What Do Hebbian Learners Learn?

Reduction Axioms for Iterated Hebbian Learning

Caleb Schultz Kisby,

with Saúl Blanco, Larry Moss

Indiana University

AAAI 2024

February 22, 2024

Foundations for Neuro-Symbolic AI

From van Harmelen (2022):

> “What are the possible interactions between knowledge and learning? Can reasoning be used as a symbolic prior for learning . . . Can symbolic constraints be enforced on data-driven systems to make them safer? Or less biased? Or can, vice versa, learning be used to yield symbolic knowledge? And if so, how to manage the inherent uncertainty that comes with such learned knowledge . . .”
>
> “. . . **neuro-symbolic systems currently lack a theory that even begins to ask these questions, let alone answer them.**”

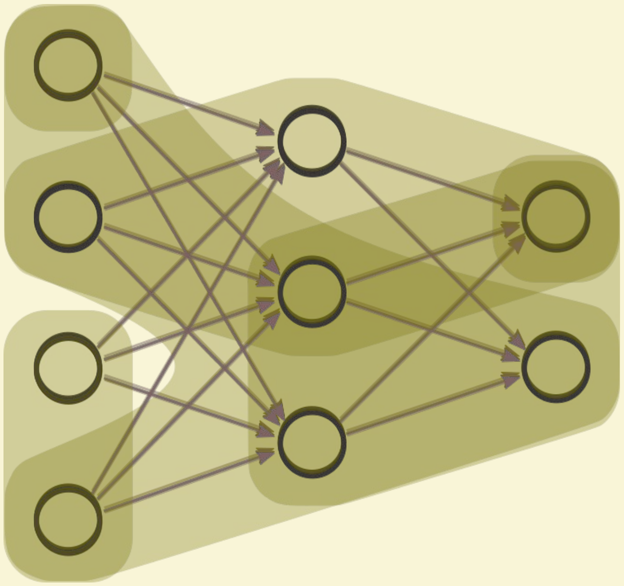
F. Harmelen. “Preface: The 3rd AI Wave Is Coming, and It Needs a Theory”. In: Neuro-Symbolic Artificial Intelligence. Ed. by P. Hitzler and M. Sarker. IOS Press BV, 2022.

A Brief Timeline

Defeasible Conditionals

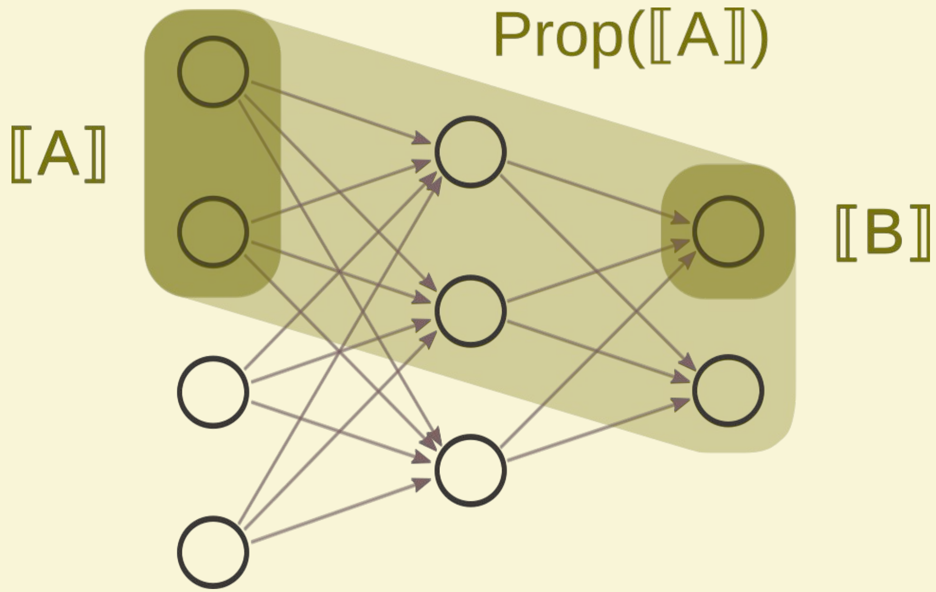
Neural Network Semantics

- **Key Idea:** Neural networks are not merely black boxes! Instead, think of nets as a kind of possible-worlds model; its activation patterns (states) contain information about its conditional beliefs.



- **We assume:** The network is the standard weighted feed-forward net; binary activations (states are just sets of neurons); fully-connected

Neural Network Semantics (Contd.)



