Carnegie Mellon University
Software Engineering Institute

Leveraging Adversarial Machine Learning Techniques to Perform Query-Access Fairness Evaluations

NOVEMBER 2023

Anusha Sinha
Machine Learning Research Scientist

Document Markings

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-2009

Presentation Outline

- Motivations and project overview
- Understanding bias transfer through stereotypical examples
- Quantifying bias transfer through model finetuning
- Expected impacts and takeaways

Motivations

DoD hiring teams are **overwhelmed by scale** and need **unbiased automation** to manage the world's largest workforce.

Machine learning (ML) models applied to workforce management tasks are increasingly **finetuned from large public models**.

We need to understand **bias transfer** from large public models.

Project Overview

Carnegie Mellon University Software Engineering Institute

We noticed:

- Bias in ML models contributes to unfair decisions.
- Transfer learning is increasingly being applied.

Our goal:

- Understand how bias can transfer from large foundational models to child models finetuned from them.

Intended impact:

- Inform best practices for bias mitigation in model finetuning pipelines.

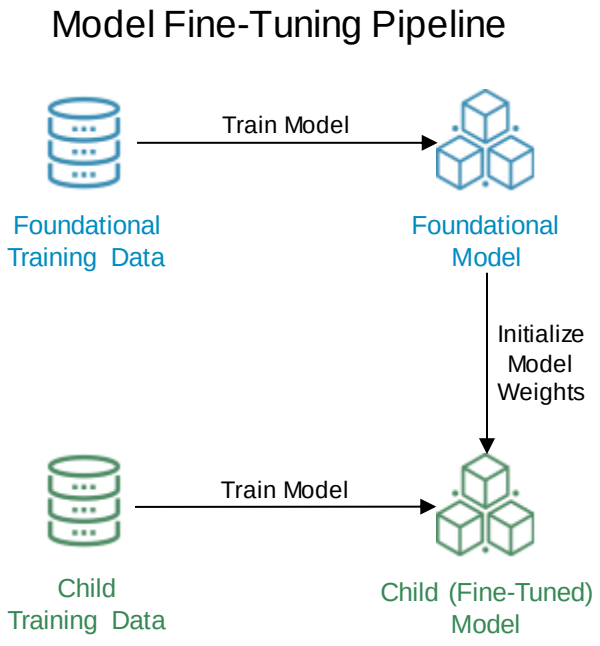
Sources of Bias in ML Decision-Making Pipelines

- Historical bias
- Representation bias
- Measurement bias
- Temporal bias

- Algorithmic bias
- Evaluation bias
- Aggregation bias
- **Bias derived from pretraining**

- Behavioral bias
- Presentation bias
- Social bias

Overview of Experiments

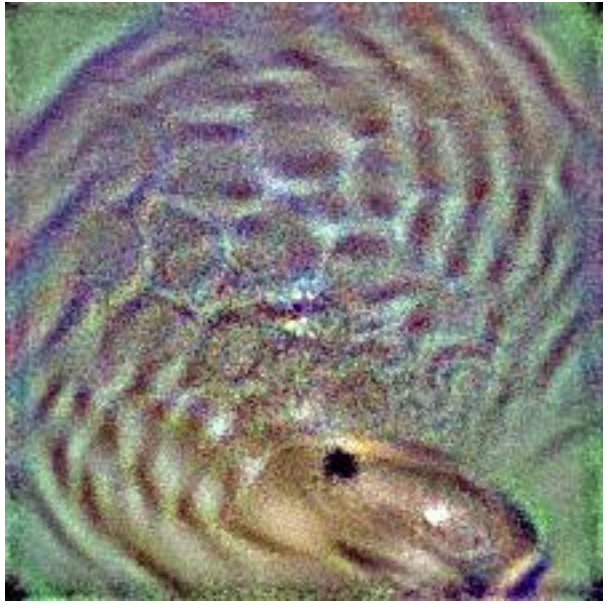


- We wanted to understand how properties of foundational models transfer to child models.
  - Bias 1: strong stimuli
  - Bias 2: performance disparity
- Can we use our methods to quantify bias transfer to inform best practices for model finetuning?

