

# Reasoning about Neural Networks with Dynamic Logic

## Research Proposal

Caleb Schultz Kisby

December 13, 2024

In the last 15 years, modern artificial intelligence (AI) systems have shown unprecedented success at learning from data with little human guidance. Consider for example large language models such as Llama and GPT [1; 12; 34], which have taken the world by storm with their ability to learn to converse in English merely from unstructured text data they scrape off the web. Or consider AlphaGo [29], which learned to play Go at a human expert level by repeatedly playing against itself. These breakthroughs in machine learning are thanks to the widespread use of neural networks—brain-inspired computational models that excel at learning from unstructured data.

But the danger of neural networks is that they come with no safety, fairness, or correctness guarantees. If you play with systems like GPT long enough, you eventually realize that they carry all sorts of prejudices and misconceptions, make silly logical mistakes, and are quite happy to spew out disinformation [13; 24; 30]. Neural networks also lack transparency, which means diagnosing and correcting these errors is not feasible. In practice, neural networks are often treated as ‘black-boxes’ whose biases, mistakes, and correct inferences are impossible to predict or control.

How can we better reason about, understand, and guide the behavior of neural networks? The answer lies in symbolic (logic) systems, which were commonly used to model reasoning and intelligent behavior prior to the rapid growth of neural network systems. In contrast with neural networks, symbolic systems provide transparent rules for their reasoning in a human-interpretable language. Historically, logics have suffered from being unable to model flexible learning or update (known as the *frame problem* in AI [23; 28]). One way to escape this problem, while preserving the benefits of logic, is through *dynamic logic*.

## 1 Previous Work

Dynamic logic can be seen as a general set of tools for reasoning about many different kinds of actions and effects. It has been used to model a wide range of dynamic scenarios, including programming effects [26], quantum computation [9], multi-agent communication [31; 33], and social networks [5; 11]. It's not surprising that dynamic logic is also a natural choice for reasoning about learning [6; 8; 10], and in particular learning over neural networks.

To see this, first consider that a logical language can be interpreted directly over neural networks [4; 16; 19; 20; 21; 25]. This is done by interpreting some operator  $\langle \mathbf{T} \rangle \varphi$  as the forward propagation (or diffusion) of input  $\varphi$  through the net. Formulas  $\mathbf{T}\varphi \rightarrow \psi$  then express constraints on neural network *inference*, i.e., the input-output behavior of the net.

In the dynamic logic setting, we can similarly interpret a dynamic operator  $[P]$  directly as neural network *update*. In my previous work, I did this using the Logic of Hebbian Learning [17; 18]. In this logic, formulas express the effects of a simple learning policy, iterated Hebbian update, on a neural network. For example,  $(\mathbf{T}\varphi \rightarrow \psi) \wedge [P](\mathbf{T}\varphi \rightarrow \psi)$  says that the network classifies input  $\varphi$  as  $\psi$ , and iterated Hebbian learning of  $P$  preserves that fact.

## 2 Proposed Work

The dynamic logic approach to neural networks is in its early stages, and there are many questions left to be answered. Recall two questions I mentioned from the attached cover letter:

- How powerful and reliable are neural network learning algorithms?
- Is provably correct neural network alignment possible?

The goal of this project is to take steps towards answering these, using tools from dynamic logic.

### 2.1 Comparing Neural Network Update with Dynamic Updates

The use of dynamic logic to model neural network learning opens up the possibility of comparing the power and properties of neural network updates against previously-known dynamic update operators. For instance, the Hebbian update operator  $\text{Hebb}^*$  resembles certain belief revision policies over plausibility models [18; 31; 32]. This leads to a number of questions. Which neural network updates can be simulated by which plausibility updates, and vice-versa? Which properties of learners does  $\text{Hebb}^*$  satisfy, and can it learn a data stream in the limit (see [7])?

These neural network logics also bear striking resemblance to logics for *social networks* [2; 5; 11]. But whereas neural network logics model updates inspired by neural networks, social network logics model changes in social links between agents. Here again lie many mysteries and possible connections. Can different neural network updates simulate social network updates, and vice-versa? Can we give a unified account of neural and social network semantics together?

### 2.2 Building Aligned Neural Networks using Dynamic Logic

If our dynamic logic interpreted on neural networks is *complete*, this means we can build a neural network that obeys constraints on its behavior before and after learning. For example, the Logic of Hebbian Learning is indeed complete, and so we can build neural networks that obey constraints such as  $(\mathbf{T}\varphi \rightarrow \psi) \wedge [P](\mathbf{T}\varphi \rightarrow \psi)$ . But the Logic of Hebbian Learning is a simplified setting; it doesn't model learning used in practice, and falls short of the rich language we would need to state useful rules.

First, I will consider a dynamic logic which models the effects of the most widely used neural network learning algorithm: gradient descent, implemented as back-propagation [27]. This first requires an account of “supervised” updates  $[P; Q]$  in dynamic logic, i.e., observations with an expected answer. Second, I will consider a richer constraint language: First-Order Logic (FOL). Existing neuro-symbolic systems also use FOL to reason about and build neural networks, but it is still an open problem to prove that any such neural network mapping to FOL is sound.

