Neural Network Models: Thesis Outline

1. **Introduction**

    - Introduce in the modern context of GPT & transformers not being able to reason / explain their reasoning, “black boxes” making important decisions, of “AI Alignment”, interpretability, explainability, self-driving cars making mistakes due to lack of reasoning. There have been many, many proposals for fixing this, but the choices are disparate and it’s not clear how they relate to each other, or if there are any mathematical guarantees we can make, etc. Emphasize the need for principled, mathematical foundations of neural networks, especially of integrating *learning and reasoning*, as well as *neural and symbolic systems*. (*So many* speakers have emphasized this point, draw on all of the ones I know) This thesis stands in the center of this tension. (Make claims about what this thesis does. Since the consequences of this work are interdisciplinary, I’ll follow this up with different introductions depending on where the reader is coming from (literally just tell the reader to skip to their section first, then return to the others):)
    - Main Thesis
        - **Simple Idea:** Take ordinary neural networks, attach a valuation function (interpretation) to them, and then apply standard logic techniques (treat neural networks as a model for logic). The claim is that this is a fruitful approach for understanding neural networks — many things a logician wants to know turn out to correspond to things a machine learning researcher wants to know:
            - Static Logic:
                - Soundness : Formally verifying properties of neural network inference
                - Completeness : Neural network model building (encoding desired inferences in a net)
                - Expressive Power — “What is the minimal logic that can express closure properties of neural networks” is a sort of descriptive complexity version of “What functions can neural networks encode/represent?” (It would be cool to relate this work to Will & Lena & et al’s work on this.)
            - Dynamic Logic:
                - Soundness : Formally verifying properties of neural network *learning*
                - Completeness : Neural network model building with *learning* constraints (one of the goals of AI Alignment)
                - Expressive Power — Somehow related to “What functions can neural networks *learn*”. Though be careful here; most properties of learning probably won’t be expressible *in* the logic per se.
    - Logic

  - The bulk of the work in this thesis really is standard logic methodology stuff — we define models for a language, give formal semantics, prove soundness and completeness of axioms, etc. We expect logicians to feel the most at home here. Consider neural network models as an *alternative choice* to, e.g., possible-worlds models, KLM-style plausibility models, neighborhood models, etc. This work spans all sorts of different logic systems, including conditional logics, modal logics, epistemic and doxastic logics, dynamic logics, and belief revision.

- Machine Learning
  - This work is about interfacing logic with neural networks. I do this using the standard logic methodology, but it turns out that issues like soundness, completeness, minimal models, etc. correspond directly to issues about the verification and construction of neural networks with certain properties! The best way to think about this work is a map for building neural networks that have learning constraints, as well as proving general characterizations of different learning methods.
- Cognitive Science
  - Mental representations, symbol grounding, neuro-philosophy and connectionism, frame problem
- Computational Learning Theory
  - Emphasize that these are two different theoretical perspectives on understanding machine learning. Say what the benefits of each are, and then foreshadow to the connection/bridge between them.
- Epistemology
  - Explain what an epistemologist gets from this; what a neuro-philosopher gets from this; what a philosopher of mind gets from this
- Dynamical Systems
  - I can't do much for the dynamical systems mathematician, but I do want to encourage more of these people to consider working on these problems! Point them to *specific* open problems that rely on dynamical systems expertise! (e.g. the stability of forward propagation, stable descriptions of iterated learning methods, etc.)

2. **Basics of Neural Network Models**

   2.1. Neural Network Preliminaries

   2.2. State and Forward Propagation

   - The idea is very simple: Neural network models are just ordinary neural networks, plus its current 'state'
   - Clarify what we mean by state in a general sense, assumptions we make on state (allowing both binary and fuzzy)
   - Some examples! Clarify what we mean by the net's "inference"

    - The state of a neural network changes (according to its activation function)
    - Assumptions we make about the forward propagation closure operator (e.g. it stabilizes).
    - properties of propagation

  2.3. Neural Network Semantics for Belief

    - Give several languages in increasing order of expressive power: Belief, conditional belief, and ‘best’/prototypes. Since the ‘best’ language can express both belief and conditional belief, I’ll stick with it for the rest of this thesis.

  2.4. Inference and Axioms

    - These neural network semantics satisfy exactly the axioms of ‘best’ given in Appendix A (prove them each individually here, although it’s just a quick check)

  2.5. Model Building and Completeness

    - Include the modifications to the completeness proof for each conditional axiom we could satisfy.

  2.6. Reflections on Methodology

    - The main point: Forward propagation is a sort of prototype — generally, identify *what closure operators over the network are important*. For forward propagation, we mapped it to the ‘best’ modality
    - (new subsection) Graph Reachability is another good example — show that our completeness proof extends to **K**: Reach (Alexandru said he is skeptical of this point, so I should clarify that the network flips in the construction, so “worlds above” also flips)
    - Determining which closure operators are most relevant for understanding a neural network architecture is an art. For feed-forward nets (in general, terminating nets), it’s clear that forward propagation carries the full information of its inference. What about unstable/oscillating nets? What about first-order quantifiers? etc.
    - Our story doesn’t end at the dynamics of inference/forward propagation. In fact, the main contribution of this thesis is an account for *learning* on neural network models. The trick is essentially the same, extending it with the DEL methodology (will explain)

3. **Dynamics on Neural Network Models**

  3.1. Hebbian Learning Inspired Update

    - Explain the idea behind Hebbian learning (we’re using it as a simple update on neural networks)
    - Give the four updates that I’ve come up with (“make neurons wire together”; “only if they fired together”; “iterated Hebbian learning”; “single-step Hebbian learning”) and the relationships between these methods

- Reduction Axioms and Completeness

3.2. Properties of Hebbian Update

3.3. Reduction Axioms and Completeness

- Give reduction axioms for all three of the methods, and then also think about completeness for single-step update!

3.4. Expressive Power of Neural Network Update

- Answer the questions: Are there any classical updates (over plausibility models) corresponding to our three neural network updates (turns out to be no! — consider graded plausibility and in the worst-case think about neighborhood models)? Are there any neural network updates corresponding to plausiblity updates (conditioning, lexicographic, conservative)?

3.5. Reflections on Interpretability and Alignment

- give an explicit example! Show an actual neural network with learning guarantees!
- Mention the caveat on interpretability, which is that we *don't* have a classical model corresponding to Hebbian learning

4. **Bridges to Related Work**

4.1. Bridge to Other Neural-Symbolic Proposals

- Understand the recent survey by Simon & Artur
- Understand & relate Logic Tensor Networks
- Understand & relate Neural ProbLog

4.2. Bridge to Social Networks

4.3. Bridge to Neural Network Extraction

- We assumed that the interpretation of neurons is given to us. But the task of neural network extraction is identifying these variables! But this work does give some perspective to neural network extraction (think of it as coming up with a valuation function without knowing it a priori). Think about how this relates to Thomas Icard's work

4.4. Bridge to Computational Learning Theory

- Here it would be good to formalize the learning method corresponding to neural network update (see Alexandru, Sonja, and Nina's paper, and model it in a similar way)

4.5. Bridge to Cognitive Science

- A simple idea that connects neuroscience/connectionism with psychology/conceptual cognition. Does this say anything about linguistics? What exactly does this suggest / what assumptions are we making about how we should interpret the conceptual contents of brains?

5. **Future Directions and Open Problems**

5.1.

Appendix