Keegan Hines, Ph.D.

keegan.hines@gmail.com http://keeganhin.es linkedin.com/in/keeganhines/

Experience

Principal Applied Scientist - Microsoft - 2023 to Present

- Leading research on the safety and security of generative AI systems; building and deploying robust defensive mechanisms to protect against attacks and threats to AI models.

Vice President of Machine Learning - ArthurAI - 2020 to 2023

- At Arthur, we've built the first model monitoring platform for the enterprise. Joined at seed stage, and helped build Arthur's capabilities and success with Fortune 50 customers. Helped grow company from 5 employees to over 60, from seed stage through Series B funding.
- Lead product strategy and the development of ML capabilities focused on Data Drift, Explainable AI, and Algorithmic Fairness and Bias Mitigation. Led ML research, development of IP portfolio, and product integration.
- Built out teams and best practices for ML Sales Engineering and Customer Success. Worked closely with our sales organization throughout the account lifecycle and engage with enterprise executive stakeholders as well as practitioner product users.

Co-founder and Chair - CAMLIS - 2017 to Present

- The Conference on Applied Machine Learning for Information Security is an annual conference for discussing new developments in machine learning as applied to problems in cybersecurity and defense. The conference meets annually with ~200 person attendance, with hundreds more on live stream.
- Coordinate efforts in executing our annual event and serving our community. This includes securing corporate sponsorship, venue, program and speakers, and event management.

Director of Machine Learning Research - Capital One - 2017 to 2020

- Built and led the ML Research team within the Center for Machine Learning. Worked with a phenomenal team to develop novel applications of ML in critical financial services areas such Explainable AI, Graph Representation Learning, and Computer Vision.
- Established and demonstrated the strategic importance of ML Research at Capital One by delivering publications at top conference and workshops, strong patent portfolio, and release of open source projects. Developed strong partnerships with business lines and oversaw the production deployment of novel ML systems in areas such as fraud, marketing, and security.
- Worked with leadership team on the development and strategy of Capital One's Center for Machine Learning. Helped grow the Center from 10 employees to over 150 through recruiting, interviewing, and mentoring engineers and data scientists.

Data Scientist - IronNet Cybersecurity - 2015 to 2017

- Worked with a team of brilliant data scientists and developers to build powerful algorithms for anomaly detection on computer networks.

Education

University of Texas at Austin
PhD in Neuroscience, 2014
Doctoral Advisor: Richard W. Aldrich

Washington and Lee University
BS in Physics, *magna cum laude*, 2009

Teaching Experience

Adjunct Assistant Professor - Georgetown University - 2017 to 2021

- Develop and teach graduate coursework in Georgetown's Masters in Data Science program.
- *ANLY-512 Statistical Learning Theory*: topics include classification and regression, model evaluation, parametric and nonparametric methods, regularization, and unsupervised methods.
- *ANLY-590 Neural Networks and Deep Learning*: topics include artificial neural networks, optimization and gradient descent, backpropagation, convolutional neural networks, recurrent neural networks, autoencoders, embeddings, generative methods.

Select Publications Google Scholar

K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, E. Kiciman. 2024. Defending Against Indirect Prompt Injection Attacks With Spotlighting. https://arxiv.org/abs/2403.14720

Yi, J., Y. Xie, B. Zhu, **K. Hines**, E. Kiciman, G. Sun, X. Xie, F. Wu. 2023. Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models. *https://arxiv.org/abs/2203.07490*

Kwegyir-Aggrey, K, J. Dai, J. Dickerson, **K. Hines**. 2022. Achieving Downstream Fairness with Geometric Repair. *https://arxiv.org/abs/2203.07490*

Kumar, I. E., **K. Hines**, J. Dickerson 2022. Equalizing Credit Opportunity in Algorithms: Aligning Algorithmic Fairness Research with U.S. Fair Lending Regulation. *AIES.*

Verma, S., **K. Hines**, J. Dickerson 2021. Amortized Generation of Sequential Counterfactual Explanations for Black-box Models *AAAI. https://arxiv.org/abs/2106.03962*

Verma, S., J. Dickerson, **K. Hines** 2020. Counterfactual Explanations for Machine Learning: A Review. *NeurIPS RSA Wkshp. https://arxiv.org/abs/2010.10596*. Best Paper Award.

C.B. Bruss, A. Khazane, J. Rider, R. Serpe, A. Gogoglou, **K. Hines** 2019. Embedding Graphs of Financial Transactions. *IEEE ICMLA. https://arxiv.org/pdf/1907.07225.pdf*. Spotlight Talk.

R. Sarshogh and **K. Hines**. 2019. A Multitask Network for Localization and Extraction of Text From Images. *IEEE ICDAR. https://arxiv.org/pdf/1906.09266.pdf*

Hines, K. 2015. A Primer On Bayesian Inference For Biophysical Systems. *Biophysical Journal. 108(9) 2103-2113.*

Hines, K., J. Bankston, R. Aldrich. 2015. Analyzing Single Molecule Time Series Via Nonparametric Bayesian Inference. *Biophysical Journal. 108(3) 540-556.*

Hines, K. 2013. Inferring Subunit Stoichiometry from Single Molecule Photobleaching. *Journal of* General Physiology. 141(6):737-746.