Finally, I plan to develop a software suite that performs the neural network verification and model building. The user will provide learning constraints in a generous language of FOL alongside dynamic operators for neural network updates. Across the range of neuro-symbolic systems, including Logic Tensor Networks [3], Distributed Alignment Search [15], DeepProbLog [22], and neural network fibering [14], this will be the first ever such system that places constraints on the net's behavior before and after learning—and it will be exciting to put this feature to the test! For this, I would like to work with the Epistemological and Ethical ‘Explainable AI’ team to put together realistic, human-interpretable ethical constraints and formalize them in this language.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat et al. GPT-4 technical report. *ArXiv preprint arXiv:2303.08774*, 2023.
- [2] Edoardo Baccini, Zoé Christoff, and Rineke Verbrugge. Dynamic logics of diffusion and link changes on social networks. *Studia Logica*, pages 1–71, 2024.
- [3] Samy Badreddine, Artur d’Avila Garcez, Luciano Serafini, and Michael Spranger. Logic Tensor Networks. *Artificial Intelligence*, 303:103649, 2022.
- [4] Christian Balkenius and Peter Gärdenfors. Nonmonotonic inferences in neural networks. In *KR*, pages 32–39. Morgan Kaufmann, 1991.
- [5] Alexandru Baltag, Zoé Christoff, Rasmus K Rendsvig, and Sonja Smets. Dynamic epistemic logics of diffusion and prediction in social networks. *Studia Logica*, 107:489–531, 2019.
- [6] Alexandru Baltag, Nina Gierasimczuk, Aybüke Özgün, Ana Lucia Vargas Sandoval, and Sonja Smets. A dynamic logic for learning theory. *Journal of Logical and Algebraic Methods in Programming*, 109:100485, 2019.
- [7] Alexandru Baltag, Nina Gierasimczuk, and Sonja Smets. Truth-tracking by belief revision. *Studia Logica*, 107:917–947, 2019.
- [8] Alexandru Baltag, Dazhu Li, and Mina Young Pedersen. On the right path: A modal logic for supervised learning. In *International Workshop on Logic, Rationality and Interaction*, pages 1–14. Springer, 2019.
- [9] Alexandru Baltag and Sonja Smets. The logic of quantum programs. In P. Selinger, editor, *Proceedings of the 2nd International Workshop on Quantum Programming Languages (QPL2004)*, volume 33 of *TUCS General Publication*, pages 39–56. Turku Center for Computer Science, June 2004.
- [10] Alexandru Baltag and Sonja Smets. Keep changing your beliefs, aiming for the truth. *Erkenntnis*, pages 1–16, 2011.
- [11] Zoé Christoff and Jens Ulrik Hansen. A logic for diffusion in social networks. *Journal of Applied Logic*, 13(1):48–77, 2015.
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan et al. The Llama 3 herd of models. *ArXiv preprint arXiv:2407.21783*, 2024.
- [13] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in Large Language Models: A survey. *Computational Linguistics*, pages 1–79, 2024.
- [14] Artur SD’Avila Garcez, Luis C Lamb, and Dov M Gabbay. *Neural-Symbolic Cognitive Reasoning*. Springer Science & Business Media, 2008.
- [15] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR, 2024.
- [16] Laura Giordano, Valentina Gliozzi, and Daniele Theseider Dupré. A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps. *Journal of Logic and Computation*, 32(2):178–205, 2022.
- [17] Caleb Kisby, Saúl Blanco, and Lawrence Moss. The logic of Hebbian learning. In *The International FLAIRS Conference Proceedings*, volume 35. 2022.
- [18] 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*, volume 38, pages 14894–14901. 2024.
- [19] Hannes Leitgeb. Nonmonotonic reasoning by inhibition nets. *Artificial Intelligence*, 128(1-2):161–201, 2001.
- [20] Hannes Leitgeb. Nonmonotonic reasoning by inhibition nets II. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(supp02):105–135, 2003.
- [21] Hannes Leitgeb. Neural network models of conditionals. In *Introduction to Formal Philosophy*, pages 147–176. Springer, 2018.
- [22] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence*, 298:103504, 2021.

- [23] Drew McDermott. A critique of pure reason. *Computational intelligence*, 3(3):151–160, 1987.
- [24] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.
- [25] Simon Odense and Artur S. d'Avila Garcez. A semantic framework for neural-symbolic computing. *ArXiv*, abs/2212.12050, 2022.
- [26] John C. Reynolds. Separation logic: a logic for shared mutable data structures. In *Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science, LICS '02*, pages 55–74. USA, 2002. IEEE Computer Society.
- [27] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. *Biometrika*, 71(599-607):6, 1986.
- [28] Murray Shanahan. The frame problem. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016.
- [29] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.
- [30] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the capabilities, limitations, and societal impact of Large Language Models. *ArXiv preprint arXiv:2102.02503*, 2021.
- [31] Johan Van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.
- [32] Johan Van Benthem and Sonja Smets. Dynamic logics of belief change. In H. Van Ditmarsch, J. Halpern, W. van der Hoek, and B. Kooi, editors, *Handbook of Epistemic Logic*, pages 313–393. College Publications, London, UK, 2015.
- [33] Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*, volume 337. Springer, 2007.
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.