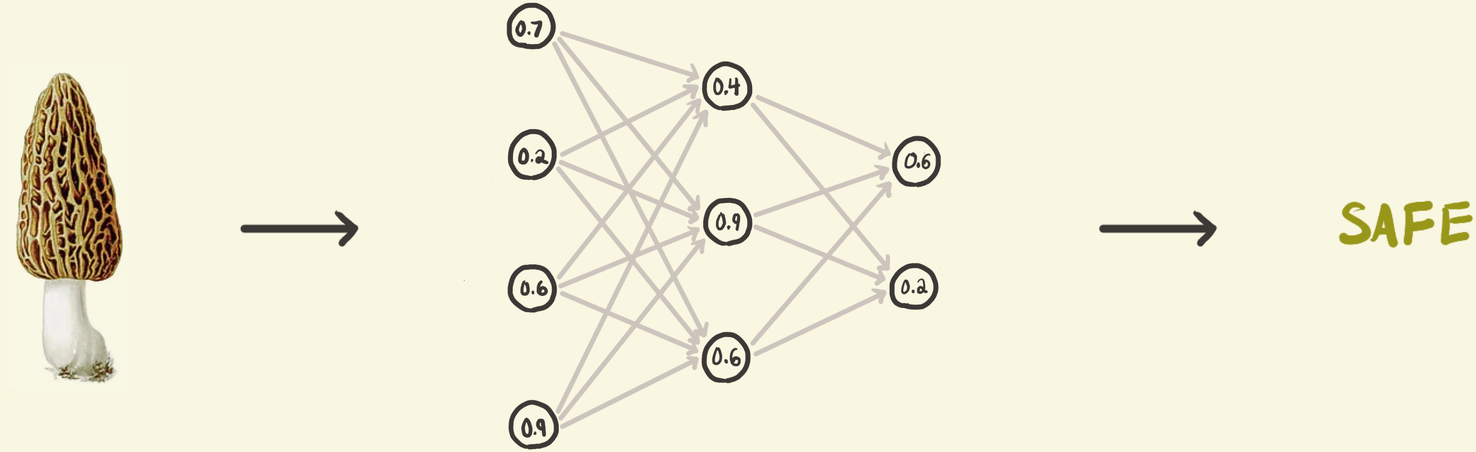
Binary (or "Saturated")



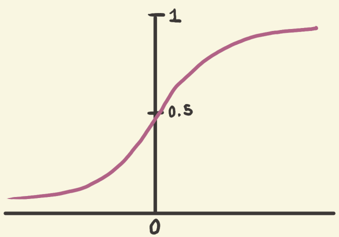
Terminating



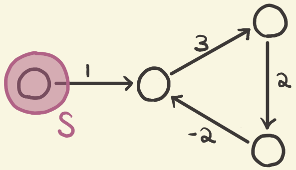
Brief Intro to Neural Networks



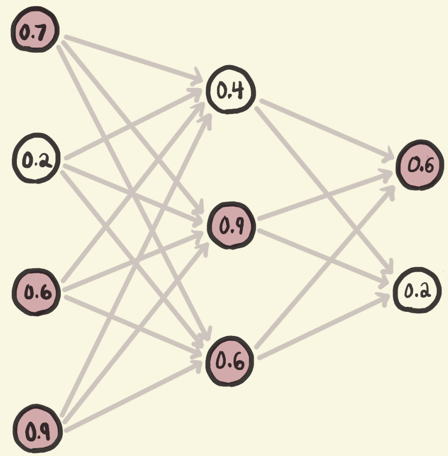
Binary (or “Saturated”)



Terminating

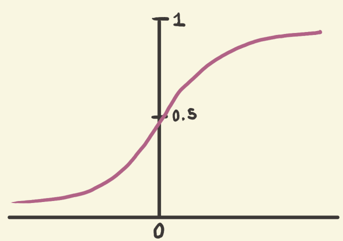


Brief Intro to Neural Networks

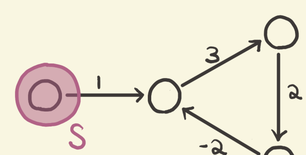


SAFE

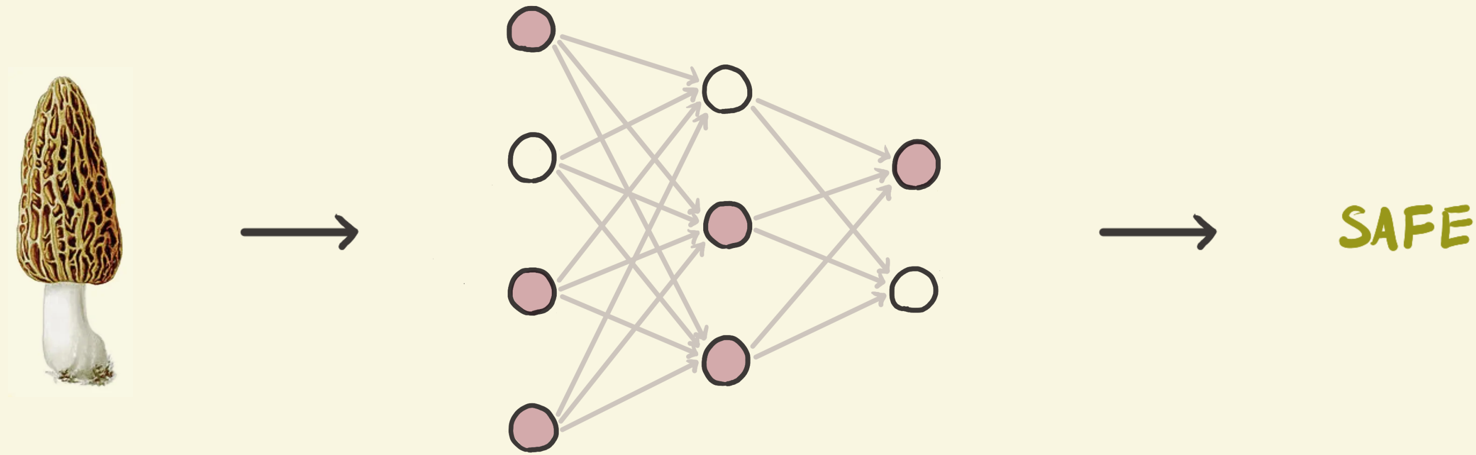
Binary (or “Saturated”)