Soundness and Completeness

Soundness

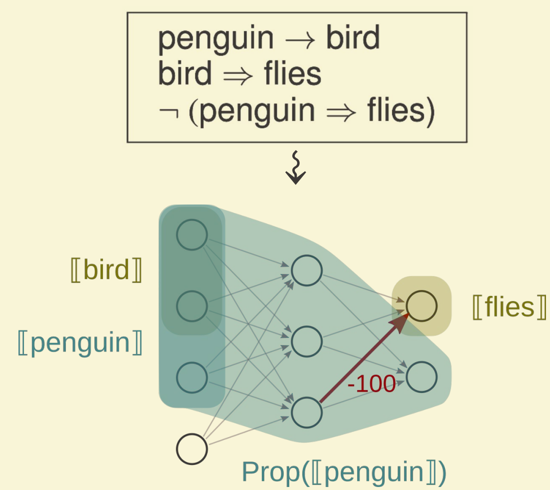
$\Gamma \vdash A$ implies $\Gamma \models A$

- **Not:** An explanation of a *particular* neural network's behavior
- **But instead:** Sound rules give *high-level* properties for *all* neural networks (of a certain architecture)

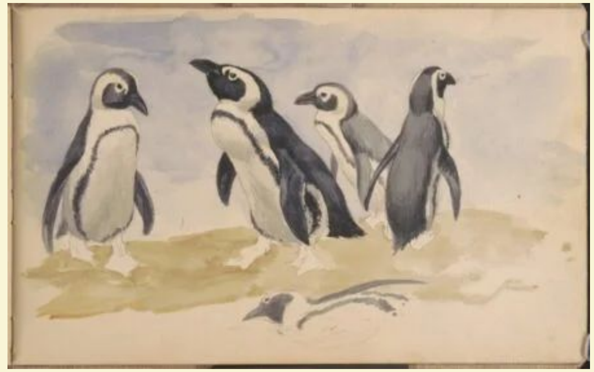
Completeness

$\Gamma \models A$ implies $\Gamma \vdash A$

- **Equivalently:** Can we build a neural network satisfying the set Γ of constraints?

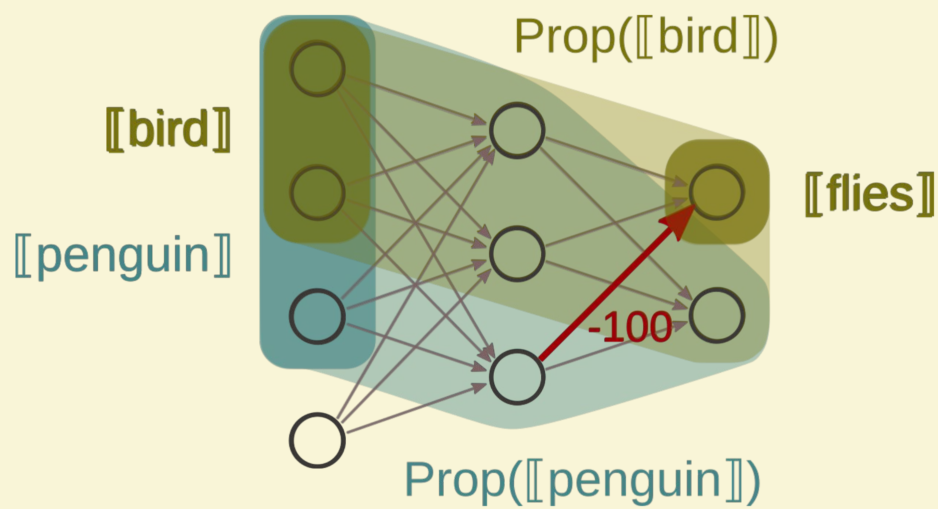


Example: Building a Neural Network



$$\begin{array}{l} \text{penguin} \rightarrow \text{bird} \\ \text{bird} \Rightarrow \text{flies} \\ \neg\,(\text{penguin} \Rightarrow \text{flies}) \end{array}$$

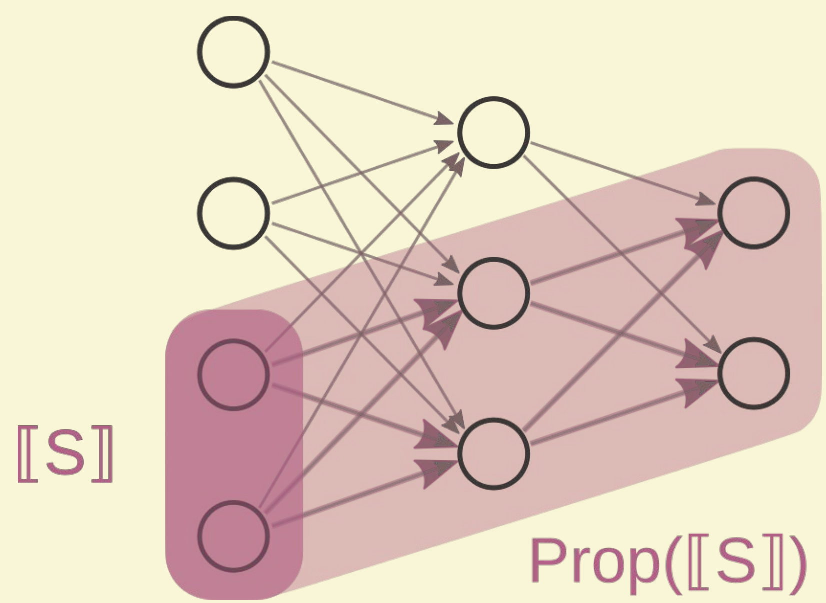
$\rightsquigarrow$



Louis Agassiz Fuertes. Penguins (1895). Watercolor and pencil on paper. Image from Cornell University Lab of Ornithology Art Collection.

