

## **Applied Physics 293 Explainable AI**

**Instructor: Surya Ganguli**

**Stanford University**

### **General review/perspective articles**

- Review articles
  - [Foundation models in neuroscience](#)
  - [A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models](#)
    - See also [ICML 2025 Tutorial on Mechanistic Interpretability for Language Models](#)
  - [Mechanistic Interpretability for AI Safety: A Review](#)
  - [Post-hoc Interpretability for Neural NLP: A Survey](#)
  - [The Shapley value in machine learning](#)
  - [The Quest for the Right Mediator: Mechanistic Interpretability via Causal Mediation Analysis](#)
  - [Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review](#)
  - [Towards Unified Attribution in Explainable AI, Data-Centric AI, and Mechanistic Interpretability](#)
  - [Explaining by removing: A unified framework for model explanation](#)
  - [Training Data Influence Analysis and Estimation: A Survey](#)
  - [A Primer on the Inner Workings of Transformer-based Language Models](#)
  - [The Explainability of Transformers: Current Status and Directions](#)
  - [Circuit analysis research landscape](#)
- Perspective pieces
  - [Position: Principles of Animal Cognition to Improve LLM Evaluations](#)
  - [Testing methods of neural systems understanding](#)
  - [Multilevel Interpretability Of Artificial Neural Networks: Leveraging Neuroscience](#)
  - [Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?](#)
  - [Assessing skeptical views of interpretability research](#)
- Roadmaps
  - [How To Become A Mechanistic Interpretability Researcher](#)
  - [Open problems in mechanistic interpretability](#)
- Paper lists
  - [List of Explainable AI papers](#)
  - [Awesome Interpretability in Large Language Models](#)
  - [Opinionated list of mechanistic interpretability papers](#)

### **Motivations: Foundation models in neuroscience: big data, big models, but understanding?**

- Task trained models in neuroscience across the years
  - [A back-propagation programmed network that simulates posterior parietal neurons](#)
  - [What Does the Retina Know about Natural Scenes?](#)
  - [The emergence of multiple retinal cell types through efficient coding of natural movies](#)
  - [Emergence of simple-cell receptive fields by learning a sparse code for natural images](#)
  - [Performance-optimized hierarchical models predict neural responses in higher visual cortex](#)
  - [Context-dependent computation by recurrent dynamics in prefrontal cortex](#)
- Complex models fit to neural data, including foundation models
  - EEG
    - [Neuro-GPT: Towards A Foundation Model for EEG](#)
  - fMRI
    - [Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data](#)
    - [BrainLM: A foundation model for brain activity recordings](#)
  - Single-cell electrophysiology
    - [Inferring single-trial neural population dynamics using sequential auto-encoders](#)
    - [Interpreting the retinal neural code for natural scenes: From computations to neurons](#)
    - [A Unified, Scalable Framework for Neural Population Decoding](#)

- [Multi-session, multi-task neural decoding from distinct cell-types and brain regions](#)
  - [Generalizable, real-time neural decoding with hybrid state-space models](#)
  - [Representation learning for neural population activity with neural data transformers](#)
  - [Neural Data Transformer 2: Multi-context Pretraining for Neural Spiking Activity](#)
  - [Towards a "universal translator" for neural dynamics at single-cell, single-spike resolution](#)
  - [Neural encoding and decoding at scale](#)
  - [Foundation model of neural activity predicts response to new stimulus types](#)
  - [Compact deep neural network models of visual cortex](#)
- Basic theories of transfer learning explaining how data from other sessions/subjects/species might help
  - [An analytic theory of generalization dynamics and transfer learning in deep linear networks](#)
  - [Features are fate: a theory of transfer learning in high-dimensional regression](#)