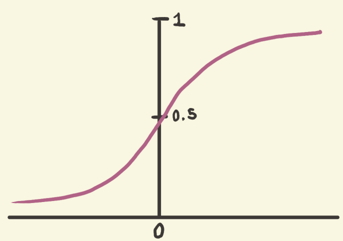
Terminating



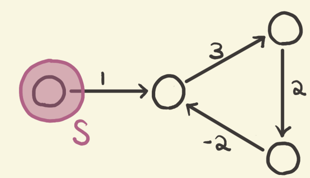
Brief Intro to Neural Networks



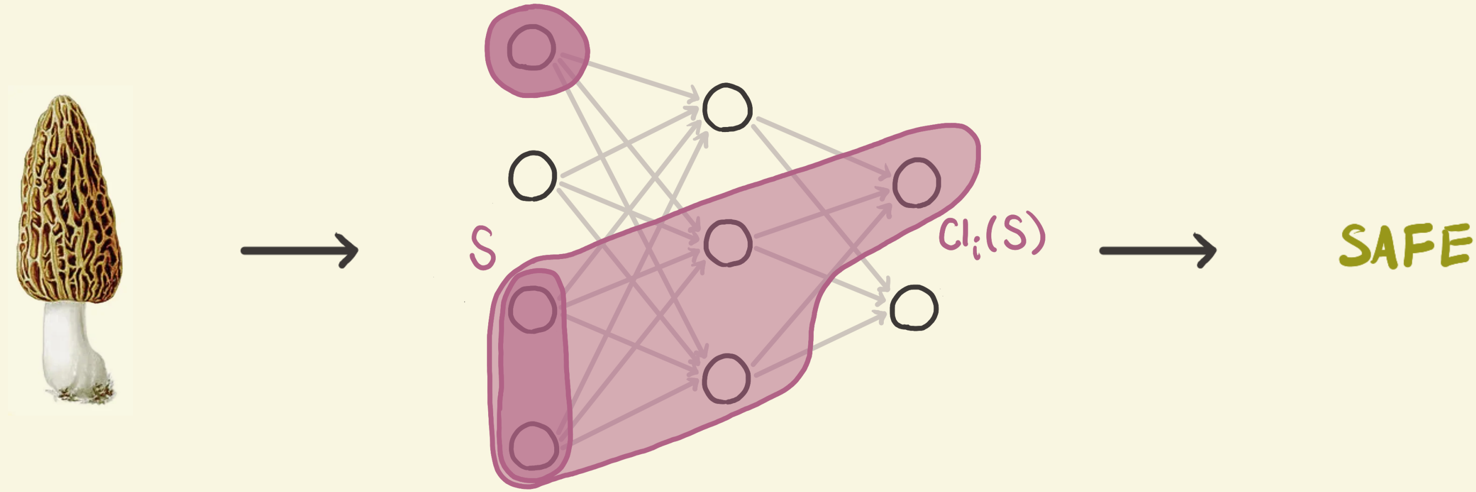
Binary (or “Saturated”)



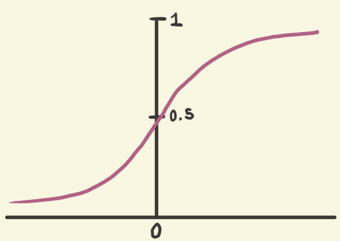
Terminating



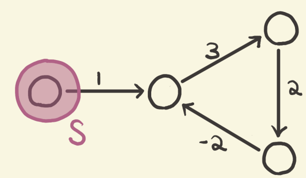
Brief Intro to Neural Networks



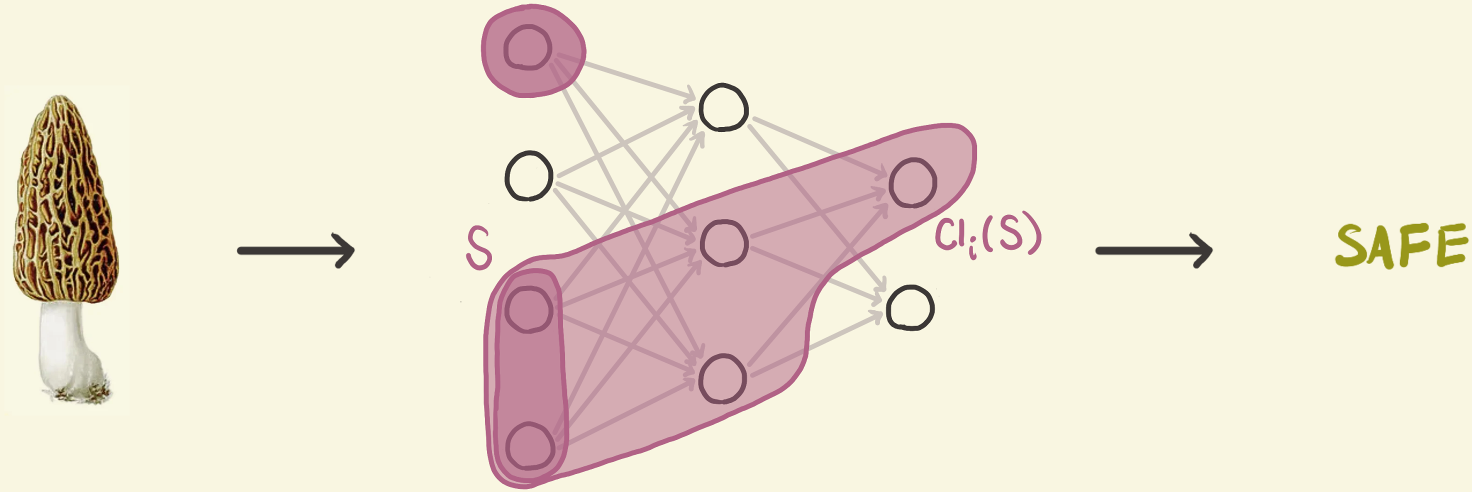
Binary (or “Saturated”)



