Alex Oesterling

aoesterling@g.harvard.edu, alexoesterling.com

EDUCATION

Harvard University 2022 – Present
Ph.D. in Computer Science
Advised by Prof. Flavio du Pin Calmon, Prof. Hima Lakkaraju

Duke University 2018 – 2022
B.S.E. in Electrical and Computer Engineering, Computer Science GPA: 3.98
Summa Cum Laude
Advised by Prof. Guillermo Sapiro, Prof. Cynthia Rudin

FELLOWSHIPS & AWARDS

National Science Foundation Graduate Research Fellowship, 2022
Funding for three years of graduate studies, covering cost of tuition and stipend
Duke George Sherrerd III Memorial Award, 2022
Awarded annually by ECE faculty to the senior with the highest scholastic achievement and service to the community.
Tau Beta Pi, 2021
National engineering honor society member with high academic standing and exemplary character.
IEEE Eta Kappa Nu, 2021
IEEE honor society for electrical engineers with high academic standing and leadership potential.
Pratt Fellowship, 2020
Competitive research fellowship culminating in graduation with distinction.
Huang Fellowship, 2019
Competitive fellowship focused on science and its intersection with society.

WORK EXPERIENCE

Apple AI/ML Research 05/2025 - 08/2025
Research Scientist Intern
Working with Apple's Human Centered Machine Intelligence team to develop AI Interpretability tools to improve safety of Apple Intelligence

Amazon Alexa 06/2021 - 08/2021
Software Development Engineering Intern
Improved ambient Alexa experience by introducing new use case for customers owning multiple Alexa devices including one device with a screen

PUBLICATIONS

"Leveraging the Sequential Nature of Language for Interpretability"
A. Oesterling*, U. Bhalla*, C.M. Verdun, F.P. Calmon, H. Lakkaraju.
Oral*, Workshop on Assessing World Models, ICML, 2025*

"Inference Time Reward Hacking in Large Language Models"
H. Khalaf, C.M. Verdun, **A. Oesterling**, H. Lakkaraju, F.P. Calmon
Spotlight*, NeurIPS 2025*
Prev. at Workshop on Models of Human Feedback for AI Alignment, ICML, 2025

"Soft Best-of-n for Model Alignment"
A. Oesterling*, C.M. Verdun*, H. Lakkaraju, F.P. Calmon
ISIT 2025

“Multi-group Proportional Representations in Text-to-Image Models.”
S. Jung, **A. Oesterling**, C.M. Verdun, S. Vithana, F.P.Calmon.
CVPR 2025
Prev. at Algorithmic Fairness through the Lens of Metrics and Evaluation Workshop, NeurIPS, 2024

“Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE).”
A. Oesterling*, U. Bhalla*, S. Srinivas, F.P. Calmon, H. Lakkaraju.
NeurIPS, 2024

“Multi-Group Proportional Representation in Retrieval.”
A. Oesterling*, C.M. Verdun*, C.X. Long, A. Glynn, L.M. Paes, S. Vithana, M. Cardone, F.P. Calmon.
NeurIPS, 2024

“Fair Machine Unlearning: Mitigating Disparities during Data Deletion.”
A. Oesterling, J. Ma, F. P. Calmon, H. Lakkaraju.
AISTATS 2024
Prev. at Data Centric Machine Learning Workshop, ICML, 2023

“Operationalizing the Blueprint for an AI Bill of Rights: Recommendations for Practitioners, Researchers, and Policy Makers.”
A. Oesterling*, U. Bhalla*, S. Venkatasubramanian, H. Lakkaraju
Preprint

“All Roads Lead to Rome? Exploring Representational Similarities Between Latent Spaces of Generative Image Models”.
C. Badrinath, U. Bhalla, **A. Oesterling**, S. Srinivas, H. Lakkaraju.
HiLD, GRaM, SPIGM Workshops, ICML, 2024

“Multitask Learning for Citation Purpose Classification.”
A. Oesterling*, A. Ghosal*, H. Yu*, R. Xin*, Y. Baig*, L. Semenova, C. Rudin.
Second Workshop on Scholarly Document Processing (SDP), NAACL, 2021

“Distributionally Robust Group Backwards Compatibility.”
M. Bertran, N. Martinez, **A. Oesterling**, and G. Sapiro.
DistShift Workshop, NeurIPS, 2021

“Detecting Motion in a Room Using a Dynamic Metasurface Antenna.”
A. Oesterling, M. Imani, O. Mizrahi, J. Gollub, and D. Smith.
IEEE Access, 2020

“Integrated Single-cell Multiomic Analysis of HIV Latency Reversal Reveals Novel Regulators of Viral Reactivation.”
A. Manickam, J. Peterson, Y. Harigaya, D. Murdoch, D. Margolis, **A. Oesterling**, Z. Guo, C. Rudin, Y. Jiang, E. Browne.
Genomics, Proteomics & Bioinformatics, 2024

“PP 1.33 – 00167 Integrated single-cell multiomic profiling of HIV latency reversal.”
A. Manickam, J. Peterson, W. Mei, D. Murdoch, D. Margolis, **A. Oesterling**, Z. Guo, C. Rudin, Y. Jiang, E. Browne.
Journal of Virus Eradication, 2022

** denotes equal contribution*

INVITED TALKS

Workshop on Assessing World Models, ICML 7/2025

Algorithmic Alignment Lab, MIT 12/2024

TEACHING

Teaching Fellow, *CS 2901/2: Seminar on Effective Research Practices and Academic Culture* Harvard, Fall 2025, Spring 2026

Teaching Fellow, *ES 156: Signals and Communications* Harvard, Spring 2024

Head Teaching Assistant, *First Year Design (EGR 101)* Duke, Fall 2020-2021

Teaching Assistant, *First Year Design (EGR 101)* Duke, Fall 2019

SERVICE

Program Chair
RegML @ NeurIPS 2024

Area Chair
RegML @ NeurIPS 2025

Conference Reviewer
NeurIPS, ICML, ICLR, FAccT

Top Reviewer Award
Complimentary Registration, NeurIPS 2025

Journal Reviewer
IEEE Journal on Selected Areas in Information Theory

LEADERSHIP

Brownstone Residential Living Group 01/2019 – 05/2022
President, Historian

Duke Pureun KPop Dance Team 09/2019 – 05/2022
Social Chair and Small Group Leader