**Feature attribution:** How does a network output depend on input features?

- Perturbation based approaches
  - [Visualizing and Understanding Convolutional Networks](#)
  - [A Unified Approach to Interpreting Model Predictions](#)
  - [The many Shapley values for model explanation](#)
- Gradient based approaches
  - [Deep inside convolutional networks: visualizing saliency maps](#)
  - [Axiomatic Attribution for Deep Networks](#) (Integrated gradients)
  - [SmoothGrad: removing noise by adding noise](#)
  - [Time-series attribution maps with regularized contrastive learning](#)
  - [TIMING: Temporality-Aware Integrated Gradients for Time Series Explanation](#)
- Approximation based approaches
  - [Why Should I Trust You?": Explaining the Predictions of Any Classifier](#) (LIME)
  - [Significance Tests for Neural Networks](#)
- Unified view and perspectives
  - [Which Explanation Should I Choose? A Function Approximation Perspective](#)
  - [From Shapley Values to Generalized Additive Models and back](#)

**Data Attribution:** Which training data points support a test prediction?

- [Understanding Black-box Predictions via Influence Functions](#)
- [Data Shapley: Equitable Valuation of Data for Machine Learning](#)
- [Datamodels: Predicting Predictions from Training Data](#)
- Scaling up
  - [Studying Large Language Model Generalization with Influence Functions](#)
  - [TRAK: Attributing Model Behavior at Scale](#)
  - [DataInf: Efficiently Estimating Data Influence in LoRA-tuned LLMs and Diffusion Models](#)

**Discovery of Concepts**

- [Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors \(TCAV\)](#)
- [Towards automatic concept based explanations](#)
- [We Can't Understand AI Using our Existing Vocabulary](#)
- [Neural representational geometry underlies few-shot concept learning](#)
- [A mathematical theory of semantic development in deep neural networks](#)

**Introduction to Interpretability in transformers**

- Introductory articles
  - [Attention Is All You Need](#)
  - [The Illustrated Transformer](#)

- [Attention and kernel smoothing](#)
- Early Interpretation of transformers
  - [A Mathematical Framework for Transformer Circuits](#)
    - [One-layer transformers aren't equivalent to a set of skip-trigrams](#)
    - [Some common confusion about induction heads](#)
  - [In-context Learning and Induction Heads](#)
- RASP interpretation
  - [Thinking Like Transformers](#)
  - [Tracr: Compiled Transformers as a Laboratory for Interpretability](#)
- Connections to modern Hopfield model
  - [Hopfield networks is all you need](#)
  - [Dense associative memory for pattern recognition](#)
  - [On a model of associative memory with huge storage capacity](#)
  - [Exponential capacity of dense associative memories](#)
  - [The Capacity of Modern Hopfield Networks under the Data Manifold Hypothesis](#)

## **Sparse Autoencoders**

- [Towards Monosemanticity: Decomposing Language Models With Dictionary Learning](#)
- [Interpreting Attention Layer Outputs with Sparse Autoencoders](#)
- [Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control](#)
- [Interpretability Illusions with Sparse Autoencoders: Evaluating Robustness of Concepts](#)
- [The Geometry of Concepts: Sparse Autoencoder Feature Structure](#)
- [CRISP: Persistent Concept Unlearning via Sparse Autoencoders](#)
- Scaling up
  - [Scaling and evaluating sparse autoencoders](#)
  - [Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet](#)

## **Causal analysis, editing and control**

- Perturbation based approaches
  - [Direct and Indirect Effects](#)
  - [Investigating gender bias in language models using causal mediation analysis](#)
  - [Locating and Editing Factual Associations in GPT](#) (Causal tracing)
  - [How to use and interpret activation patching](#)
  - [Neuron Shapley: Discovering the Responsible Neurons](#)
- Gradient based approaches
  - [Attribution patching: Activation patching at industrial scale](#)
- Approximation based approaches
  - [Decomposing and Editing Predictions by Modeling Model Computation](#) (COAR)
- Causal abstractions
  - [Causal abstractions of neural networks](#)
  - [Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations](#)
  - [An Interpretability Illusion for Subspace Activation Patching](#)
  - [A Reply to Makelov et al. \(2023\)'s "Interpretability Illusion" Arguments](#)
  - [The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?](#)
- More model editing
  - [Editing factual knowledge in language models](#)
  - [Fast model editing at scale](#)
  - [Does localization inform editing? Surprising differences](#)
- Model steering
  - [Representation engineering: A top-down approach to AI transparency](#)
  - [The Geometry of Truth: Emergent Linear Structure](#)
  - [Truth is universal: Robust detection of lies in LLMs](#)