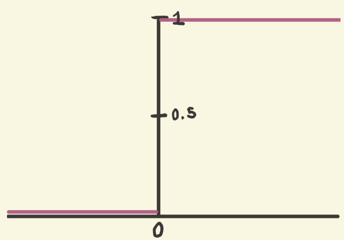
Terminating



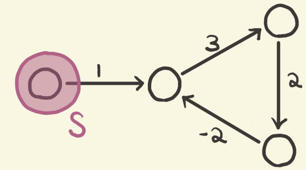
Brief Intro to Neural Networks



Binary (or "Saturated")



Terminating

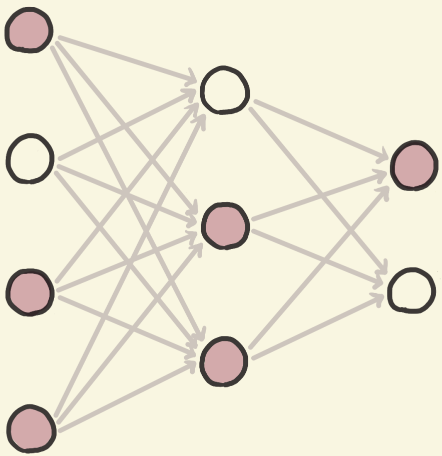


Neural Networks as Models

- A neural network model is $\mathcal{N} = \langle N, \text{bias}, \{E_i\}_{i \in \boldsymbol{I}}, \{W_i\}_{i \in \boldsymbol{I}}, \{A_i\}_{i \in \boldsymbol{I}}, V \rangle$
  - N is the set of nodes, bias a "bias" node, $V : \text{Proposition} \to \mathcal{P}(N)$ is a valuation.
  - E_i are edges, $W_i(u, v)$ are edge weights, $A_i : \mathbb{Q} \to \{0, 1\}$ are activation functions

- A **state** is a possible activation pattern of the net: active (1) or not (0)
  The bias node is always active!

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



- Each choice of E_i, W_i, A_i specifies a transition function from State $\to$ State:

$$F_i(S) = S \cup \left\{ w \mid A_i\left(\sum_{u \in \text{preds}(w)} W_i(u, w) \cdot \chi_S(u) \right) = 1 \right\}$$

Neural Networks as Models (Contd.)

- Moreover, we assume the net is **terminating**! i.e. for all $S \in \text{State}$, $i \in \boldsymbol{I}$, we assume

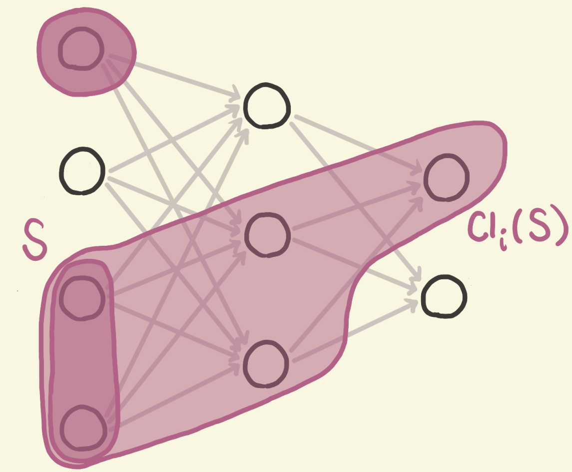
$$S, F_i(S), F_i(F_i(S)), \ldots, F_i^k(S), \ldots$$

  has a *finite, unique* fixed point. Let $\mathbf{Net}$ be the class of all such neural network models.

- Let the **closure** $\mathsf{Cl}_i(S)$: State $\rightarrow$ State be this fixed point.
  - $\mathsf{Cl}_i(S)$ is what I mean by the *spread* (or diffusion, or forward propagation) of the signal S through the net.

Semantics:

$\mathcal{N}, w \Vdash p$	iff	$w \in V(p)$
$\mathcal{N}, w \Vdash \neg\varphi$	iff	$\mathcal{N}, w \nVdash \varphi$
$\mathcal{N}, w \Vdash \varphi \wedge \psi$	iff	$\mathcal{N}, w \Vdash \varphi$ and $\mathcal{N}, w \Vdash \psi$
$\mathcal{N}, w \Vdash \diamond_i \varphi$	iff	$w \in \mathsf{Cl}_i(\llbracket\varphi\rrbracket)$

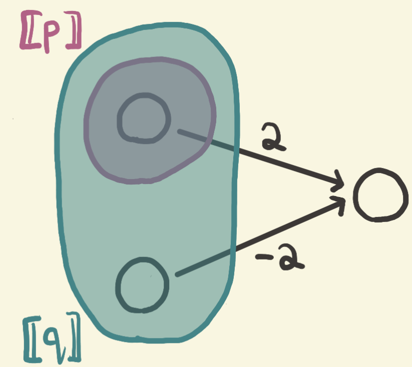


Neural Network Inference is $\Box_i \varphi \to \psi$

- I claim that modal logic is rich enough to express neural network inference
  - In particular, I claim that $\Box_i \varphi \to \psi$ expresses "the net classifies φ as ψ"