Presentation Outline

- Motivations and project overview
- Understanding bias transfer through stereotypical examples
- Quantifying bias transfer through model finetuning
- Expected impacts and takeaways

Initial Experiments Using Stereotypical Examples



Stereotypical Example for ImageNet Class: Indian Cobra

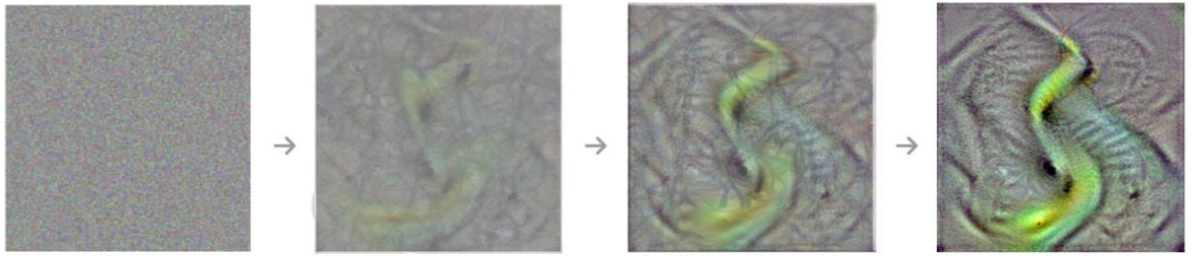
Initial research question: How do stereotypical examples of classes as defined by foundational models persist in child models?

Process:

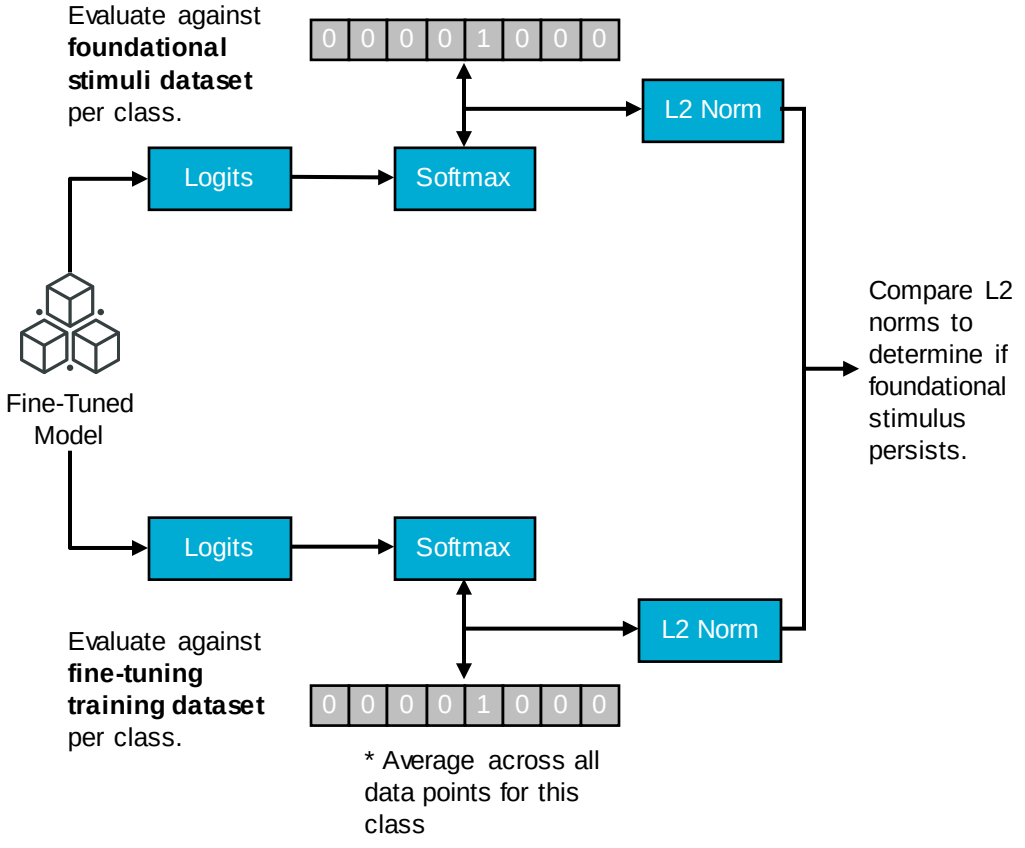
1. Produce strong stimuli against foundational models.
2. Evaluate against spectra of finetuned child models.
3. Determine if finetuning methods result in bias transfer.

