Michael Feffer

mfeffer@andrew.cmu.edu
mfeffer.github.io

Education

Carnegie Mellon University (CMU) Pittsburgh, PA
Doctor of Philosophy in Societal Computing.
GPA: 4.1/4.0 (September 2021 – May 2026)

Massachusetts Institute of Technology (MIT) Cambridge, MA
Master of Engineering in Electrical Engineering and Computer Science.
GPA: 5.0/5.0 (September 2017 – June 2018)

Massachusetts Institute of Technology (MIT) Cambridge, MA
Bachelor of Science in Computer Science and Minor in Music.
GPA: 5.0/5.0 (August 2014 – June 2018)

Honors

Phi Beta Kappa Honor Society Cambridge, MA
Elected for membership of the Xi (Massachusetts) chapter of the national honor society based on academic standing and pursuit of the sciences and liberal arts. (June 2018 – present)

Tau Beta Pi Engineering Honor Society Cambridge, MA
Invited to apply to MIT's chapter of the national honor society based on overall academic standing, after which community service requirements were completed to become initiated as a full member of the chapter. (February 2017 – present)

Fellowships and Awards

GEM Fellowship (2021 – present)
ARCS Scholarship (2021 – 2024)
AIES 2024 Best Paper Award (with Sinha A., Deng W.H., Lipton Z.C., Heidari H.)
for "Red-Teaming for Generative AI: Silver Bullet or Security Theater?"

Research Experience

CMU School of Computer Science Pittsburgh, PA
Work as a PhD student researcher in the Approximately Correct Machine Intelligence (ACMI) Lab under the supervision of Prof. Zachary Chase Lipton and Prof. Hoda Heidari. Analyze problems where AI, society, and the humanities interact and intersect. Research areas include algorithmic fairness, music information retrieval, data science for social good, and participatory approaches to machine learning. Submit research to top conferences. (August 2021 – present)

Software Engineering Institute (SEI) Pittsburgh, PA
Work as a threat analysis intern as part of the CERT Division of the SEI. Reconcile the differences between traditional red-teaming in cybersecurity and red-teaming for generative AI systems. Alongside collaborators, craft literature reviews and position papers arguing for more rigorous AI red-teaming and the adoption of AI expertise by cybersecurity teams. (December 2024 – present)

Spotify New York, NY
Worked as a graduate research intern on improving automated speech transcription, namely by exploring modifications to OpenAI's Whisper family of open-source language transcription models, with the goal of obtaining better transcripts for Spotify's podcasts. Conducted experiments informed by recent research where approaches ranged from model prompting to model fine-tuning via low-rank adaptation (LoRA). Documented code and noted ideas for future work for team. (June – August 2024)

IBM Research Yorktown Heights, NY

Worked on two exploratory research projects related to large language models (LLMs) as part of the Responsible and Inclusive Tech team. The first investigated novel processes for prompting LLMs. The second involved prototyping user interfaces that safeguard against malicious prompts while recommending beneficial prompts. Both projects utilized open-source LLMs and featured quantitative and qualitative analyses of text generations. Paper detailing LLM prompting approach accepted to 2024 COLM conference. (May – August 2023)

IBM Research (virtual) Yorktown Heights, NY

Developed software and performed machine learning research as a summer research intern in the AI Engineering organization. Implemented enhancements and fixed bugs in Lale, a Python package for "semi-automated data science" compatible with both scikit-learn and IBM's AI Fairness 360 Toolkit. After reviewing existing literature in the fairness in ML space, conducted original research exploring the usage of algorithmic bias mitigation techniques in conjunction with ensemble learning across a myriad of datasets to examine conditions in which fairness generalizes well. Papers detailing findings accepted to 2022 ICML DataPerf workshop and 2023 AutoML conference. (May – August 2021)

MIT Media Lab Cambridge, MA

Worked as a graduate researcher in the Affective Computing Group under the supervision of Dr. Ognjen (Oggi) Rudovic and Prof. Rosalind W. Picard. Explored personalized machine learning techniques that perform human affect estimation with the end goal of creating personalized systems to detect valence and arousal levels from video and images. Read existing literature for architecture inspiration and wrote code to test architectures and hyperparameters with multiple datasets. Also assisted other students in the lab when possible and left behind documented code to allow for running experiments after leaving the lab. Work yielded master's thesis. (September 2017 – June 2018)

