

PROJECT SUMMARY

Overview

We present a new brain-inspired computing paradigm, *neuromorphic metacognition*, in which, **when a human is interacting with an AI system**, an independent neuromorphic architecture observes and evaluates that interaction and provides a graded concern signal when the information, advice, or direction provided by the AI warrants human reconsideration. This research builds upon five progressive stages of our prior work: Fuzzy Set and Rough Set computation on embedded NVIDIA/CUDA hardware; mathematical modeling of human connectome development; physical FPGA implementation of synthetic synaptogenesis and self-learning neuromorphic computation; development of a larger neuroCMOS/FPGA architecture; and the theoretical development of neuromorphic metacognition. These results establish that the fundamental components have been demonstrated mathematically, computationally, and physically. The remaining challenge is scale. For this FET research effort, we will construct and experimentally evaluate a large-scale computational model integrating these mechanisms into an architecture comprising millions of synthetic states and dynamically developing synaptic connections. Rather than constructing another AI to supervise an existing AI, we investigate **two independent computational systems**: the external generative, adaptive, or agentic AI that provides intelligence, and our brain-inspired neuromorphic architecture that independently provides metacognition. The architecture will be evaluated to determine whether uncertainty, ambiguity, inconsistency, contextual conflict, and anomalous structure can collectively produce a meaningful graded concern signal advising the human user, “You may want to reconsider this.”

Intellectual Merit

The intellectual merit of this research is the development and experimental evaluation of a fundamentally different computational approach to AI oversight. The proposed architecture integrates Fuzzy Set and Rough Set mathematics for reasoning under uncertainty with a dynamically developing connectome, synthetic synaptogenesis, and self-learning neuromorphic computation. Unlike approaches in which one AI evaluates another, our architecture remains computationally independent of the AI it observes and is modeled upon principles of human synaptic development. We will determine whether mechanisms demonstrated in our physical neuromorphic prototypes remain viable at the scale of millions of synthetic states and whether their collective behavior can independently identify conditions warranting human reconsideration. Modeling will progress from PyTorch/CUDA implementation of the neuromorphic mathematics through hardware-oriented architectural modeling toward FPGA prototyping and eventual neuroCMOS/FPGA implementation. Success will establish neuromorphic metacognition as a new direction in unconventional computing and demonstrate whether a brain-inspired developmental architecture can perform an independent metacognitive function alongside large AI systems.

Broader Impacts

The fundamental broader impact of this research is a new means of preserving human judgment as people increasingly rely upon AI for information, advice, and decision support. The proposed architecture will neither replace human judgment nor determine what a person should believe or do; instead, it will provide an independent indication that an AI interaction may warrant reconsideration. If successful, the research will establish a pathway toward an independent neuromorphic safeguard capable of complementing future generative, adaptive, and agentic AI systems while remaining computationally distinct from them. The project will also provide interdisciplinary student research opportunities spanning artificial intelligence, neuromorphic computing, neuroscience-inspired computation, and FPGA/CMOS systems. Resulting models, methods, and experimental findings will be disseminated to the scientific community, supporting future physical implementation and further research into human-centered neuromorphic oversight of AI.