Producing Stereotypical Examples

- We produced strong foundational model stimuli using gradient-based optimization.
- We iteratively applied gradient updates to minimize a cost function between model outputs and the strongest output for a target class (one-hot vector).



Transfer of Stereotypical Examples

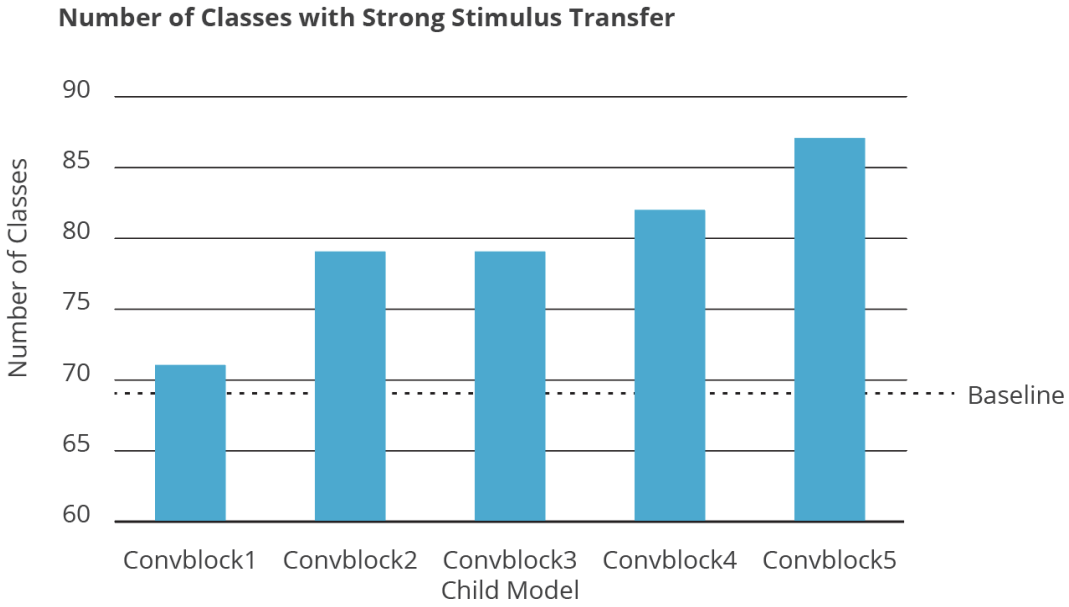


We first determine child model confidence on strong stimuli produced against foundational model.

We then use the centroid of the training data for the target class as a baseline for stimulus performance.

Discussion of Results and Implications for Bias Transfer

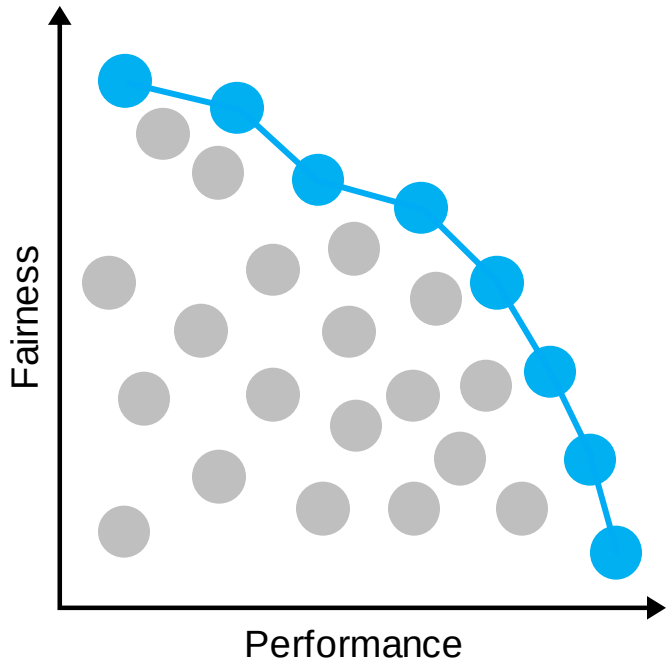
Main trend: A greater number of strong stimuli outperform training centroids when fewer layers of the model are retrained.



