Dr. Alex J. Mercer

Research Statement

Department of Computer Science · Stanford University
amercer@stanford.edu · alexmercer.github.io · +1 (650) 555-0199

Research Vision: My research lies at the intersection of deep learning theory and robust systems engineering. Specifically, I design algorithms that bridge the gap between theoretical guarantees in machine learning and the unpredictable reality of open-world deployments. Over the next five years, my goal is to build **provably safe, resource-efficient, and self-adaptive autonomous systems** that can safely interact with dynamic environments.

1 Overview

Modern artificial intelligence has achieved remarkable success in controlled laboratory settings. However, when deployed in high-stakes, real-world environments—such as autonomous driving, healthcare, and robotics—these systems frequently fail due to distribution shifts, adversarial inputs, and hardware constraints.

My research program addresses these vulnerabilities through a two-pronged approach:

1. **Robustness and Safety Foundations:** Developing mathematical frameworks that provide formal out-of-distribution (OOD) generalization guarantees.
2. **Systems-Aware Deep Learning:** Co-designing neural architectures and training algorithms that optimize latency and energy efficiency on edge hardware without sacrificing performance.

To date, my research has led to 8 publications in top-tier venues (NeurIPS, ICML, CVPR, and ICLR), including one Outstanding Paper Award. My open-source libraries have been adopted by over 50 academic labs and industrial research teams.

2 Past & Current Research

My work is characterized by bridging theoretical machine learning principles with practical systems engineering.

2.1 Robustness under Distribution Shift

Standard machine learning assumes that training and testing data are drawn from the same distribution. In practice, this assumption is regularly violated. To tackle this, I developed a framework for *Invariant Feature Attribution* [1]. By constraining neural networks to learn causal features rather than spurious correlations, we improved out-of-distribution accuracy by 14% on safety-critical computer vision benchmarks. This work mathematically formalizes the connection between gradient-based attribution methods and causal invariance, providing the first generalization bounds under covariate shift.

2.2 Computational Efficiency for Vision-Language Models

Large-scale vision-language models (VLMs) offer rich representations but are computationally prohibitive for edge devices. In [2], I introduced *SparseFlow*, a dynamic structural pruning algorithm that compresses VLMs during runtime. Unlike static pruning, which permanently removes parameters, SparseFlow selectively activates sub-networks depending on the input complexity. This achieves a $3.5\times$ reduction in computational latency and a 60% reduction in memory footprint, enabling real-time deployment on standard mobile processors while retaining 98.5% of the original model's accuracy.

3 Research Agenda & Future Directions

As a faculty member, I will build an interdisciplinary research group focusing on the lifecycle of trustworthy machine learning, spanning from mathematical proofs to deployment.

3.1 Short-Term Focus (Years 1–3): Provably Safe Foundation Models

As foundation models are increasingly integrated into interactive environments (e.g., LLM agents in robotics), ensuring safety is paramount. I plan to develop *Neural Guardrails*—a framework that wraps around pre-trained foundation models and monitors their outputs in real-time. By utilizing conformal prediction and runtime verification, these guardrails will guarantee, with user-defined probability, that the system's actions remain

within a safe state space. This project will be supported by collaborations with roboticists and control theorists to translate safety bounds to physical actuators.

3.2 Long-Term Vision (Years 4+): Autonomic Self-Healing ML Systems

The ultimate goal of my research is to create machine learning systems that can continuously adapt to new environments without human intervention. Standard fine-tuning often leads to catastrophic forgetting. I will investigate methods for *Continual Modular Learning*, where the model dynamically allocates new sub-networks to represent new concepts while freezing existing modules. By formulating this as a Bayesian task-routing problem, we aim to build systems that automatically detect when their performance degrades, isolate the failure mode, and patch their own weights using streaming data.

4 Broader Impacts

4.1 Education, Mentorship, and Diversity

I am deeply committed to broadening participation in computer science. Throughout my PhD and postdoc, I have mentored 5 undergraduate students (including 3 from underrepresented backgrounds), resulting in two co-authored workshop papers. As a faculty member, I will design a new curriculum on *Trustworthy Systems-Aware Machine Learning* to teach students how to balance algorithmic efficiency with societal safety. I will also partner with local high schools to run summer workshops introducing AI ethics and coding to underrepresented groups.

4.2 Open-Source Software and Industrial Collaboration

Democratizing research is key to accelerating scientific progress. I actively maintain the `RobustTorch` library, which simplifies the evaluation of model robustness under distribution shift. I will continue to foster collaborations with industry partners (having previously worked with researchers at Nexa DeepMind and Quanta) to ensure that the theoretical frameworks developed in my lab address real-world engineering bottlenecks and are quickly translated into production.

5 Selected Publications

[1] **Alex J. Mercer**, Yu-An Chen, and Li-Min Tseng. "Provably Robust Foundation Models via Neural Guardrails." *International Conference on Machine Learning* (*ICML*), 2025.
Introduced a conformal-prediction framework to guarantee safety bounds in foundation models. **[Outstanding Paper Award]**

[2] **Alex J. Mercer** and Sarah Jenkins. "SparseFlow: Dynamic Pruning for Vision-Language Models." *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2024.
Designed a dynamic sub-network activation scheme for edge computing, achieving $3.5\times$ *latency reduction.*

[3] **Alex J. Mercer**, Marcus Vance, and David K. Harrison. "Invariant Feature Attribution for Causal Out-of-Distribution Generalization." *Advances in Neural Information Processing Systems* (*NeurIPS*), 2023.
Proposed causal invariance constraints to improve neural network generalization under covariate shifts.

[4] **Alex J. Mercer** and David K. Harrison. "Quantifying and Mitigating Spurious Correlations in Deep Classifiers." *International Conference on Learning Representations* (*ICLR*), 2022.
Identified structural vulnerabilities in deep vision classifiers and proposed adversarial training mitigations.