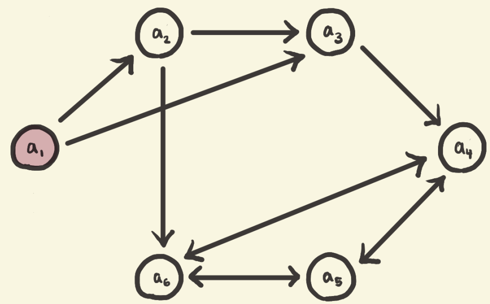
- Note that $\Box_i \varphi \to \psi$ is equivalent to $\psi \to \Diamond_i \varphi$
  - $\psi \to \Diamond_i \varphi$ says "the set of nodes activated by $[\![\varphi]\!]$ includes all nodes in state $[\![\psi]\!]$
  - Or: The state $[\![\psi]\!]$ is activated by input $[\![\varphi]\!]$
  - Or: The net *classifies* φ as ψ
  - Example: $\Box_i \mathrm{morel} \to \mathrm{safe}$

- Recall that for **Plaus**, we used $\mathbf{B}^{\varphi}\psi := \Box_i \varphi \to \psi$ to express conditional belief.
  - Like **Plaus**, neural networks can model nonmonotonicity!



Aside: Social Network Models

- Some of you have taken a similar logical approach to social network models[11]. How do social networks and neural networks compare?
- Consider a basic threshold model, where agents adopt an opinion if $\theta \geqslant \frac{1}{2}$ of their close friends do (friendships are directed here):

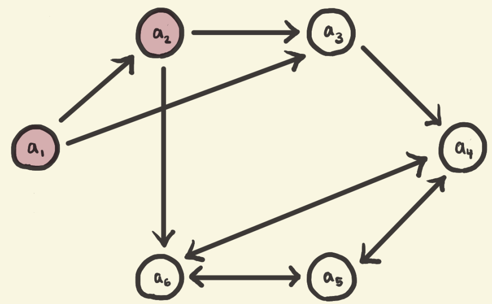


- **Both** are models of spread/diffusion of a signal, and the formalization is similar!
- Perspective: Agents vs Single Agent
- $\mathsf{Cl}_i(S)$ is a particular adoption policy

11. Baltag, A., Christoff, Z., Rendsvig, R., Smets, S. Dynamic epistemic logics of diffusion and prediction in social networks. Studia Logica, 2019.

Aside: Social Network Models

- Some of you have taken a similar logical approach to social network models[12]. How do social networks and neural networks compare?
- Consider a basic threshold model, where agents adopt an opinion if $\theta \geqslant \frac{1}{2}$ of their close friends do (friendships are directed here):

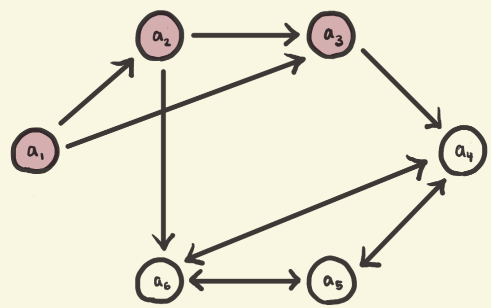


- **Both** are models of spread/diffusion of a signal, and the formalization is similar!
- Perspective: Agents vs Single Agent
- $\mathsf{Cl}_i(S)$ is a particular adoption policy

12. Baltag, A., Christoff, Z., Rendsvig, R., Smets, S. Dynamic epistemic logics of diffusion and prediction in social networks. Studia Logica, 2019.

Aside: Social Network Models

- Some of you have taken a similar logical approach to social network models[13]. How do social networks and neural networks compare?
- Consider a basic threshold model, where agents adopt an opinion if $\theta \geqslant \frac{1}{2}$ of their close friends do (friendships are directed here):



- **Both** are models of spread/diffusion of a signal, and the formalization is similar!
- Perspective: Agents vs Single Agent
- $\mathrm{Cl}_i(S)$ is a particular adoption policy

13. Baltag, A., Christoff, Z., Rendsvig, R., Smets, S. Dynamic epistemic logics of diffusion and prediction in social networks. Studia Logica, 2019.

Measuring Modeling Power

- How can we measure the relative power of two classes of models? We have many choices, each answering a different question about modeling power:
  - **Satisfiability.** What formulas *can* a class of models $\mathcal{C}$ satisfy?

$$\mathrm{Sat}(\mathcal{C}) = \{\varphi \in \mathcal{L} \mid \exists\, \mathcal{M} \in \mathcal{C} \text{ with } \mathcal{M} \vDash \varphi\}$$

  - **Validity.** What formulas *must* a class of models satisfy?

$$\mathrm{Th}(\mathcal{C}) = \{\varphi \in \mathcal{L} \mid \forall\, \mathcal{M} \in \mathcal{C} \text{ we have } \mathcal{M} \vDash \varphi\}$$

  - **Definability.** What properties can a class of models + language $\mathcal{L}$ *define*?

$$\mathrm{Def}(\mathcal{C}) = \{P \subseteq \mathcal{C} \mid \exists\, \varphi \in \mathcal{L} \text{ such that } \forall \mathcal{M} \in \mathcal{C},\ \mathcal{M} \in P \text{ iff } \mathcal{M} \vDash \varphi\}$$

- In this talk, I will use **satisfiability** to compare nets with other models
- Unfortuantely, I will not talk about **definability**. (For descriptive complexity of neural networks, see FLaNN group (flann.super.site) and this paper[14]

14. Martin, G. The descriptive complexity of graph neural networks. LICS, 2023.

Satisfiability Hierarchy

$$\begin{tikzcd}[column sep=small]
 & \mathrm{Sat}(\mathbf{Nbhd_{S4}}) & \\
\mathrm{Sat}(\mathbf{Plaus}) \arrow[ur, phantom, sloped, "\subset"] \arrow[rr, dotted, no head, "="] & & \mathrm{Sat}(\mathbf{Net}) \arrow[ul, phantom, sloped, "\subset"] \\
 & \mathrm{Sat}(\mathbf{Rel_{S4}}) \arrow[ul, phantom, sloped, "\subset"] \arrow[ur, phantom, sloped, "\subset"] &
\end{tikzcd}$$

Open Question. What is the relationship between $\mathrm{Sat}(\mathbf{Net})$ and satisfiability in social networks? What social net adoption policies can $\mathbf{Net}$ model?

