Phillip H. Kieval

Research Statement

Several fields study learning and intelligence as the ability to discover patterns in past observations that help us predict the future. But we lack certainty that the patterns observed in the past are not just chance correlations. So, what explains this ability to generalize? This ancient philosophical challenge becomes even more puzzling when we turn to *artificial* intelligences. Current AI systems can memorize arbitrary patterns in vast amounts of complex, randomly labeled data. Yet they also show abilities to use the subtle patterns they learn to make remarkably successful predictions. We observe that this ability to generalize is both impressive and unstable. For example, they sometimes fail at surprisingly simple tasks, like counting the number of R's in "strawberry." These blunders make us wonder whether we can trust AI predictions. But systems like AlphaFold2 have also solved scientific problems like predicting how proteins fold, a challenge that stumped human scientists for a century. Critics argue that AI systems are like opaque oracles that make predictions we cannot understand or explain. I argue that successful applications of AI in science reveal that modelers make interpretable design choices that constrain how models learn representations from data. I illustrate how these choices often reflect underlying regularities in the world by encoding "invariances" that preserve important properties and relations across specific changes to data. This focus on invariance has led me to connect AI's success with the notion of symmetry discussed in the philosophy of physics, applied mathematics, and causal inference. This insight motivates my broader research program, which traces how questions concerning representation and reliability that arise in scientific modeling also appear when we use AI models to understand cognition itself. What unites my work across the philosophy of science, cognitive science, and AI ethics is a commitment to understanding both artificial and natural intelligence through careful attention to the actual practices of situated inquirers, whether they are scientists using neural networks to model biological molecules, neuroscientists studying representations in the brain, or ordinary users engaging with conversational AI systems.

My dissertation comprises six interrelated chapters on the epistemology of AI in science. I extend a pragmatist, deflationary conception of representation to describe how modelers can use deep neural networks to learn about aspects of the world. I argue that deep learning models generalize well not because of abstract mathematical guarantees, but because researchers make design choices that fit the specific features of their problem. Success depends on recognizing when general modeling techniques can be appropriately applied to a particular domain. The connections I draw between AI's success and transformational invariances motivates my current research connecting philosophical work on the epistemology of applied mathematics to empirical work on the mechanistic interpretability of deep neural networks. In the short-term, I aim to revise and publish these chapters. A version of one chapter is already forthcoming in an edited volume entitled *Philosophy of Science for Machine Learning: Core Issues and New Perspectives,* edited by Juan M. Durán and Giorgia Pozzi.

Questions about representation and inductive bias in statistical modeling take on new dimensions when AI models themselves become objects of investigation in cognitive science. If neural networks succeed by learning invariances that reflect genuine regularities, what does this tell us about the representational strategies employed by biological neural networks? In my previously published work, I defend the use of deep neural networks as models of mid-level processing in specific brain areas. I am now applying my analysis of deep learning to current efforts to use large language models (LLMs) to study human cognition, evaluating such work in light of recent achievements and empirical data. I recommend increasingly explicit scrutiny of the various modeling roles that LLMs might play in developing new explanatory

Phillip H. Kieval
✉ *pkieval@ufl.edu*

theories of human cognition. Without getting clear on the precise relationship between these models and their targets, both proponents and critics of LLMs risk talking past one another. Understanding how LLMs relate to human cognition requires consensus on standards for evaluating them as cognitive models. Failure to agree on such standards threatens to stall progress in cognitive science.

My work on representational similarity has also led me to examine a more fundamental question about the format of neural representation itself. Recent neuroscientific research describes neural computation in terms of low-dimensional dynamics, giving rise to a novel conception of neural computation. I have a pair of interrelated papers that aims to clarify the explanatory status of this work. I provide a theoretical analysis and defense of the notion of computation through neural dynamics. I argue that dynamical computation appropriates the conceptual resources of dynamical systems theory to describe a distinctive kind of abstract computation involving continuous dynamics for which there is no well-defined way of breaking down transformations into discrete computational steps. In both papers, I call for greater connection between theoretical and statistical analyses of neural computations and interventionist methods such as optogenetics in neuroscience and mechanistic interpretability studies in artificial neural networks.

The representational questions that drive my work in philosophy of science and cognitive science extend into normative territory. When applied to affective and social cognition, the inflationary assumptions about representation that I critique elsewhere will distort how we design AI systems intended for social interaction. In a paper co-authored with Cameron Buckner, we argue that prevailing approaches to artificial empathy in LLMs suffer from a spectatorial conception of social cognition. This approach sees accurate representation and reliable prediction as the primary aim of mental state attribution. But empathetic engagements involve a normative dimension; we judge our emotional responses as more or less appropriate, and we regulate each other's behavior in light of these attitudes. Drawing on ecological psychology and moral sentimentalism, this paper redefines the agenda for artificial empathy research in AI agents. If we treat artificial empathy merely as a problem of creating a more satisfying chatbot, we face the danger of offering users "social junk food." Much like hyperpalatable foods developed by nutrition scientists to hijack our food preference systems, these social interactions will be much more satisfying to users in the short term but offer little in the way of genuine emotional nourishment. Over extended periods, these shallow interactions risk social and moral deskilling that may result in long-term, societal consequences for mental health and moral development. Our perspective suggests a need for fundamental shifts in LLM agent architectures and training regimes to create the sort of virtue-promoting interactions that are needed in future waves of research.

Looking ahead, I plan to develop the arguments sketched in my recent co-authored commentary on LLMs as human surrogates in social scientific research. This work will extend my core concerns about representation and induction into the high-stakes domain of social science research where AI models are used to stand in for human participants. The fundamental question of when we can trust an AI model's capacity to generalize looms large when researchers use LLMs to predict human behavior in novel experimental contexts or to generate synthetic data for social scientific studies. This research direction promises to bring together several threads from my previous work: the mechanistic interpretability methods I advocate for understanding neural networks, my pragmatist analysis of representation, and my work on the inductive reliability of neural networks. If we cannot reliably distinguish between mere statistical pattern matching and genuine process-level similarity in AI models of human behavior, we risk building social scientific theories on fundamentally unstable foundations. By developing clearer criteria for when AI models can legitimately serve as human proxies, this work aims to ensure that the computational tools in social science enhance rather than undermine our understanding of human cognition and behavior.

Phillip H. Kieval
✉ *pkieval@ufl.edu*