

Foundations of Interpretability

Monday August 31, 2026

- 8:00–8:55 *Check-in/Breakfast (hosted by IPAM)*
- 8:55–9:00 *Welcome and Opening Remarks*
- 9:00 *Session Chair Marina Melia*
- 9:00–9:45 **Been Kim** (Google DeepMind)
Changing Shapes: Interpretability, AI, and the Future of Human Agency
- 10:00–10:15 *Break*
- 10:15–11:00 **Jacob Hilton** (Alignment Research Center)
Miniature mathematical mysteries: uninterpreted models with fewer than 1,500 parameters
- 11:15–11:30 *Break*
- 11:30–12:15 **Paul Christiano** (Alignment Research Center)
Explaining computational phenomena
- 12:30–2:30 *Lunch (on your own)*
- 2:30 *Session Chair Been Kim*
- 2:30–3:15 **Marina Meila** (University of Waterloo)
Counterfactual Explanations in the statistical limit
- 3:30–4:00 *Break*
- 4:00–4:45 **Josh Batson** (Anthropic)
Interpretability: From Art Towards Science
- 5:00–5:20 *Lightning Poster Round*
- 5:20–6:30 *Poster Session & Reception (Hosted by IPAM)*

Tuesday September 1, 2026

- 8:00–9:00 *Check-in/Breakfast (hosted by IPAM)*
- 9:00 *Session Chair Christopher Potts*
- 9:00–9:45 **Bin Yu** (University of California, Berkeley (UC Berkeley))
Not All Interpretation Methods Are Created Equal: Principled Evaluation of AI Interpretability
- 10:00–10:15 *Break*

(Tuesday schedule continued on next page)



(Tuesday schedule continued from previous page)

- 10:15–11:00 **Max Tegmark** (Massachusetts Institute of Technology)
Neural network interpretability: symmetry, geometry and formal verification
- 11:15–11:30 *Break*
- 11:30–12:15 **Jacob Steinhardt** (University of California, Berkeley (UC Berkeley))
Oversight Foundation Models
- 12:30–2:30 *Lunch (on your own)*
- 2:30 *Session Chair Bin Yu*
- 2:30–3:15 **Ryan Cotterell** (ETH Zurich)
Two views of a language model
- 3:30–4:00 *Break*
- 4:00–4:45 **Christopher Potts** (Stanford University)
Creating and detecting subliminal learning effects on demand

Wednesday September 2, 2026

- 8:00–9:00 *Check-in/Breakfast (hosted by IPAM)*
- 9:00 *Session Chair Terry Tao*
- 9:00–9:45 **Sam Eisenstat** (Machine Intelligence Research Institute)
Concepts, information, and objectivity
- 10:00–10:15 *Break*
- 10:15–11:00 **Yusu Wang** (University of California, San Diego (UCSD))
Neural Network Generalization through an algorithmic lens
- 11:15–11:30 *Break*
- 11:30–12:15 **Michael Winer** (The Alignment Research Center)
Mechanistic Understanding of Multilayer Perceptrons
- 12:30–12:40 *Group Photo*
- 12:40–2:00 *Lunch (on your own)*
- 2:00 *Session Chair Gitta Kutyniok*
- 2:00–2:45 **Elias Bareinboim** (Columbia University)
Towards Causal AI: From Mechanism to Understanding
- 3:00–3:15 *Break*

(Wednesday schedule continued on next page)

(Wednesday schedule continued from previous page)

- 3:15–4:00 **Gal Mishne** (University of California, San Diego (UCSD))
From Explanations to Mechanisms: Interpreting Computation in Graph Neural Networks
- 4:15–4:30 *Break*
- 4:30–5:15 **Gyorgy Turan** (University of Illinois at Chicago)
Error-correcting codes: a sandbox for learning, interpretability and optimization

Thursday September 3, 2026

- 8:00–9:00 *Check-in/Breakfast (hosted by IPAM)*
- 9:00 *Session Chair Elias Bareinboim*
- 9:00–9:45 **Gitta Kutyniok** (Ludwig-Maximilians-Universität München)
From Mathematical Guarantees to Computational Limits of AI Interpretability
- 10:00–10:15 *Break*
- 10:15–11:00 **Mahdi Soltanolkotabi** (University of Southern California (USC))
Interpreting Generative Models for Better Steering: From Visual Generation to Verifiable Reasoning
- 11:15–11:30 *Break*
- 11:30–12:15 **Dmitry Vaintrob** (Principles of Intelligence, PIBBSS)
Statistical theory of learning sparse structure
- 12:30–2:30 *Lunch (on your own)*
- 2:30 *Session Chair Thurs PM TBD*
- 2:30–3:15 **Jacob Tsimerman** (University of Toronto)
Some new results in Open Source Game Theory
- 3:30–4:00 *Break*
- 4:00–5:00 *Panel Discussion - Moderator: Been Kim, Panelists: Terry Tao, Chris Potts, Thomas Icard, Eunice Jun*

Friday September 4, 2026

- 8:00–9:00 *Check-in/Breakfast (hosted by IPAM)*
- 9:00 *Session Chair Thomas Icard*
- 9:00–9:45 **Eunice Jun** (University of California, Los Angeles (UCLA))
Imagining a new future for science
- 10:00–10:15 *Break*

(Friday schedule continued on next page)

(Friday schedule continued from previous page)

10:15–11:00 **Paul Riechers** (Simplex)
The shape of beliefs and abstraction in neural networks

11:15–11:30 *Break*

11:30–12:15 **Terence Tao** (University of California, Los Angeles (UCLA))
Reflections on the workshop