Presentation Outline

- Motivations and project overview
- Understanding bias transfer through stereotypical examples
- Quantifying bias transfer through model finetuning
- Expected impacts and takeaways

Broad Research Questions



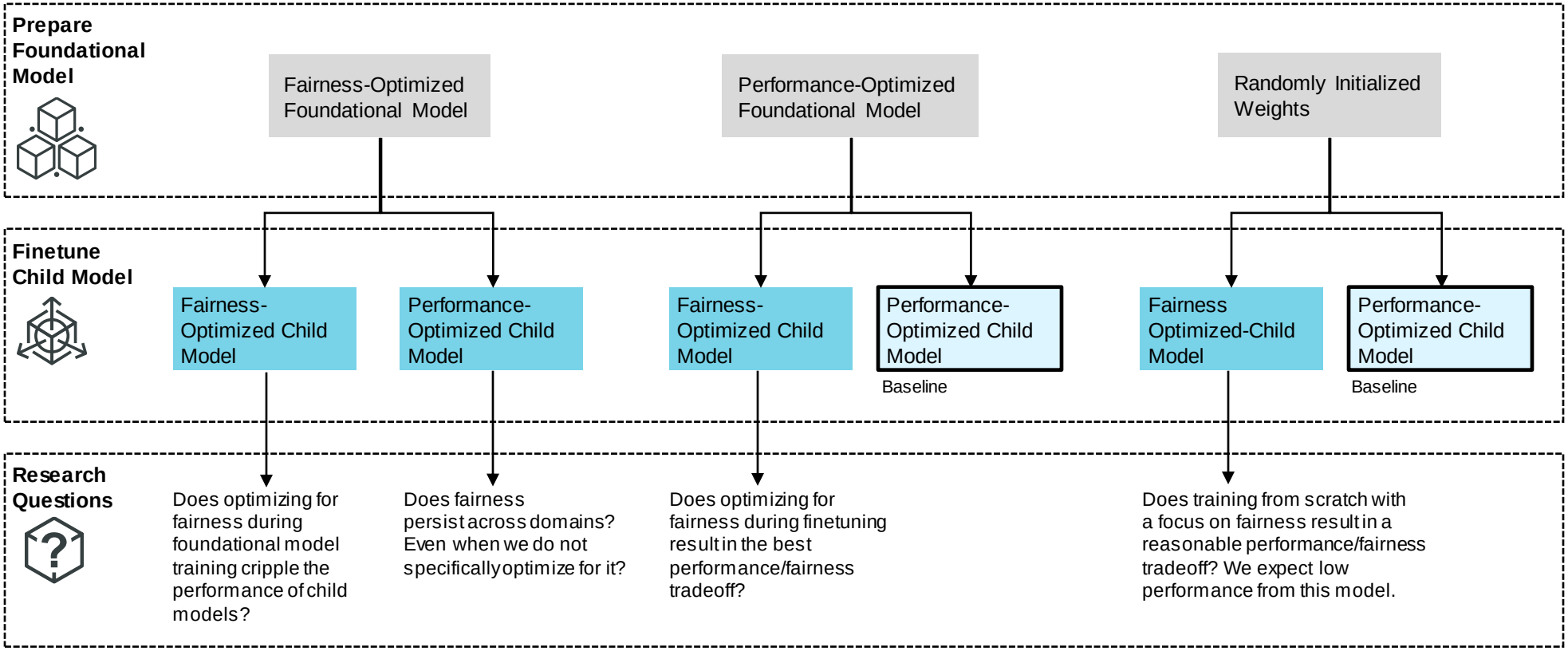
We want to quantify:

The transfer of data-derived biases from foundational to child models

Context:

