In the last 15 years, modern machine learning systems based on neural networks have shown unprecedented success at learning from data with little human guidance. [To do: Give examples: ChatGPT, AlphaZero, ...] [To do: Advantages of neural network systems] Neural networks are flexible and excel at learning from unstructured data. These methods come with no safety or reliability guarantees, and they notoriously lack transparency. In practice, a computational learner is often a 'black-box' whose correct inferences, mistakes, and biases lack interpretation and explanation.

Prior to the widespread use of neural networks in artificial intelligence, the use of symbolic (logic) systems. In contrast, symbolic (logic) systems in artificial intelligence are pretty good at sophisticated explainable reasoning. They provide explicit rules for their reasoning in a human-interpretable language. But historically, symbolic systems are brittle, and fail to model actions that change systems over time. This is the well-known frame problem in AI [To do: cite SEP]. [To do: Bring in Critique of Pure Reason, failure of logic systems]

There is some hope that we might be able to integrate symbolic systems and neural networks while retaining the advantages of both. In particular, how can we better understand and control the seemingly black-box behavior of neural networks using logic? In response to this possibility, the field of neuro-symbolic AI has emerged to do just that [To do: cite Bader and Hitzler 2005; Besold et al. 2017; Sarker et al. 2022]. [To do: Mention journal, conference, book] There is now a wide range of neuro-symbolic systems, including Logic Tensor Networks [To do: cite], Distributed Alignment Search [To do: cite], DeepProbLog [To do: cite], Logic Explained Networks [To do: cite], and neural network fibring [To do: cite]. But these systems are disparate, and together form a scattered picture. In the inspiring preface to the recent book *Neuro-Symbolic Artificial Intelligence: The State of the Art*, Frank van Harmelen writes [To do: cite]:

> 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? ... **neuro-symbolic systems currently lack a theory that even begins to as**

1 Here's a section

Theorem 1 *blah blah blah*

In this thesis, I propose a bridge between neural networks and logic that sheds light on existing neuro-symbolic systems, providing a unifying perspective. The critical idea is

---

[Clear need for a unifying theory — quote [**?**]] [Give explicit questions about neural networks that we don't have a framework for answering]

[Many neuro-symbolic systems are contained within a larger umbrella; explain the history of treating neural networks directly as a class of models in formal logic]

Thesis: [The point of this dissertation is to explore and develop these *neural network semantics* — to take this idea as far as it can go, and see what we get] [Over the course of this development, I will show how foundational questions about neural networks that were once elusive become natural and answerable questions in logic:]

Thesis: Neural networks can be treated as a class of models in formal logic, simply by adding an interpretation function. By doing so, foundational questions about neural network inference and learning that were once elusive become natural and answerable questions in logic:

Talk Abstract: [In this talk I will present one such proposal that is close to the hearts of modal and epistemic logicians: Treat (binary) neural networks as a class of models in modal logic by (1) adding a valuation of propositions (as sets of neurons), and (2) interpreting $\Diamond\varphi$ as the forward propagation (or diffusion) of input φ through the net. We can then do "business as usual," using neural networks as our models. To cement this idea, I will compare the modeling power of neural networks with other classes of models, in particular: relational, plausibility, neighborhood, and social network models. If time permits, I will mention recent work in which we "dynamify" this logic, in the spirit of modeling neural network update and learning.]

Old paper blurb: In fact, there is an up-and-coming foundational theory for neuro-symbolic systems, which we call neural network semantics. Its key insight is that neural networks can be taken as models for a formal logic. Moreover, logical operators can be mapped to operators on neural network states. Alternatively, we can semantically encode classical model operators into neural operators (and vice-versa). ... We refer the reader to the landmark survey (Odense and d'Avila Garcez 2022), which shows that this framework encompasses a wide class of neuro-symbolic systems.

In response to this dichotomy, the field of neuro-symbolic AI has emerged — a community-wide effort to integrate neural and symbolic systems, while retaining the advantages of both. Despite the many different proposals for neuro-symbolic AI (too many to list! See (Bader and Hitzler 2005; Besold et al. 2017; Sarker et al. 2022)), there is little agreement on what the interface between the two ought to be. There is a clear need for a unifying theory that can explain the relationship between neural networks and symbolic systems (Harmelen 2022). In fact, there is an up-and-coming foundational theory for neuro-symbolic systems, which we call neural network semantics. Its key insight is that neural networks can be taken as models for a formal logic. Moreover, logical operators can be mapped to operators on neural network states. Alternatively, we can semantically encode classical model operators into neural operators (and vice-versa).

historically, logics are not designed to capture dynamics of the system and fail to explain actions that change systems over time. This is one of the faces of the well-known frame problem in AI [49]. One way to escape it, while preserving the benefits of logic, is through dynamic logic.