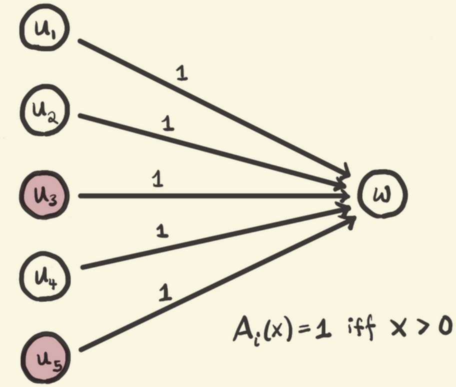
$\mathrm{Sat}(\mathbf{Rel}_{S4}) \subset \mathrm{Sat}(\mathbf{Net})$

Proof:

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

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

  - $N = W$, bias arbitrary, keep V the same, and edges $E_i = R_i$
  - Weights $W_i(u, w) = 1$ for all uE_iw
  - Activation: $A_i(x) = 1$ iff $x > 0$



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

- **Lemma.** For all $\varphi, w, \mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
  - By induction, key case is $\mathcal{M}, w \Vdash \Diamond_i \varphi$ iff $\mathcal{N}, w \Vdash \Diamond_i \varphi$

- So any $\varphi \in \mathrm{Sat}(\mathbf{Rel}_{S4})$ is also $\varphi \in \mathrm{Sat}(\mathbf{Net})$
- This inclusion is strict, since nonmonotonicity $\neg(\Box_i(\varphi \wedge \psi) \rightarrow (\Box_i \varphi \wedge \Box_i \psi))$ is in $\mathrm{Sat}(\mathbf{Net})$ but not in $\mathrm{Sat}(\mathbf{Rel}_{S4})$

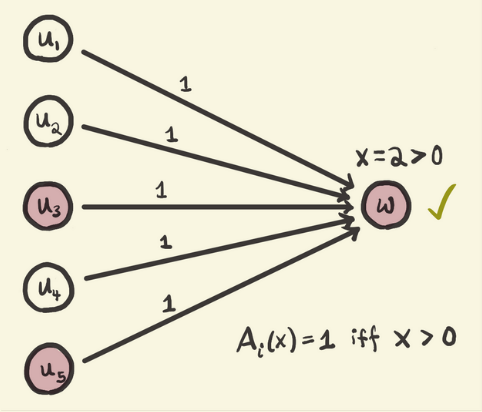
$\mathbf{Sat(Rel_{S4}) \subset Sat(Net)}$

Proof:

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

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

  - $N = W$, bias arbitrary, keep V the same, and edges $E_i = R_i$
  - Weights $W_i(u, w) = 1$ for all uE_iw
  - Activation: $A_i(x) = 1$ iff $x > 0$



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

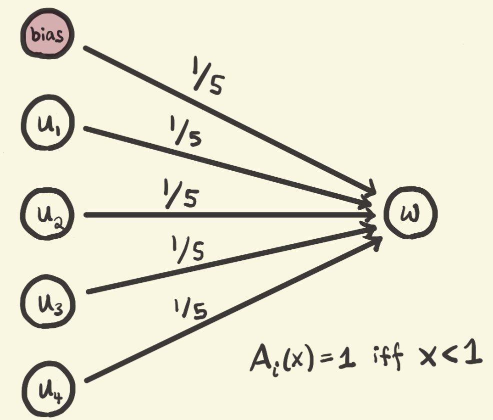
- **Lemma.** For all $\varphi, w, \mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
  - By induction, key case is $\mathcal{M}, w \Vdash \Diamond_i \varphi$ iff $\mathcal{N}, w \Vdash \Diamond_i \varphi$

- So any $\varphi \in \text{Sat}(\mathbf{Rel}_{S4})$ is also $\varphi \in \text{Sat}(\mathbf{Net})$
- This inclusion is strict, since nonmonotonicity $\neg(\Box_i(\varphi \wedge \psi) \rightarrow (\Box_i \varphi \wedge \Box_i \psi))$ is in $\text{Sat}(\mathbf{Net})$ but not in $\text{Sat}(\mathbf{Rel}_{S4})$

$\mathbf{Sat}(\mathbf{Plaus}) \subseteq \mathbf{Sat}(\mathbf{Net})$

Proof:[15]

- To show inclusion: Given $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle \in \mathbf{Plaus}$, construct
$$\mathcal{N} = \langle N, \text{bias}, \{E_i\}_{i \in \boldsymbol{I}}, \{W_i\}_{i \in \boldsymbol{I}}, \{A_i\}_{i \in \boldsymbol{I}}, V \rangle$$
as follows. If $\diamond_i \varphi$ has relational semantics, do the $\mathbf{Rel}_{S4}$ construction. If $\diamond_i \varphi$ has plausibility semantics:
  - $N = W$, keep V the same, and let bias be a fresh node.
  - $E_i = R_i$, but we also create an edge from bias to every node that is not E_i-minimal.
  - Weights $W_i(u, w) = \dfrac{1}{|\text{preds}(w)| + 1}$ for all $u E_i w$
  - Activation: $A_i(x) = 1$ iff $x < 1$



