Joshua Nathaniel Williams

⚑ US Citizen ✉ jnwillia@andrew.cmu.edu joshua-n-williams jnwilliams

EDUCATION

Carnegie Mellon University **August 2018 - August 2025**
Doctorate of Compute Science (PhD) *Pittsburgh, PA*

Hampton University **August 2012 - May 2016**
Bachelor of Science in Mathematics *Hampton, Virginia*
Summa Cum Laude

SKILLS

- **Programming Languages**: Python, Matlab
- **Machine Learning Frameworks**: PyTorch, Scikit-learn
- **Tools**: Git, Unix
- **Technical Specialities:** Generative Modeling, Model Evaluation & Red-Teaming, Large Language Models (LLMs), Explainable AI (XAI), AI Safety & Alignment, Data Crowdsourcing

DISSERTATION RESEARCH

Carnegie Mellon University **August 2025**
Advised by: Zico Kolter *Pittsburgh, PA*

My research focuses on explainability methods for generative models. I develop and analyze algorithms that identify prompts capable of reproducing a given image using a specified image generator. These discovered prompts offer valuable insights into the behavior and decision-making processes of the analyzed models.

WORK EXPERIENCE

Student Researcher **August 2023 - March 2024**
Google *Remote*

- Developed and tested methodologies to understand the impact of dialect variations on generative image modeling.
- Created a custom dataset of dialect-based image prompts hand-derived from internal data sources.
- Built a Python-based tool to efficiently crowdsource image labels, enabling broad analysis of dialect on generated data.

Summer Associate - Adjunct Staff **June 2023 - August 2023**
RAND Corporation *Pittsburgh, PA*

- Developed protocols for integrating machine learning into Air Force human resource management systems.
- Analyzed several classes of ML models to identify potential risks and additional considerations in AI-driven HR processes.
- Presented key findings and recommendations to senior Air Force leadership, influencing strategic decision-making.

Freelance **June 2021 - August 2021**
American Civil Liberties Union *Pittsburgh, PA*

- Collaborated with stakeholders to refine data interpretation and support policy recommendations.
- Analyzed judicial bail data for a statewide report on pretrial release decisions, identifying trends and disparities.
- Reviewed student in-school arrest data to assess patterns contributing to the school-to-prison pipeline.

Post-Baccalaureate Researcher **September 2016 - June 2018**
University of California Irvine - Beckman Laser Institute *Irvine, CA*

- Designed algorithms for processing and analyzing multiphoton microscopy images to study skin structures.
- Wrote MATLAB-based neural networks for detecting and classifying structures within dermatological images.
- Created computational methods to quantify collagen fiber orientation and assess skin abnormalities for clinical applications.

CONFERENCE & WORKSHOP ORGANIZATION

Workshop on Responsible AI **May 2021**
International Conference on Learning Representations (ICLR) *Virtual*

- Organized workshop paper submission process, recruited paper reviewers and area chairs.
- Facilitated virtual poster session and spotlight talks for accepted papers.

Workshop on AI-Based Policing **Feb 2020 & Feb 2021**
Pittsburgh Racial Justice Summit *Pittsburgh, PA*

- Developed presentations and activities on AI-based policing solutions for non-technical audiences.

SELECTED PUBLICATIONS

Proposed a distance metric tailored for counterfactuals, differentiating them from adversarial points

- ***Williams, Joshua Nathaniel**, Anurag Katakkar, Hoda Heidari, and J Zico Kolter (2024). "Rethinking Distance Metrics for Counterfactual Explainability". In:* arXiv preprint arXiv:2410.14522

Studied properties and convergence rates of discrete prompt optimizers

- ***Williams, Joshua Nathaniel**, Avi Schwarzschild, and J Zico Kolter (2024). "Prompt recovery for image generation models: A comparative study of discrete optimizers". In:* arXiv preprint arXiv:2408.06502

Proposed a method for preserving gradients of one model's loss wrt another model's embeddings

- ***Williams, Joshua Nathaniel** and J Zico Kolter (2024). "FUSE-ing Language Models: Zero-Shot Adapter Discovery for Prompt Optimization Across Tokenizers". In:* First Conference on Language Modeling

Analyzed reactivity of Stable Diffusion's skin-tone representations to the user's dialect

- ***Williams, Joshua N**, Molly FitzMorris, Osman Aka, and Sarah Laszlo (2024). "DrawL: Understanding the Effects of Non-Mainstream Dialects in Prompted Image Generation". In:* arXiv preprint arXiv:2405.05382

Developed protocols for integrating ML into US Air Force human resource management systems.

- *David Schulker, Matthew Walsh, Joshua Snoke, and **Williams, Joshua** (2024). "Safe Use of Machine Learning for Air Force Human Resource Management: Volume 4, Evaluation Framework and Use Cases". In:* RAND Corporation

SELECTED HONORS AND AWARDS

Carnegie Mellon Graduate Student Service Award **July 2021**
Ford Foundation Predoctoral Fellowships **September 2019 - August 2022**