Alexander Wei

Research Scientist

awei@eecs.berkeley.edu
www.alexwei.org

Experience

OpenAI, *Member of Technical Staff* January 2024–

Education

UC Berkeley, *Ph.D. in Computer Science* 2020–2023
Advisors: Nika Haghtalab, Michael I. Jordan, and Jacob Steinhardt

Harvard University, *S.M. in Computer Science* 2019–2020

Harvard University, *A.B. summa cum laude in Computer Science and Mathematics* 2016–2020
Advisors: Jelani Nelson and Scott Kominers
Secondary field: Economics

Phillips Exeter Academy, *Diploma* 2012–2016

Awards and Honors

Meta Research PhD Fellowship, *Economics & Computation* 2022–2023

NSF Graduate Research Fellowship 2020–2023

International Olympiad in Informatics, *Gold Medal* 2015

Internships

Meta AI, *Research Intern* Summer 2022

Microsoft Research, *Undergraduate Research Intern* Summer 2020

D. E. Shaw & Co., *Quantitative Research Intern* Summer 2019

Google, *Software Engineering Intern* Summer 2017

Journal Papers

** denotes alphabetical order.*

∗ Learning Equilibria in Matching Markets from Bandit Feedback. *Journal of the ACM*, 2023.
Meena Jagadeesan*, Alexander Wei*, Yixin Wang, Michael I. Jordan, and Jacob Steinhardt.

∗ Human-Level Play in the Game of *Diplomacy* by Combining Language Models with Strategic Reasoning. *Science*, 2022.
Meta Fundamental AI Research Diplomacy Team, Anton Bakhtin*, Noam Brown*, Emily Dinan*, Gabriele Farina*, Colin Flaherty*, Daniel Fried*, Andrew Goff*, Jonathan Gray*, Hengyuan Hu*, Athul Paul Jacob*, Mojtaba Komeili*, Karthik Konath*, Minae Kwon*, Adam Lerer*, Mike Lewis*, Alexander H. Miller*, Sasha Mitts*, Adithya Renduchintala*, Stephen Roller*, Dirk Rowe*, Weiyan Shi*, Joe Spisak*, Alexander Wei*, David Wu*, Hugh Zhang*, and Markus Zijlstra*.

∗ Designing Approximately Optimal Search on Matching Platforms. *Management Science*, 2022.
Nicole Immorlica*, Brendan Lucier*, Vahideh Manshadi*, and Alexander Wei*.
INFORMS Auctions & Market Design Rothkopf Prize, 3rd place.

∗ Optimal Las Vegas Approximate Near Neighbors in ℓ_p. *ACM Transactions on Algorithms*, 2022.
Alexander Wei.

∗ An Interscholastic Network To Generate LexA Enhancer Trap Lines in Drosophila. *G3: Genes, Genomes, Genetics*, 2019.
Lutz Kockel et al. [including Alexander Wei.]

Conference Papers

** denotes alphabetical order.*

∗ Covert Malicious Finetuning: Challenges in Safeguarding LLM Adaptation. *Proceedings of the 41st International Conference on Machine Learning* (*ICML 2024*).
Danny Halawi*, Alexander Wei*, Eric Wallace, Tony Tong Wang, Nika Haghtalab, and Jacob Steinhardt.

∗ Jailbroken: How Does LLM Safety Training Fail? *Proceedings of the 37th Conference on Advances in Neural Information Processing Systems* (*NeurIPS 2023*).
Alexander Wei, Nika Haghtalab, and Jacob Steinhardt.
NeurIPS 2023 Oral Presentation.

∗ TCT: Convexifying Federated Learning using Bootstrapped Neural Tangent Kernels. *Proceedings of the 36th Conference on Advances in Neural Information Processing Systems* (*NeurIPS 2022*).
Yaodong Yu, Alexander Wei, Sai Praneeth Karimireddy, Yi Ma, and Michael I. Jordan.

∗ More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize. *Proceedings of the 39th International Conference on Machine Learning* (*ICML 2022*).
Alexander Wei, Wei Hu, and Jacob Steinhardt.

∗ Predicting Out-of-Distribution Error with the Projection Norm. *Proceedings of the 39th International Conference on Machine Learning* (*ICML 2022*).
Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt.

