Kai Fronsdal

kaif@stanford.edu • (650) 554-1750

Education

Stanford University
Bachelor Candidate in Mathematics June 2024
Masters Candidate in Computer Science June 2025 (expected)
GPA: 4.08

Woodside High School
Valedictorian June 2020
GPA: 4.00

Experience

ML Alignment & Theory Scholars June 2024 – Present
AI Safety Researcher Berkeley, CA

- AI agents with sophisticated self-reasoning abilities pose significant risks to AI alignment and safety as they may be able to circumvent safety measures, engage in deceptive behavior, and sandbag.
- Together with DeepMind's David Lindner, we create an open source, agentic evaluation suite to detect different kinds of instrumental self-reasoning (accepted at NeurIPS and in review at ICLR).
- Explore elicitation and control methods to prevent instrumental self-reasoning.

STAIR, Stanford University May 2023 – Present
AI Researcher Stanford, CA

- Current reasoning benchmarks rely on a single, boxable final answer to verify LLM outputs.
- We are working towards creating better metrics to measure LLM reasoning abilities that don't rely on having a single numerical answer to compare to such as proofs (accepted at NeurIPS and in review at ICLR).
- Separately, working towards the auto-formalization of informal mathematical statements in order to enhance the accuracy and reliability of automated proof checking.
- Leveraging interactive proof assistants and languages to create training environments, accelerating the model refinement process.

Supervised Program for Alignment Research June 2024 – September 2024
AI Safety Researcher Berkeley, CA

- Working with Anthropic's Nina Panickssery on using synthetic data to address model deficiencies.
- Creating simple methods to dramatically reduce sycophancy, over refusal, and other undesirable behaviors.
- Testing methods to reduce the effectiveness of many-shot jailbreak red teaming methods.

Stanford University September 2024 – Present
Teaching Assistant Stanford, CA

- Teaching assistant for CS 362: Research in AI Alignment
- Facilitate discussions about each weeks lecture and provide feedback on students weekly assignments

Stanford University May 2024 – Present
Student Lecturer Stanford, CA

- Co-creating and co-teaching Stanford AI Club's (SAIC) Intro to Applied PyTorch class at Stanford.
- ML classes at Stanford tend to focus more on theory, rather than creating practical ML projects.
- We teach students how to build an ML pipeline from the ground up with all of the messy details that tend to get skipped over in classes (like handling data).
- Focusing on a hands-on/interactive lecture format.

QuantCo June 2023 – September 2023
ML Researcher Intern Boston, MA

- Spearheaded a comprehensive evaluation of deep learning approaches to long-horizon hierarchical demand forecasting, comparing to the current solution using traditional statistical techniques such as ARIMA and gradient-boosted methods.

- Engineered a robust and reproducible experiment pipeline, streamlining model development, evaluation, and saving processes, enhancing team efficiency and knowledge management.
- Achieved a 23% reduction in mean squared error along with better handling of extreme events, and created more robust predictive models.

IRIS Lab, Stanford University March 2022 – February 2023
AI Researcher Stanford, CA

- Solving language-conditioned robotic manipulation tasks through meta-learning techniques and multi-modal models.
- Orchestrated the curation and combination of extensive robotic datasets, culminating in the creation of a large-scale robotic manipulation dataset with the long-term goal of constructing a foundation model in this regime.
- Focus on creating machine learning models capable of operating on multiple robot systems and generalizing to few-shot and zero-shot learning new tasks.

Stanford Data and Mapping for Society August 2021 – December 2022
Data Analyst Stanford, CA

- Conducted in-depth analysis and mapping of global energy flow trends, providing crucial insights into the dynamics of energy distribution worldwide.
- Tracked and visualized fossil fuel exports and imports between countries, creating intuitive and informative visuals to effectively communicate complex data.

Stanford Intelligent Systems Laboratory March 2021 – July 2021
AI Researcher Stanford, CA

- Developing a more robust adversarial reinforcement learning algorithm integrating principles from safety validation for enhanced system reliability with Mykel Kochenderfer.
- Employed deep learning techniques to accurately approximate the distribution of probable failure modes, enabling effective sampling for robust system training and evaluation.
- Measured a 2% reduction in failure modes during testing across diverse disturbance scenarios outperforming current state-of-the-art methodologies.

Afiniti June 2020 – September 2020
Applied Research Intern Washington, D.C.