MIT Computer Science and Artificial Intelligence Laboratory Cambridge, MA

As an undergraduate researcher, applied various AI and machine learning techniques for Prof. Randall Davis to improve automatic written-digit recognition with the overall goal of assessing medical patients' mental health based on handwritten responses to specific tests. Iteratively developed functionality by experimenting and evaluating changes in performance and created proper documentation for each code contribution. Integrated a neural net platform into the project at the end of my time in the lab, paving the way for future research with more novel techniques.
(June 2016 – May 2017)

Publications

Bukey I., Feffer M., Donahue C. (2024) Just Label the Repeats for In-The-Wild Audio-to-Score Alignment. International Society for Music Information Retrieval Conference (ISMIR), 2024.

Feffer M., Sinha A., Deng W.H., Lipton Z.C., Heidari H. (2024) Red-Teaming for Generative AI: Silver Bullet or Security Theater? AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2024.

Feffer M., Xu R., Sun Y., Yurochkin M. (2024) Prompt Exploration with Prompt Regression. Conference on Language Modeling (COLM), 2024.

Feffer M., Lipton Z.C., Donahue C. (2023) DeepDrake ft. BTS-GAN and TayloRVC: A Survey of Musical Deepfake Models. 2nd Workshop on Human-Centric Music Information Research (HCMIR@ISMIR), 2023.

Feffer M., Martelaro N., Heidari H. (2023) The AI Incident Database as an Educational Tool to Raise Awareness of Harms: A Classroom Exploration of Efficacy, Limitations, & Future Design Improvements. ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), 2023.

Feffer M., Hirzel M., Hoffman S.C., Kate K., Ram P., Shinnar A. (2023) Searching for Fairer Machine Learning Ensembles. The International Conference on Automated Machine Learning (AutoML), 2023.

Hirzel M., Feffer M. (2023) A Suite of Fairness Datasets for Tabular Classification. arXiv, 2023.

Feffer M., Skirpan M., Heidari H., Lipton Z.C. (2023) From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2023.

Feffer M., Heidari H., Lipton Z.C. (2023) Moral Machine or Tyranny of the Majority? The AAAI Conference on Artificial Intelligence (AAAI), 2023.

Feffer M., Lipton Z.C., Donahue C. (2022) Assistive Alignment of In-The-Wild Sheet Music and Performances. Late-Breaking Demo for the International Society for Music Information Retrieval Conference (ISMIR), 2022.

Feffer M., Rudovic O., Picard R.W. (2018) A Mixture of Personalized Experts for Human Affect Estimation. The International Conference on Machine Learning and Data Mining (MLDM), 2018. Press Release: http://news.mit.edu/2018/helping-computers-perceive-human-emotions-0724

Working Papers

Agnew W., Barnett J., Chu A., Hong R., Feffer M., Netzorg R., Awumey E., Das S. (2024) Sound Check: Auditing Audio Datasets.

Teaching Experience

CMU Machine Learning, Ethics, and Society Pittsburgh, PA

Served as a teacher's assistant (TA) for a course covering societal impacts and effects of machine learning, such as fairness, accountability, transparency, and ethics. Responsibilities included holding office hours, grading assignments, and guiding final project work. (January 2023 – May 2023)

CMU Math and Computational Foundations for Machine Learning Pittsburgh, PA

Served as a teacher's assistant (TA) for two half-semester courses that cover math and computer science fundamentals for machine learning. Responsibilities included maintaining the course websites, holding office hours, leading several recitation sessions, and creating and grading homework assignments and quizzes. (August 2022 – December 2022)

MIT Intro to Machine Learning Cambridge, MA

Served as a course lecturer's assistant (LA) for one semester for an intro to machine learning class. Obligations included writing solutions for homework problems and ensuring that students understood key concepts. Served as a teacher's assistant (TA) the following semester for the same class. Had the same obligations as before, plus was additionally responsible for proctoring and grading exams as well as responding to student questions online. (September 2017 – May 2018)