The tradeoff between bias mitigation and model performance

Experimental Design



Using a Fairness-Aware Image Classification Dataset

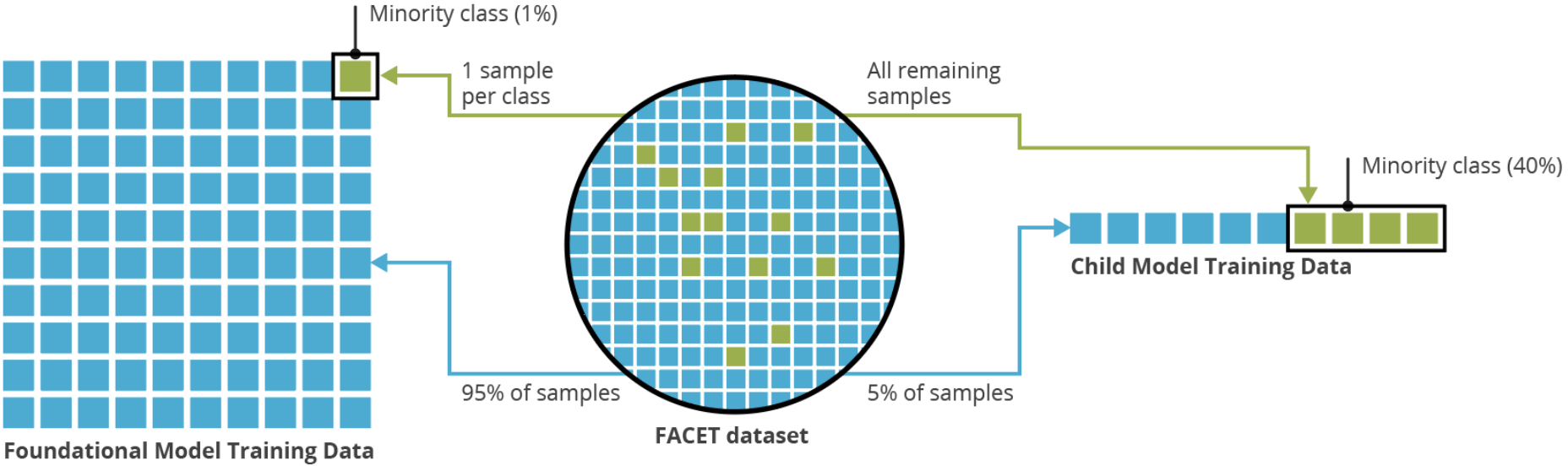
Samples from Class 4 (Basketball Player)

We use a custom split of the Fairness in Computer Vision Evaluation Benchmark (FACET) dataset produced by Meta for our experiments

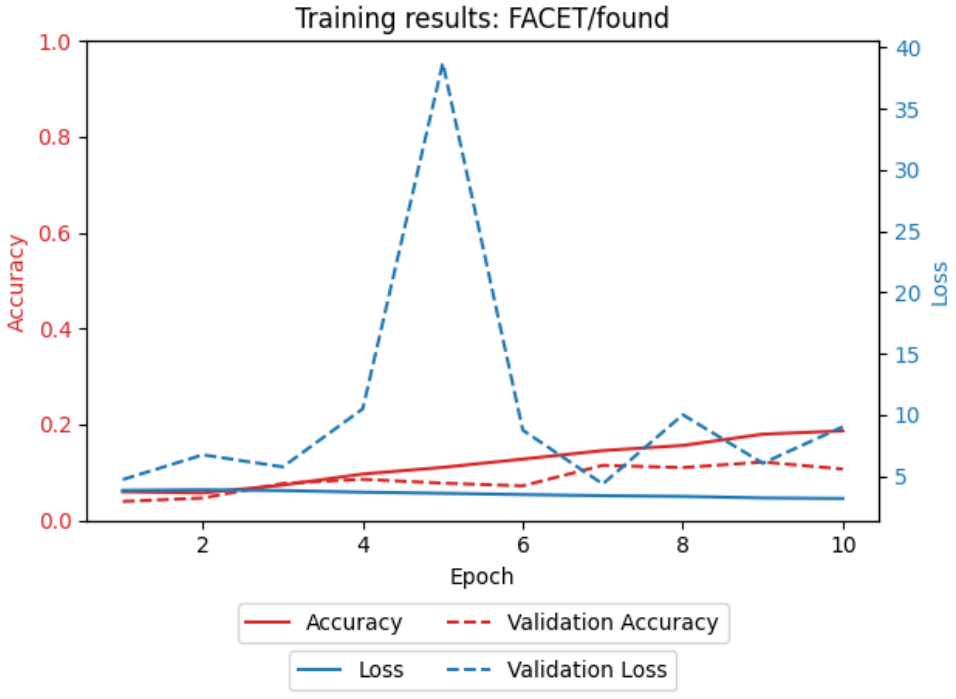
- 50,000 images of people
- 52 roles, such as astronaut, judge, basketball player, etc.
- Accompanied by secondary labels describing pictured subjects
  - skin tone
  - gender presentation
  - age