- Developed causal machine learning algorithms enabling precise estimation of the heterogeneous treatment effect for call center outcomes.
- Conducted comprehensive evaluation and visualization of hyperparameter combinations for R-learners, optimizing model performance and enhancing predictive accuracy.
- Designed and evaluated strategies for the next-generation pairing algorithms, leveraging predictive modeling of future events to drive innovation in customer-agent pairings.

Team 100 Robotics October 2016 – June 2020
Head of Vision Processing Woodside, CA

- Held a key leadership role within the FRC Team 100 Robotics, contributing to the strategic decision-making process and fostering a culture of excellence.
- Spearheaded the prototyping and design efforts for competition mechanisms, ensuring precision and innovation in the face of challenging constraints.
- Led the development of vision processing techniques and real-time path planning for autonomous robot control.

Relevant Courses

- Deep Learning (CS 230), Machine Learning (CS229), NLP (CS 224N), Deep Multi-task and Meta-Learning (CS 330), Decision Making Under Uncertainty (CS 238), Sequence Modeling (CS 229B), Data for Sustainable Development (CS 325B)
- Probability Theory (MATH 230A, MATH 230B), Mathematics of AI (MATH 275C), Algebraic Topology (MATH 215A), Mathematical Problems in Machine Learning (MATH 276)
- Statistical Inference (STATS 200), Applied Statistics (STATS 305A), Nonparametric Statistics (STATS 205)
- Design and Analysis of Algorithms (CS 161), Optimization (CS 261), Randomized Algorithms and Probabilistic Analysis (CS 265), Data Structures (CS 166), Modern Algorithmic Toolbox (CS 168), Theory of Computation (CS 154), Convex Optimization (EE 364A)
- Computer Systems (CS 106B, CS 107, CS 111), Parallel Computing (CS 149)

- Aligning Superintelligence (MS&E 338), Algorithmic Fairness (CS 256), AI Alignment (STS 10SI), Building Trust in Autonomy (AA 120Q)

Technical Skills

Languages: Python, Julia, R, C++, C
Frameworks: PyTorch, Huggingface Ecosystem, Pandas, Dask, Sklearn, Numpy, Einops, Statsmodels, Matplotlib

Other Projects

Open Source Contributions: Inspect (agentic evals), sglang (LLM serving framework), Pandas (data manipulation), einops (readable tensor operations), Crux.jl (deep reinforcement learning library), Tabmat (efficient matrix representations for tabular data)

A Mathematical Framework of Goodharts Law (MS&E 338): Characterised Goodharts Law and its categories through dynamics of causal models. Using this framework, I was able to prove under what conditions some variants of Goodharts Law will occur and bound how bad they can be.

Can AI Self-Correct? (CS 224N): Analyze the capabilities of models to find errors in their own reasoning chains. We found that current SOTA LLMs struggle to correct calculation errors much more than general reasoning errors indicating a potential avenue for making LLMs more robust.

Wildfire Mapping (CS 325B): Creating state-of-the-art scalable segmentation models using satellite imagery to map out burned areas. Burned areas are a critical component to many climate models as well as disaster prepardness and response.

Transformers for Time Series Modeling (CS 229B): Recently there has been a large debate over the efficacy of transformer based methods for time series modelling. This project showed that transformers are better able to capture covariate dependencies compared to other state-of-the-art methods, especially in the low data regime.

Autotune: Automatic Surgical Fine-Tuning (CS 330): Building on the work of surgical fine-tuning, we develop a method based on multi-armed bandits to automatically fine-tune the layers of any neural network to match the performance of the best surgical fine-tuning.

Large Scale Behavior Cloning: Creation and collation of a robotic manipulation dataset to take advantage of high-capacity models. Design and train a general representation model for robotic manipulation tasks to reduce train time/training samples required and improve generalization for downstream tasks.

Predicting Chaotic Electrical Responses (CS 229): Exploring approaches such as reservoir computing and RNNs to predict measurably chaotic time series data in axons.

Generic Tree Augmentations to Speed up Maximum Flow (CS 166): An analysis and pedagogic explanation of the Top Tree data structure for creating a generic tree augmentation with an application to the maxflow algorithm and a discussion on possible routes of generalization to any graph.

DOLFiN: A more robust approach to adversarial reinforcement learning through approximation of and sampling from the distribution of likely failure modes.

Deep Learning Approaches for Predicting Drug Mechanisms of Action (CS 230): Discovering a drug's mechanisms of action (MOAs) is a critical first stage in the drug discovery process; however, this process is typically costly and time-intensive. We explored various approaches to use deep learning to predict MOAs taking advantage of correlations in the data.