MIT MISTI Global Teaching Labs (GTL) Brazil Recife, Brazil

Taught a series of workshops focusing on science, engineering, and communication to Brazilian high school students. Led hands-on exercises in computer programming, circuit construction, chemistry, and public speaking, and aided students when necessary. Also discussed life in the US and gave presentations on applying to MIT. (January 2017)

MIT Experimental Study Group (ESG) Cambridge, MA

Helped teach both introductory biology and chemistry as an undergraduate teacher's assistant (TA). Held office hours for both biology and chemistry to review concepts and reinforce understanding of the material covered in recent lectures. Also led recitations for chemistry by going over practice problems to enable students to learn by example and address any issues they might have had. (September 2015 – May 2016, September 2016 – December 2016)

Service

Organizer

- Volunteer for GenLaw Center's workshop on Evaluating Generative AI Systems: the Good, the Bad, and the Hype (April 2024)
- Volunteer for CMU expert convening on Supporting NIST's Development of Guidelines on Red-teaming for Generative AI (March 2024)
- Co-leader of CMU Fairness, Ethics, Accountability, and Transparency (FEAT) in Machine Learning Reading Group (February 2022 – May 2023)

Program Committee

- AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) (June 2025)
- AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) (May 2024)

Reviewer

- Artificial Intelligence Review (November 2024)
- ACM Conference on Human Factors in Computing Systems (CHI) (October 2024)
- NeurIPS Workshop on Red Teaming GenAI (September 2024)
- ICLR Workshop on Secure and Trustworthy Large Language Models (SeT LLM) (February 2024)
- Topics in Cognitive Science (topiCS) (July 2022 and July 2023)
- NeurIPS Workshop on Human Evaluation of Generative Models (HEGM) (October 2022)

Academic Presentations & Discussions

- FAccT Tutorial Presenter (with Manish Nagireddy and Ioana Baldini) (June 2025)
  "Participatory & Periodic Red-Teaming of LLMs"
- NAACL Tutorial Presenter (with Manish Nagireddy and Ioana Baldini) (May 2025)
  "DAMAGeR: Deploying Automatic and Manual Approaches to GenAI Red-teaming"
- AAAI Tutorial Presenter (with Manish Nagireddy and Ioana Baldini) (February 2025)
  "DAMAGeR: Deploying Automatic and Manual Approaches to GenAI Red-teaming"
- Speaker for CMU "Current Topics in Privacy" Seminar (February 2025)
- Speaker at ARCS Pittsburgh Scholar Showcase (April 2024)
- Speaker/panelist during "Artificial Intelligence: Friend or Foe?" event of ARCS Forward series (February 2024)
- Speaker for UT Austin class "Social Applications and Impact of NLP" (February 2024)
- Speaker for CS Reading Group at Qazvin Islamic Azad University (QIAU) in Iran (February 2024)

Work Experience

Mastercard Arlington, VA

Developed software for the Data and Services division. Tackled work ranging from fullstack and frontend-facing features to backend API design and research of data science techniques. Worked with a variety of different languages and frameworks as a result (including but not limited to R, C#, React Typescript, and SQL). Quickly fixed bugs, addressed techdebt, and responded to client issue tickets in addition to normal development responsibilities. (August 2018 – April 2021)

IMC Financial Markets Chicago, IL

Wrote code as a software engineering intern over the course of a summer and collaborated with another intern to revolutionize trading software configuration of the company. Work involved customizing Docker containers with tools created using Go, Javascript, and MongoDB. Gave final presentation to interns and full-time employees to summarize improvements. (June – August 2017)

Codecademy Cambridge, MA

As an advisor, taught users of the Codecademy platform how to code via online chat. Routine tasks involved helping students understand lessons and working with them to troubleshoot coding errors. Also offered advice to members regarding which languages to learn according to their individual goals and desired work environments. (August – September 2016)

United States Department of Defense Elkridge, MD

Worked as a software developer intern over the course of a summer alongside two other interns. Conducted vulnerability analysis and wrote tools in Python, PHP, and C and wrote documentation and usage guides. Also engaged in the code review process with other interns. (June – August 2015)

Website and Mobile Application Developer State College, PA

Worked as a self-employed developer to design and update websites and apps for optimal end-user experience. Five websites were contracted over eight years. Also conceived, designed, and developed *Route Maker*, an iOS app formerly available on the App Store. (February 2010 – June 2018)