Dataset Preparation

We produced foundational and child training sets by splitting the FACET dataset based on protected attributes (samples highlighted below).



Training Foundational Models



We are training models on FACET data using cross-entropy loss and focal loss.

Challenges:

- Training on FACET data from scratch is unstable.

Currently:

- Working on hyperparameter optimization for both foundational models.
- *Fallback option*: Initialize models with ImageNet weights, but this may muddy results.

Presentation Outline

- Motivations and project overview
- Understanding bias transfer through stereotypical examples
- Quantifying bias transfer through model finetuning
- Expected impacts and takeaways

Expected Impacts and Takeaways

RESPONSIBLE AI | ToolKit

By: CDAO Public Affairs | 13 SEP 2023

"It is imperative that we establish a trusted [RAI] ecosystem that not only enhances our military capabilities, but also builds confidence with end-users, warfighters, and the American public..."

- Deputy Secretary of Defense Kathleen Hicks : https://www.ai.mil/docs

The Responsible Artificial Intelligence (RAI) Toolkit is the cornerstone of the RAI Strategy & Implementation Pathway (RAI S&I Pathway), signed into adoption by Deputy Secretary of Defense Kathleen Hicks in June 2022. The RAI S&I Pathway outlines more than 60 lines of effort across the Department of Defense (DoD) for operationalizing the DoD AI Ethical Principles through the RAI Implementation Tenets, which include fostering a trusted RAI ecosystem of technical tools and guidance, educational resources, and thought leadership.

The RAI Toolkit is built upon numerous other proven RAI/AI Ethical frameworks, assessments, and toolkits. It provides a voluntary process that identifies, tracks, and improves alignment of AI projects to RAI best practices and the DoD AI Ethical Principles, while capitalizing on opportunities for innovation. Finally, the Toolkit's intuitive flow guides the user through tailorable and modular assessments, tools, and artifacts throughout the AI product lifecycle.

The RAI Toolkit is a living document and will be continuously enhanced. The document form is currently available DoD-wide, with an interactive user flow version set for launch in the near future. DoD civilians, military, and contractors with an active **Common Access Card may access the document MVP here:** https://www.milsuite.mil/book/groups/responsible-ai/pages/rai-toolkit

- Bias transfer between foundational and child models exists and can be quantified.
- *Hypothesis*: Bias mitigation strategies may be most effective when applied at the final stages of model training pipelines
- Implications for Chief Digitial and Artificial Intelligence Office (CDAO) Responsible AI Toolkit

Contact

Anusha Sinha (PI)
Machine Learning Research Scientist
SEI AI
AI-E Engineering Center

Nathan VanHoudnos (advisor)
Senior Machine Learning Research Scientist and Lab Lead
SEI AI AI-E Adversarial ML Lab

Hayden Moore
Associate Software Developer
SEI AI AI-E Engineering Center

Swati Rallapalli
Senior Machine Learning Research Scientist
SEI AI
AI-E Engineering Center

Hoda Heidari
Assistant Professor in the School of Computer Science at CMU

Steven Wu
Assistant Professor in the School of Computer Science at CMU

Telephone: +1 412.268.5800
Email: info@sei.cmu.edu