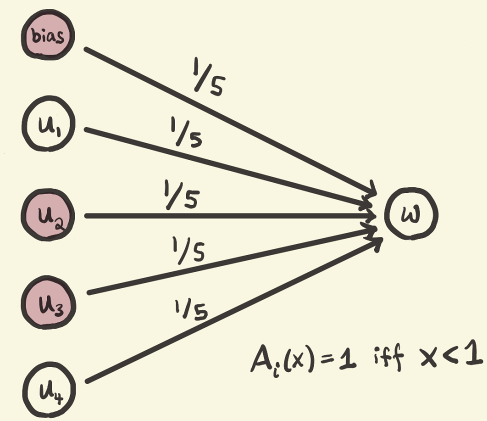
- **Lemma.** $\text{Cl}_i(S) = \overline{\text{best}_i(\overline{S})}$
- **Lemma.** For all $\varphi, w, \mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
- So any $\varphi \in \text{Sat}(\mathbf{Plaus})$ is also $\varphi \in \text{Sat}(\mathbf{Net})$

15. Leitgeb, H. Nonmonotonic reasoning by inhibition nets. In Artificial Intelligence, 2001.

$\mathbf{Sat(Plaus)} \subseteq \mathbf{Sat(Net)}$

Proof:[16]

- To show inclusion: Given $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V\rangle \in \mathbf{Plaus}$, construct
$$\mathcal{N} = \langle N, \text{bias}, \{E_i\}_{i \in \boldsymbol{I}}, \{W_i\}_{i \in \boldsymbol{I}}, \{A_i\}_{i \in \boldsymbol{I}}, V\rangle$$
as follows. If $\diamond_i \varphi$ has relational semantics, do the $\mathbf{Rel}_{S4}$ construction.
If $\diamond_i \varphi$ has plausibility semantics:
    - $N = W$, keep V the same, and let bias be a fresh node.
    - $E_i = R_i$, but we also create an edge from bias to every node that is not E_i-minimal.
    - Weights $W_i(u, w) = \dfrac{1}{|\text{preds}(w)| + 1}$ for all uE_iw
    - Activation: $A_i(x) = 1$ iff $x < 1$



- **Lemma.** $\text{Cl}_i(S) = \overline{\text{best}_i(\overline{S})}$
- **Lemma.** For all $\varphi, w, \mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
- So any $\varphi \in \text{Sat}(\mathbf{Plaus})$ is also $\varphi \in \text{Sat}(\mathbf{Net})$

16. Leitgeb, H. Nonmonotonic reasoning by inhibition nets. In Artificial Intelligence, 2001.

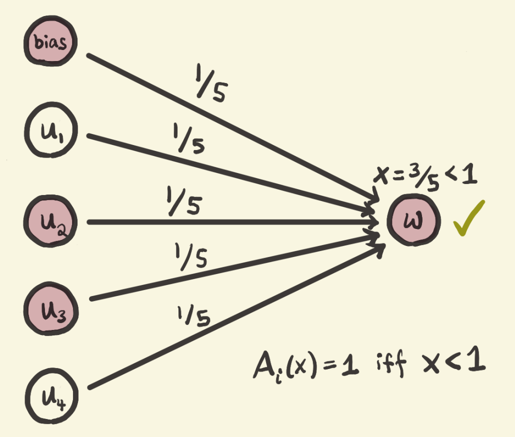
$\mathrm{Sat}(\mathbf{Plaus}) \subseteq \mathrm{Sat}(\mathbf{Net})$

Proof:[17]

- To show inclusion: Given $\mathcal{M} = \langle W, \{R_i\}_{i \in \boldsymbol{I}}, V \rangle \in \mathbf{Plaus}$, construct

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

  as follows. If $\diamond_i \varphi$ has relational semantics, do the $\mathbf{Rel}_{S4}$ construction.
  If $\diamond_i \varphi$ has plausibility semantics:
  - $N = W$, keep V the same, and let bias be a fresh node.
  - $E_i = R_i$, but we also create an edge from bias to every node that is not E_i-minimal.
  - Weights $W_i(u, w) = \dfrac{1}{|\text{preds}(w)| + 1}$ for all $u E_i w$
  - Activation: $A_i(x) = 1$ iff $x < 1$



- **Lemma.** $\mathrm{Cl}_i(S) = \overline{\mathrm{best}_i(\overline{S})}$
- **Lemma.** For all φ, w, $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
- So any $\varphi \in \mathrm{Sat}(\mathbf{Plaus})$ is also $\varphi \in \mathrm{Sat}(\mathbf{Net})$

17. Leitgeb, H. Nonmonotonic reasoning by inhibition nets. In Artificial Intelligence, 2001.

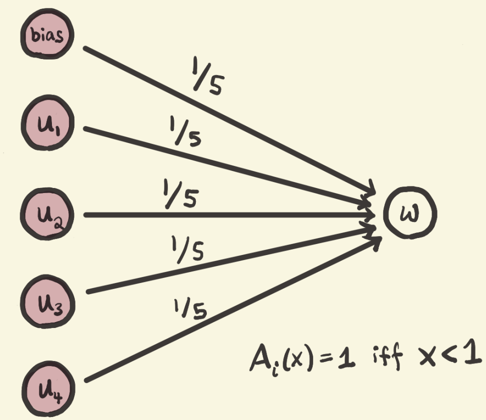
Sat(Plaus) ⊆ Sat(Net)

Proof:[18]

- To show inclusion: Given $\mathcal{M} = \langle W, \{R_i\}_{i\in \boldsymbol{I}}, V\rangle \in \mathbf{Plaus}$, construct

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

  as follows. If $\diamond_i \varphi$ has relational semantics, do the $\mathbf{Rel}_{S4}$ construction.
  If $\diamond_i \varphi$ has plausibility semantics:
  - $N = W$, keep V the same, and let bias be a fresh node.
  - $E_i = R_i$, but we also create an edge from bias to every node that is not E_i-minimal.
  - Weights $W_i(u, w) = \dfrac{1}{|\text{preds}(w)| + 1}$ for all uE_iw
  - Activation: $A_i(x) = 1$ iff $x < 1$