The two dominant paradigms of AI, connectionist neural networks and symbolic systems, have long seemed irreconcilable. Symbolic systems are well-suited for giving explicit inferences in a human-interpretable language, but are brittle and fail to adapt to new situations. On the other hand, neural networks are flexible and excel at learning from unstructured data, but are considered black-boxes due to how difficult it is to interpret their reasoning. In response to this dichotomy, the field of neuro-symbolic AI has emerged — a community-wide effort to integrate neural and symbolic systems, while retaining the advantages of both.

An answer could be given by symbolic systems which are good at sophisticated explainable reasoning. Their logical formulae are readable, come with a well-defined meaning, and are equipped with explicit rules for their reasoning. Consequently, in TCS in general, and in the field of neuro-symbolic AI in particular, logic is used to describe and reason about the behavior of computational (learning) systems

In recent years, modern machine learning systems have shown unprecedented success at learning

from data with little human guidance. Algorithms used to provide solutions to societal problems in the public sphere are often based on neural networks and large amounts of data, and so these technologies affect increasingly larger populations (see, e.g., [43]). At the same time, these methods come with no safety or reliability guarantees, and they notoriously lack transparency. In practice, a computational learner is often a 'black-box' whose correct inferences, mistakes, and biases lack interpretation and explanation

Thesis: [The point of this dissertation is to explore and develop these *neural network semantics* — to take this idea as far as it can go, and see what we get] [Over the course of this development, I will show how foundational questions about neural networks that were once elusive become natural and answerable questions in logic:]

Soundness	answers	"How can we formally verify that a class of neural networks have certain properties?
Completeness	answers	"How can we build a neural network that aligns with constraints?"
Expressivity	answers	"What kinds of functions are neural networks capable of representing?"

Moreover, we can extend this approach to a *dynamic logic* perspective, where we can answer questions about neural network *learning*: [say more! I have to explain what this 'dynamic logic' means!]

Soundness	answers	"How can we formally verify that a class of neural network *learning policies* have certain properties?
Completeness	answers	"How can we build a neural network that aligns with constraints on its behavior *before and after learning takes place*?"
Expressivity	answers	"What kinds of *learning policies* are neural networks capable of supporting?"

[Then give a sort of outline for the rest of the thesis] [probably background?] [Part I: Neural Network Semantics for Inference] [Early on, give Hannes' semantics, and then explain how this underlying idea (using Simon Odense's survey) is the basic structure for many different neuro-symbolic systems] [Then in another chapter, we can generalize to modal logic] [Rethink this structure (?)] [Part II: Neural Network Semantics for Update]

The idea that neural networks can be viewed as models for logic dates back to (McCulloch and Pitts 1943). But the neural network semantics we present here builds on a recent reimagining of this (Balkenius and Gardenfors 1991; Leitgeb 2018), where logical formulas are ¨ mapped to states of the net rather than to individual neurons (thus avoiding the "grandmother cell" problem (Gross 2002)). Early work established the formal correspondence between forward propagation and conditional belief (Balkenius and Gardenfors 1991; Leitgeb 2001, 2003; Blutner 2004). ¨ Note that all of this early work focuses on binary nets. More recently, (Giordano and Theseider Dupre 2021) and (Gior-´ dano, Gliozzi, and Theseider Dupre 2022) prove soundness ´ for forward propagation over fuzzy neural networks. And as mentioned above, (Kisby, Blanco, and Moss 2022) shows soundness — but not completeness — for a simple Hebbian learning policy

---

The two dominant paradigms of AI, connectionist neural networks and symbolic systems, have long seemed irreconcilable. Symbolic systems are well-suited for giving explicit inferences in a human-interpretable language, but are brittle and fail to adapt to new situations. On the other hand, neural networks are flexible and excel at learning from unstructured data, but are considered black-boxes due to how difficult it is to interpret their reasoning. In response to this dichotomy, the field of neuro-symbolic AI has emerged — a community-wide effort to integrate neural and symbolic systems, while retaining the advantages of both. Despite the many different proposals for neuro-symbolic AI (too many to list! See (Bader and Hitzler 2005; Besold et al. 2017; Sarker et

al. 2022)), there is little agreement on what the interface between the two ought to be. There is a clear need for a unifying theory that can explain the relationship between neural networks and symbolic systems (Harmelen 2022). In fact, there is an up-and-coming foundational theory for neuro-symbolic systems, which we call neural network semantics. Its key insight is that neural networks can be taken as models for a formal logic. Moreover, logical operators can be mapped to operators on neural network states. Alternatively, we can semantically encode classical model operators into neural operators (and vice-versa).