Iterated Hebbian Learning

Neurons that fire together wire together

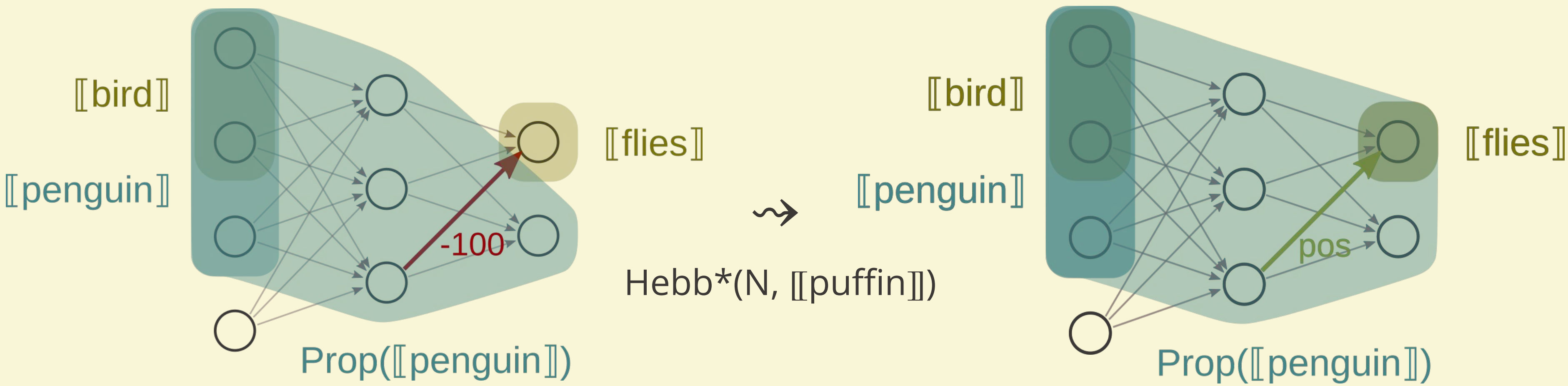


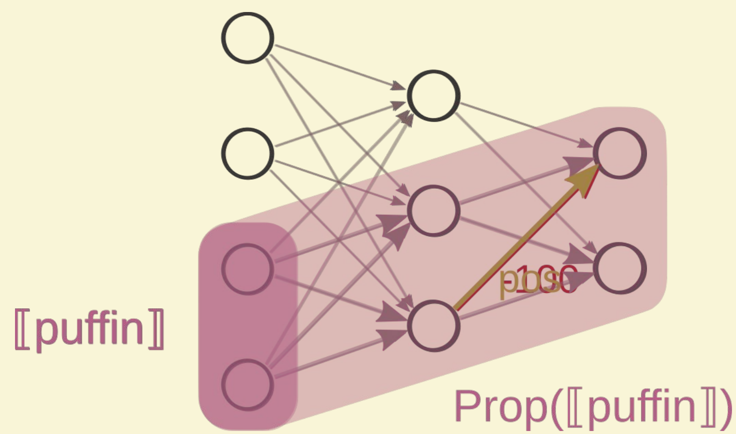
Repeat this update until a fixed point!
i.e. until the weights are “maximally high”

We call the resulting net **Hebb*(N, ⟦S⟧)**

D. Hebb. The Organization of Behavior. Psychology Press, 1949.

Example: Learning Wrecks the Model!





Louis Agassiz Fuertes. Atlantic Puffin (1932). Watercolor and pencil on paper. From *Portraits of New England Birds*. Commonwealth Of Massachusetts, 1932. Edited by Sabrina Schultz Kisby.

Logic & Formal Semantics

Main Results

Theorem. The following axioms are sound:

$$
\begin{array}{lcll}
[\varphi]p & \leftrightarrow & p & \text{for propositions } p \\
[\varphi]\neg\psi & \leftrightarrow & \neg[\varphi]\psi & \\
[\varphi](\psi \wedge \rho) & \leftrightarrow & [\varphi]\psi \wedge [\varphi]\rho & \\
[\varphi]\mathbf{K}\psi & \leftrightarrow & \mathbf{K}[\varphi]\psi & \\
[\varphi]\mathbf{T}\psi & \leftrightarrow & \mathbf{T}([\varphi]\psi \wedge (\mathbf{T}\varphi \vee \mathbf{K}(\mathbf{T}\varphi \vee \mathbf{T}[\varphi]\psi))) &
\end{array}
$$

Theorem. **Assuming** model building for the base language: For all consistent $\Gamma \subseteq \mathcal{L}$ there is a net $\mathcal{N}$ such that $\mathcal{N} \models \Gamma$.

Theorem. **Assuming** completeness for the base language: $[\varphi]$ is completely axiomatized by the reduction axioms from before.

Future Work

References

Axioms for The Base Logic

A Complete Reduction for Hebb*
(Explained!)