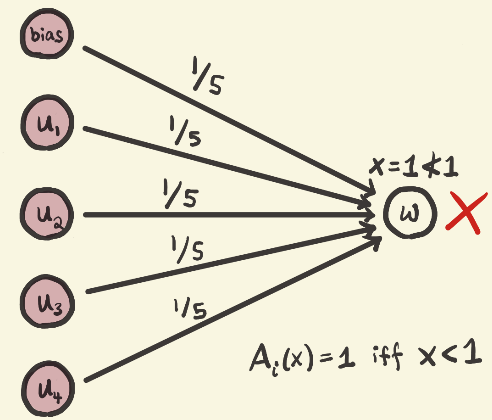
- **Lemma.** $\text{Cl}_i(S) = \overline{\text{best}_i(\overline{S})}$
- **Lemma.** For all φ, w, $\mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
- So any $\varphi \in \text{Sat}(\mathbf{Plaus})$ is also $\varphi \in \text{Sat}(\mathbf{Net})$

18. Leitgeb, H. Nonmonotonic reasoning by inhibition nets. In Artificial Intelligence, 2001.

Sat(Plaus) ⊆ Sat(Net)

Proof:[19]

- To show inclusion: Given $\mathcal{M} = \langle W, \{R_i\}_{i\in\boldsymbol{I}}, V\rangle \in \mathbf{Plaus}$, construct
$$\mathcal{N} = \langle N, \text{bias}, \{E_i\}_{i\in\boldsymbol{I}}, \{W_i\}_{i\in\boldsymbol{I}}, \{A_i\}_{i\in\boldsymbol{I}}, V\rangle$$
as follows. If $\diamond_i\varphi$ has relational semantics, do the $\mathbf{Rel}_{S4}$ construction. If $\diamond_i\varphi$ has plausibility semantics:
    - $N = W$, keep V the same, and let bias be a fresh node.
    - $E_i = R_i$, but we also create an edge from bias to every node that is not E_i-minimal.
    - Weights $W_i(u, w) = \dfrac{1}{|\text{preds}(w)| + 1}$ for all uE_iw
    - Activation: $A_i(x) = 1$ iff $x < 1$



- **Lemma.** $\text{Cl}_i(S) = \overline{\text{best}_i(\overline{S})}$
- **Lemma.** For all $\varphi, w, \mathcal{M}, w \Vdash \varphi$ iff $\mathcal{N}, w \Vdash \varphi$
- So any $\varphi \in \text{Sat}(\mathbf{Plaus})$ is also $\varphi \in \text{Sat}(\mathbf{Net})$

19. Leitgeb, H. Nonmonotonic reasoning by inhibition nets. In Artificial Intelligence, 2001.

$\mathrm{Sat}(\mathbf{Net}) \subseteq \mathrm{Sat}(\mathbf{Plaus})$

Proof: (Not constructive!)

$$
\begin{array}{rlll}
\varphi \in \mathrm{Sat}(\mathbf{Net}) & \longrightarrow & \exists\, \mathcal{N} \in \mathbf{Net} \text{ such that } \mathcal{N} \vDash \varphi & \\
& \longrightarrow & \exists\, \mathcal{N} \in \mathbf{Net} \text{ such that } \mathcal{N} \nvDash \neg\varphi & \\
& \longrightarrow & \nvDash_{\mathbf{Net}} \neg\varphi & \text{(defn of } \vDash_{\mathbf{Net}}) \\
& \longrightarrow & \nvdash_{\mathbf{C}} \neg\varphi & \text{(since } \vDash_{\mathbf{Net}} \text{ is sound wrt } \mathbf{C}) \\
& \longrightarrow & \nvDash_{\mathbf{Plaus}} \neg\varphi & \text{(since } \vDash_{\mathbf{Plaus}} \text{ is complete wrt } \mathbf{C}) \\
& \longrightarrow & \exists\, \mathcal{M} \in \mathbf{Plaus} \text{ such that } \mathcal{M} \nvDash \neg\varphi & \text{(defn of } \vDash_{\mathbf{Plaus}}) \\
& \longrightarrow & \exists\, \mathcal{M} \in \mathbf{Plaus} \text{ such that } \mathcal{M} \vDash \varphi &
\end{array}
$$

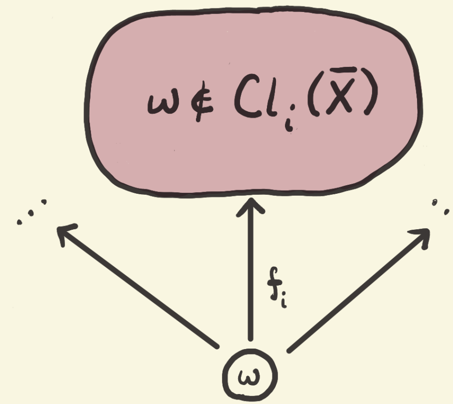
(This satisfies φ, but it doesn't satisfy me…)

Open Question. Is there a constructive proof of $\mathrm{Sat}(\mathbf{Net}) \subseteq \mathrm{Sat}(\mathbf{Plaus})$?

$\mathbf{Sat}(\mathbf{Net}) \subset \mathbf{Sat}(\mathbf{Nbhd})$

Proof:

- To show inclusion: Given $\mathcal{N} = \langle N, \text{bias}, \{E_i\}_{i\in\boldsymbol{I}}, \{W_i\}_{i\in\boldsymbol{I}}, \{A_i\}_{i\in\boldsymbol{I}}, V\rangle \in \mathbf{Net}$, construct
$$\mathcal{M} = \langle W, \{f_i\}_{i\in\boldsymbol{I}}, V\rangle \in \mathbf{Nbhd}_{S4}$$
  - $W = N$, keep V the same
  - $f_i(w) = \{X \subseteq W \mid w \notin \mathrm{Cl}_i(\overline{X})\}$