∗ Learning in Stackelberg Games with Non-myopic Agents. *Proceedings of the 2022 ACM Conference on Economics and Computation* (*EC 2022*).
Nika Haghtalab*, Thodoris Lykouris*, Sloan Nietert*, and Alexander Wei*.

∗ Learning Equilibria in Matching Markets from Bandit Feedback. *Proceedings of the 35th Conference on Advances in Neural Information Processing Systems* (*NeurIPS 2021*).
Meena Jagadeesan*, Alexander Wei*, Yixin Wang, Michael I. Jordan, and Jacob Steinhardt.
NeurIPS 2021 Spotlight Presentation.

∗ Designing Approximately Optimal Search on Matching Platforms. *Proceedings of the 2021 ACM Conference on Economics and Computation* (*EC 2021*).
Nicole Immorlica*, Brendan Lucier*, Vahideh Manshadi*, and Alexander Wei*.
INFORMS Auctions & Market Design Rothkopf Prize, 3rd place.

* Optimal Robustness-Consistency Trade-Offs for Learning-Augmented Online Algorithms. *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems* (*NeurIPS 2020*).
Alexander Wei* and Fred Zhang*.

* Better and Simpler Learning-Augmented Online Caching. *Proceedings of the 23rd International Conference on Approximation Algorithms for Combinatorial Optimization Problems* (*APPROX 2020*).
Alexander Wei.

* Allocation for Social Good: Auditing Mechanisms for Utility Maximization. *Proceedings of the 2019 ACM Conference on Economics and Computation* (*EC 2019*).
Taylor Lundy, Alexander Wei, Hu Fu, Scott Duke Kominers, and Kevin Leyton-Brown.

* Optimal Las Vegas Approximate Near Neighbors in ℓ_p. *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (*SODA 2019*).
Alexander Wei.
SODA 2019 Best Student Paper.

* Varying the Number of Signals in Matching Markets. *Proceedings of the 14th Conference on Web and Internet Economics* (*WINE 2018*).
Meena Jagadeesan* and Alexander Wei*.

Teaching

UC Berkeley, *Teaching Assistant for CS 170: Efficient Algorithms and Intractable Problems* Fall 2023

Harvard University, *Teaching Fellow for CS 124: Data Structures and Algorithms* Spring 2018
Awarded *Certificate of Distinction in Teaching* by the Derek Bok Center.

IDEA MATH, *Teaching Assistant* [*Advanced topics for math competitions.*] Summer 2015

Presentations

* Jailbroken: How Does LLM Safety Training Fail?
  - FAR.AI Bay Area Alignment Workshop 2024
  - Advances in Neural Information Processing Systems (NeurIPS) 2023.

* More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize.
  - International Conference on Machine Learning (ICML) 2022.

* Learning in Stackelberg Games with Non-myopic Agents.
  - ACM Conference on Economics and Computation (EC) 2022.

* Learning Equilibria in Matching Markets from Bandit Feedback.
  - Advances in Neural Information Processing Systems (NeurIPS) 2021.

* Designing Approximately Optimal Search on Matching Platforms.
  - Rotman Young Scholar Seminar @ University of Toronto.
  - INFORMS Annual Meeting 2021.
  - ACM Conference on Economics and Computation (EC) 2021.

- [Poster] Marketplace Innovation Workshop (MIW) 2021.

* Better and Simpler Learning-Augmented Online Caching.
  - Approximation Algorithms for Combinatorial Optimization Problems (APPROX) 2020.

* Optimal Las Vegas Approximate Near Neighbors in ℓ_p.
  - ACM-SIAM Symposium on Discrete Algorithms (SODA) 2019.

* Varying the Number of Signals in Matching Markets.
  - Conference on Web and Internet Economics (WINE) 2018.
  - Frontiers of Market Design @ EC 2018.

Professional Services

Journal reviewer for Journal of Machine Learning Research (JMLR).

Conference reviewer for International Conference on Machine Learning (ICML), International Conference on Learning Representations (ICLR), Neural Information Processing Systems (NeurIPS), ACM-SIAM Symposium on Discrete Algorithms (SODA), Innovations in Theoretical Computer Science (ITCS), and International Colloquium on Automata, Languages and Programming (ICALP).

Workshop reviewer for Machine Learning Safety @ NeurIPS 2022, Learning in Presence of Strategic Behavior @ NeurIPS 2021, and Learning and Decision-Making with Strategic Feedback @ NeurIPS 2021.

Last updated November 25, 2024.