Joscha Bach says that we need to understand AIs think to understand how we think. Generally, we need to know how AI models think, form knowledge, and potentially reinforce biases.

My project is an experiment in neural epistemology, which aims to explore how beliefs are formed and consolidated computationally. Inspired by the work of philosopher Richard Rorty, who was skeptical of the idea that there is a fixed, objective structure to knowledge that we can discover through introspection or reflection, I aim to offer insights into how neural networks might similarly construct their own “knowledge” from data, aligning with Rorty's ideas on the contingency of language and the formation of beliefs.

My goal is to design a primary network engaged in a task related to neural epistemology (forming beliefs based on input data), with a secondary metacognitive network that will reflect on the primary's belief formation process, possibly assessing the strength or consistency of those beliefs. One possibility is that the secondary network’s interpretations and evaluations of the primary network might not necessarily reveal any “objective” or “true” structure of the primary network's knowledge, but rather, it forms its own contingent beliefs or interpretations based on its architecture and training.

1. **Modeling Metacognition**: A secondary network evaluating the primary one's decisions can provide a lens into neural introspection, analogous to our human introspective capabilities. It can be seen as a computational reflection on Rorty's skepticism about epistemology.

* Start with **Neural Epistemology**. Finally, it's worth noting that, while the technical depth of your project will certainly be a factor, your unique philosophical lens and ability to critically reflect on the implications of your work will be a distinguishing feature. Emphasizing this interdisciplinary approach in your portfolio and discussions with recruiters will highlight the unique value you bring.

It’s an ethical imperative to know how AI models think, form knowledge, and potentially reinforce biases. My project is an experiment in neural epistemology, which aims to explore how beliefs are formed and consolidated computationally. My goal is to use PyTorch to design a primary network engaged in a task related to neural epistemology, such as forming beliefs based on input data. Then I will design a secondary metacognitive network that will reflect on the primary belief formation process, possibly assessing the strength or consistency of those beliefs. One possibility is that the secondary network’s interpretations and evaluations of the primary network might not necessarily reveal any “objective” or “true” structure of the primary network's knowledge, but rather, it forms its own contingent beliefs or interpretations based on its architecture and training. I aim to offer insights into how neural networks might similarly construct their own “knowledge” from data, inspired by Rorty's ideas on the contingency of language and the formation of beliefs.

FINAL PROJECT DESCRIPTION

Understanding the cognition and biases of AI models is a technological challenge and ethical imperative. With this in mind, my project dives deep into the realms of neural epistemology, seeking to decipher how computational systems, analogous to sentient beings, form and stabilize beliefs. Using PyTorch, I will design a primary neural network that interprets and synthesizes beliefs from a stream of input data. This will be complemented by a secondary metacognitive network, which critically reflects upon the belief structures formed by the primary network. This introspective layer will assess belief strength, consistency, and potential biases.

This is inspired by the philosopher Richard Rorty’s claims about the contingency of language and belief formation, bringing forth the idea that beliefs, whether human or artificial, are not merely reflections of an external truth, but constructs influenced by the medium and process of their formation. One could see how the secondary network, rather than uncovering an “objective truth” about the primary’s knowledge, might itself manifest contingent interpretations, shaped uniquely by its own architecture and learning experience.

Beyond the philosophical, the findings from this project could shed light on the inherent subjectivity of AI systems, guiding the tech industry’s approach to model interpretability, bias assessment, and ethical AI deployment.