- [Steering Language Models With Activation Engineering](#)
- [Inference-Time Intervention: Eliciting Truthful Answers from a Language Model](#)
- [Steering Out-of-Distribution Generalization with Concept Ablation Fine-Tuning](#)

## **Evaluation of model explanations**

- [Sanity checks for saliency maps](#)
- [OpenXAI: Towards a Transparent Evaluation of Model Explanations](#)
- [MIB: A Mechanistic Interpretability Benchmark](#)
- [Towards Unifying Interpretability and Control: Evaluation via Intervention](#)
- [Causal Scrubbing: a method for rigorously testing interpretability hypotheses](#)

## **Circuit discovery**

- [Initial Circuits Thread](#)
- [Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small](#)
- [How does GPT-2 compute greater-than?](#)
- [Sparse Feature Circuits: Discovering/Editing Interpretable Causal Graphs in LLMs](#)
- [Does Circuit Analysis Interpretability Scale? Multiple Choice Capabilities in Chinchilla](#)
- [Towards Automated Circuit Discovery for Mechanistic Interpretability](#)
- [Circuit Tracing: Revealing Computational Graphs in Language Models](#)
- [On the Biology of a Large Language Model](#)
- [Transcoders Find Interpretable LLM Feature Circuits](#)
- [Circuit Tracer](#)

## **Computational complexity issues in interpretability**

- [The Computational Complexity of Circuit Discovery for Inner Interpretability](#)
- [Local vs. Global Interpretability: A Computational Complexity Perspective](#)
- [Model interpretability through the lens of computational complexity](#)

## **Comparing representations across models**

- [Similarity of Neural Network Representations Revisited](#)
- [Linearly Mapping from Image to Text Space](#)
- [The Platonic Representation Hypothesis](#)

## **Discovering and understanding interesting behaviors**

- Behavior discovery through “psychology” experiments on LLMs
  - [Language Models are Few-Shot Learners](#) (In-context learning)
  - [Language Models Don't Always Say What They Think](#) (chain-of-thought unfaithfulness)
  - [Taken out of context: On measuring situational awareness in LLMs](#)
  - [Alignment faking in large language models](#)
  - [Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs](#)
  - [Introducing Docent: A system for analyzing and intervening on agent behavior](#)
- Understanding specific, interesting behaviors
  - [On the Emergence of Linear Analogies in Word Embeddings](#)
  - [Language Models use Lookbacks to Track Beliefs](#)
  - [Language Models Share Latent Grammatical Concepts Across Diverse Languages](#)
  - [Incremental Sentence Processing Mechanisms in Autoregressive Language Models](#)
  - [Emergent World Representations: Exploring a Sequence Model on a Synthetic Task](#)
  - [Progress measures for grokking via mechanistic interpretability](#)
  - [Acquisition of chess knowledge in AlphaZero](#)

## **Cautionary tales in explainability**

- [Impossibility theorems for feature attribution](#)
- [The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective](#)
- [Faithfulness vs. Plausibility: On the \(Un\)Reliability of Explanations from Large Language Models](#)
- Adversarial attacks on Interpretations
  - [Interpretation of Neural Networks is Fragile](#)
  - [Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods](#)

## **Automated Interpretability Agents**

- [A Multimodal Automated Interpretability Agent](#)

## **Reasoning**

- [On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models](#)
- [Measuring the Faithfulness of Thinking Drafts in Large Reasoning Models](#)
- [Thought Anchors: Which LLM Reasoning Steps Matter?](#)
- [All for One: LLMs Solve Mental Math at the Last Token With Information Transferred From Other Tokens](#)
- [Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs](#)