The central questions this theory aims to answer are: Soundness. What axioms are sound for neural network operators? Can neural operators be mapped to classical ones in a sound way? Note that checking soundness is equivalent to formally verifying properties of nets. Completeness. What are the complete axioms for neural network operators? This is equivalent to model building: Can we build a neural network that obeys a set of logical constraints Γ? Can we build a neural network from a classical model? We refer the reader to the landmark survey (Odense and d'Avila Garcez 2022), which shows that this framework encompasses a wide class of neuro-symbolic systems. We will discuss other work that we consider part of the core theory in the next section. The standard example is the forward propagation operator Prop over a net N . Active neurons in a state S successively activate new neurons until eventually the state of the net stabilizes — Prop(S) returns the state at the fixed point. A classic result from (Leitgeb 2001) is this: Say conditionals Then, in a binary feed-forward net, Prop is completely axiomatized by the loop-cumulative conditional laws of (Kraus, Lehmann, and Magidor 1990). The result is robust, and can be extended to different choices of conditional axioms and neural network architectures (Leitgeb 2003). The general takeaway is that forward propagation corresponds to a nonmonotonic conditional. A central challenge for this theory is to do the same for neural network learning operators. Our previous work (Kisby, Blanco, and Moss 2022) considers a simple learning policy — na¨ıve Hebbian update ("neurons that fire together wire together") — on a binary, feed-forward net. Although this work offers sound axioms for Hebbian learning, the question of completeness is left open.

Neural networks are very good at learning without human guidance, yet they're also known for making blunders that seem silly from the point of view of logic. (And this situation hasn't changed, despite modern neural network systems like GPT-4). This is a long-standing problem in artificial intelligence: How can we better understand and control neural networks using logic? In response, there have been countless proposals for "neuro-symbolic" systems that incorporate logic into neural networks, or vice versa.

In this talk I will present one such proposal that is close to the hearts of modal and epistemic logicians: Treat (binary) neural networks as a class of models in modal logic by (1) adding a valuation of propositions (as sets of neurons), and (2) interpreting $\Diamond\varphi$ as the forward propagation (or diffusion) of input φ through the net. We can then do "business as usual," using neural networks as our models. To cement this idea, I will compare the modeling power of neural networks with other classes of models, in particular: relational, plausibility, neighborhood, and social network models. If time permits, I will mention recent work in which we "dynamify" this logic, in the spirit of modeling neural network update and learning.

In recent years, modern machine learning systems have shown unprecedented suc- cess at learning from data with little human guidance. Algorithms used to provide solutions to societal problems in the public sphere are often based on neural networks and large amounts of data, and so these technologies affect increasingly larger populations (see, e.g., [46]). At the same time, these met- hods come with no safety or reliability guarantees, and they notoriously lack transparency. In practice, a computational learner is often a 'black-box' whose correct inferences, mistakes, and biases lack interpretation and explanation. This is a deep problem that cuts across artificial intelligence (AI), theo- retical computer science (TCS), and cognitive science: How can we reason about, understand, and guide computational learning processes?

Humans make intelligent decisions by seamlessly integrating both their ability to learn and their ability to reason about what they have learned. But researchers in artificial intelligence have long experienced a tradeoff between the two: Neural systems have had tremendous success learning from unstructured data, whereas symbolic systems excel at sophisticated reasoning tasks that neural systems cannot readily learn. In the last three decades, there have been countless hybrid systems that combine neural and symbolic components in a myriad of ways, hoping to strike the right balance [1, 6, 15, 22, 25]. Other authors [2, 5, 8, 7, 10, 16, 17] suggest a more principled approach: The neural and symbolic are two ways of interpreting the same agent, and we should be able to translate between the two. In fact, Garcez et. al. [8, 7] has demonstrated that we can extract (sound) knowledge from a net, as well as build a net that (completely) models existing knowledge. Equivalently, Leitgeb [16, 17] viewed neural networks as the semantics of a formal logic, and showed that this logic completely axiomatizes the behavior of the net. However, no existing neuro-symbolic proposal seamlessly interfaces learning and reasoning like humans do. Most neuro-symbolic hybrid systems, including that of Garcez et. al., treat learning as a black-box process that occurs before, after, or within a symbolic reasoner. In addition, more formal translations such as Leitgeb's do not even consider learning. As of yet, how learning relates to reasoning remains a mystery. This leads us to the central claim of my dissertation: