Symbol Manipulation in Neural Networks

Jonas Peters, Philipp Haack

August 7, 2025

Contents

1 Introduction **3**

1 Introduction

The human brain is a remarkably complex and efficient system, capable of performing an extraordinary range of cognitive tasks with ease. Its ability to learn from experience, adapt to new environments, and process vast amounts of sensory information has long fascinated scientists and engineers alike.

The study of its functionality is partially in the domain of computer science: We try to model the brain with computer algorithms, which is inspired by the idea that the human brain consists of elementary computational units, neurons, which itself can be sufficiently modeled with computers, and whose interplay creates the emergent complex behavior we see. This leads to the following premise:

Premise 1.0.1. *In theory, the human brain can be sufficiently modelled by a computer program.*

Over many decades scientists have been trying to make progress towards this monumental task of modelling the human brain. As a result, two paradigms on how to implement intelligent models emerged, they are called *connectionism* and *symbolism*. While connectionist models rely on training data, symbolist models have all their inference rules and algorithms directly coded into them. As a result, connectionist models typcilally are black boxes, i.e. we cannot retrace how the model inferred its output semantically.

Among other things, this is one Marcus' critique of contemporary LLMs. On his blog he writes:

> **G. Marcus (2025)**
>
> LLM are giant, opaque black boxes with no explicit models of the world at all. Part of what it means to say that an LLM is a black box is to say that you can't point to an articulated model of any particular set of facts inside.

To understand why Marcus is so critical of connectionist models which include LLMs, we will analyze one of his early writings, namely *The Algebraic Mind* G. F. Marcus, 2001. In this book, he argues that the popular architecture of MLPs (multi-layer perceptrons) is not sufficient to model the human brain, because it cannot represent certain algebraic structures.

In this essay, we will analyze Marcus' critique of connectionist models and his arguments against the MLP architecture, where we focus on chapter 3 *Relations between Variables*. We will also discuss the relevance of his critique in contemporary artificial intelligence research, especially in the context of LLMs.

Bibliography

Marcus, G. (2025, June). *Generative ai's crippling and widespread failure to induce robust models of the world*. Retrieved August 7, 2025, from https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. The MIT Press.