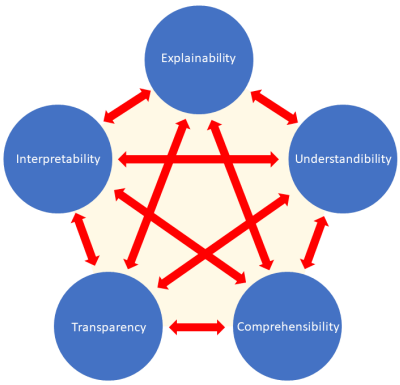
institute for pure & applied mathematics

Foundations of Interpretability



August 31 - September 4, 2026

Scientific Overview

Neural network capabilities have advanced dramatically in recent years, but it remains unclear why or how they work. The field of interpretability has sought to address these questions since the early AI (e.g., expert systems in the 80's), building inherently interpretable models in 2000's and more recently mechanistic interpretability that aims to "reverse engineer" the neural net "black boxes" and explain how they give rise to the network's behavior. Despite some success at this, the field also learned that human bias can blur our judgement regarding the effectiveness; an un-trained model can produce visually compelling explanations that is ultimately uninformative. Many open questions in foundational theory as well as about mathematical properties of neural networks remain unanswered.

The goal of this workshop is to bring together researchers in mathematics, theoretical computer science, physics, machine learning and human computer interaction working on aspects of interpretability that can be analyzed from a theoretical and practical perspective. This could involve mathematical analysis of models and techniques; modeling of phenomena and structures that arise in neural networks, such as algorithmic structures; questions about the limits of interpretability; proposed formalizations of the goals of the field and its implications for robustness and safety; and practical considerations of interpretability in a world where models are driving profound societal changes for both technical experts and the general public.

This workshop will include a poster session; a request for posters will be sent to registered participants in advance of the workshop.

Participation

Additional information about this workshop including links to register and to apply for funding, can be found on the webpage listed below. Encouraging the careers of women and minority mathematicians and scientists is an important component of IPAM's mission, and we welcometheir applications.

Organizers

Been Kim (Google DeepMind)
Marina Meila (University of Waterloo)
Alexander Oldenziel (Coefficient Giving)
Terence Tao (UCLA)

Speakers

Elias Bareinboim (Columbia University)
Ryan Cotterell (ETH Zurich)
Jacob Hilton (Alignment Research Center)
Been Kim (Google DeepMind)
Gitta Kutyniok (Ludwig-Maximilians-Universität München)
Marina Meila (University of Waterloo)
Alexander Oldenziel (Coefficient Giving)
Christopher Potts (Stanford University)
Mahdi Soltanolkotabi (USC)
Terence Tao (UCLA)
Dmitry Vaintrob (Principles of Intelligence)
Yusu Wang (UCSD)

NSF UCLA

For more information, visit the program webpage:
www.ipam.ucla.edu/IML2026