- **Lemma.** f_i is a refl and trans neighborhood function
  - Because Cl_i is inclusive and idempotent!
- **Lemma.** For all $\varphi, w, \mathcal{N}, w \Vdash \varphi$ iff $\mathcal{M}, w \Vdash \varphi$
  - By induction, key case is $\mathcal{N}, w \Vdash \Box_i\varphi$ iff $\mathcal{M}, w \Vdash \Box_i\varphi$
- So any $\varphi \in \mathrm{Sat}(\mathbf{Net})$ is also $\varphi \in \mathrm{Sat}(\mathbf{Nbhd}_{S4})$
- This inclusion is strict, since non-cumulativity $\neg((\Box_i\varphi \to \psi) \wedge (\psi \to \varphi) \to (\Box_i\varphi \leftrightarrow \Box_i\psi))$ is in $\mathrm{Sat}(\mathbf{Nbhd}_{S4})$ but not in $\mathrm{Sat}(\mathbf{Net})$

Dynamic Epistemic Logic

- We can do the DEL[20] trick and "dynamify" neural network semantics!
- This lets us compare neural network *updates* to updates over other models!

  (This is ongoing work — I have more questions than answers!)

Syntax:

$$\varphi \in p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box_i \varphi \mid [\varphi]\psi$$

Semantics:

- For each class of models $\mathcal{C}$, we pick from a menu of update functions

$$\text{Update}: \mathcal{C} \to \mathcal{L} \to \mathcal{C}$$

  We use the usual DEL semantics: for all $\mathcal{M} \in \mathcal{C}, w \in W$,

$$\mathcal{M}, w \Vdash [\varphi]\psi \qquad \text{iff} \qquad \text{Update}(\mathcal{M}, \varphi) \Vdash \psi$$

20. van Ditmarsch, H., van Der Hoek, W., and Kooi, B. *Dynamic epistemic logic*. Springer, 2007.

A Menu of Model Updates

Updates for $\mathbf{Rel}$:[21]

- **Conditioning.** Given $\mathcal{M} = \langle W, R, V\rangle \in \mathbf{Rel}$,

$$\text{Cond}(\mathcal{M}, \varphi) = \langle W \cap [\![\varphi]\!], R \cap ([\![\varphi]\!] \times [\![\varphi]\!]), V|_{[\![\varphi]\!]}\rangle$$

Updates for $\mathbf{Plaus}$:

- **Lexicographic Rev.** Lex$(\mathcal{M}, \varphi)$ *reorders* the plausibility order as follows:

  Make all $[\![\varphi]\!]$-states better than all other states,
  preserving the order within both groups.

- **Conservative Rev.** Cond$(\mathcal{M}, \varphi)$ reorders in a more conservative way:

  Make only the best$_R([\![\varphi]\!])$-states better than all other states,
  preserving the order within both groups.

21. van Benthem, J. Dynamic logic for belief revision. Journal of Applied Non-Classical Logics, 2007.

A Menu of Model Updates

Updates for $\mathbf{Nbhd}$:[22]

- **Evidence Addition** (and other evidence management operators)

$$\mathcal{M}^{+\varphi} = \langle W, f^{+\varphi}, V \rangle, \quad \text{where} \quad f^{+\varphi}(w) = f(w) \cup \{[\![\varphi]\!]\}$$

Updates for $\mathbf{Net}$:[23]

- **Iterated Hebbian Update.**
  "Neurons that fire together wire together"

$$\text{Hebb}^{\star}(\mathcal{N}, \varphi) = \langle N, \text{bias}, \{E_i\}_{i \in \boldsymbol{I}}, \{W_i^{\star}\}_{i \in \boldsymbol{I}}, \{A_i\}_{i \in \boldsymbol{I}}, V \rangle$$

where we (1) take the nodes $\mathsf{Cl}_i([\![\varphi]\!])$ activated by $[\![\varphi]\!]$, and
(2) repeatedly increase the involved weights iter number of times (learning rate $\eta \geq 0$):

$$W_i^{\star}(u, v) = W_i(u, v) + \text{iter} \cdot \eta \cdot \chi_{\mathsf{Cl}_i([\![\varphi]\!])}(u) \cdot \chi_{\mathsf{Cl}_i([\![\varphi]\!])}(v)$$

.

22. Pacuit, E. Neighborhood semantics for modal logic. Springer, 2017.
23. Schultz Kisby, C., Blanco, S., and Moss, L. What Do Hebbian Learners Learn? Reduction Axioms for Iterated Hebbian Learning. Proceedings of AAAI, 2024.

Ongoing Work & Future Directions

Open Question 1. In the static case, we know that $\mathrm{Sat}(\mathbf{Net}) = \mathrm{Sat}(\mathbf{Plaus})$. Is this true with updates as well? In particular:

- Is there some $\mathbf{Plaus}$ update that simulates Hebb*?
- Is there some $\mathbf{Net}$ update that simulates Lex? How about Consr?

Open Question 2. How does the learning power (in the limit) of $\mathbf{Net}$ updates like Hebb* compare to other updates, e.g. Cond, Lex, Consr, and evidence management?

Open Question 3. What about other neural network updates? Can we formalize backprop as a dynamic update? What laws are sound for backprop?

Open Question 4. Can we extend this to first-order logic? What neural network semantics is most appropriate for first-order quantifiers?

Thank you for your attention!

$$
\begin{tikzcd}
& \mathrm{Sat}(\mathbf{Nbhd}_{\mathbf{S4}}) \arrow[dr, phantom, "\supset" sloped] & \\
\mathrm{Sat}(\mathbf{Plaus}) \arrow[ur, phantom, "\subset" sloped] \arrow[dr, phantom, "\supset" sloped] \arrow[rr, dotted, no head, "="] & & \mathrm{Sat}(\mathbf{Net}) \\
& \mathrm{Sat}(\mathbf{Rel}_{\mathbf{S4}}) \arrow[ur, phantom, "\subset" sloped] &
\end{tikzcd}
$$

Contact:
Caleb Schultz Kisby
cckisby@gmail.com
https://ais-climber.github.io/