Benjamin Plaut, PhD

bjplaut@gmail.com Google Scholar

Education

PhD in Computer Science, Stanford University *Sept 2016 – Sept 2021*

- Thesis: *Algorithms for Fair Public and Private Resource Allocation*
- Thesis advisor: Ashish Goel

Bachelor of Science, Carnegie Mellon University *Aug 2012 – May 2016*

- Thesis: *Algorithms for Social Good: Kidney Exchange*
- Thesis advisor: Tuomas Sandholm
- Allen Newell Award for Excellence in Undergraduate Research (best thesis in Computer Science)

Work Experience

Postdoctoral Researcher, UC Berkeley *Sept 2023 – present*

- Advised by Stuart Russell
- Part of the Center for Human-Compatible AI
- Studying AI safety with a focus on caution under uncertainty

Data Scientist, Lyft *June 2021 – May 2023*

- Co-lead for multi-year initiative with 30+ people
- Created cross-org experimentation principles to ensure scientific rigor (used by 20+ people)
- Grew annual profit and revenue by millions of dollars
- Gave "Intro to Effective Altruism" talk

Research Intern, Google *Summer 2019, Summer 2020*

- 2019: Developed a fair and efficient matching algorithm for two-sided markets
- 2020: Designed and tested reinforcement learning algorithm for optimal routing

Publications

Lead author: * | Senior author: †

AI Safety

1. Safe Learning Under Irreversible Dynamics via Asking for Help
   B Plaut*, J Liévano-Karim, H Zhu, S Russell†
   Under submission.
2. Learning When Not to Learn: Risk-Sensitive Abstention in Bandits with Unbounded Rewards
   S Liaw*, **B Plaut***
   Under submission.
3. Learning to Coordinate with Experts
   M Danesh*, N Khanh, T Trinh, **B Plaut**
   Under submission.
4. Check Yourself Before You Wreck Yourself: Selectively Quitting Improves LLM Agent Safety
   V Bonagiri*, P Kumaraguru, N Khanh, **B Plaut†**
   Neural Information Processing Systems (NeurIPS) 2025 Workshops on Reliable and Regulatable ML.
5. Avoiding Catastrophe in Online Learning by Asking for Help
   B Plaut*, H Zhu, S Russell†
   International Conference on Machine Learning (ICML), 2025.

6. Probabilities of Chat LLMs Are Miscalibrated but Still Predict Correctness on Multiple-Choice Q&A
   B Plaut*, N Khanh, T Trinh
   Transactions on Machine Learning Research (TMLR), 2025.
7. Getting By Goal Misgeneralization With a Little Help From a Mentor
   T Trinh*, M Danesh, N Khanh, **B Plaut**†
   Neural Information Processing Systems (NeurIPS) 2024 Workshop on Safe and Trustworthy Agents.

Resource allocation

By field convention, author order here is mostly alphabetical.

1. Algorithms for Fair Public and Private Resource Allocation
   B Plaut.
   PhD Thesis, Stanford University, 2021.
2. Counteracting Inequality in Markets via Convex Pricing
   A Goel†, **B Plaut***
   Conference on Web and Internet Economics (WINE), 2020.
3. Almost Envy-free Repeated Matching in Two-sided Markets
   S Gollapudi†, K Kollias, **B Plaut***
   Conference on Web and Internet Economics (WINE), 2020.
4. Optimal Nash Equilibria for Bandwidth Allocation
   B Plaut.
   Conference on Web and Internet Economics (WINE), 2020.
5. Equality of Power and Fair Public Decision-making
   N Immorlica†, **B Plaut***, E G Weyl†
   Conference on Web and Internet Economics (WINE), 2019.
6. Markets Beyond Nash Welfare for Leontief Utilities
   A Goel†, R Hulett, **B Plaut***
   Conference on Web and Internet Economics (WINE), 2019.
7. Communication Complexity of Discrete Fair Division
   B Plaut*, T Roughgarden†
   SODA 2019, SIAM Journal on Computing (SICOMP), 2020.
8. Markets for Public Decision-making
   N Garg, A Goel†, **B Plaut***
   Conference on Web and Internet Economics (WINE), 2018.
9. Almost Envy-Freeness with General Valuations
   B Plaut*, T Roughgarden†
   Symposium on Discrete Algorithms (SODA), 2018; SIAM Journal on Discrete Mathematics (SIDMA), 2020.
10. Algorithms for Social Good: Kidney Exchange
    B Plaut.
    Undergraduate Honors Thesis, Carnegie Mellon University, 2016.
11. Hardness of the Pricing Problem for Chains in Barter Exchanges
    B Plaut*, JP Dickerson, T Sandholm†
    Preprint, 2016.
12. Position-Indexed Formulations for Kidney Exchange
    JP Dickerson, D Manlove†, **B Plaut**, T Sandholm†, J Trimble*
    Economics and Computation (EC), 2016.
13. Fast Optimal Clearing of Capped-Chain Barter Exchanges
    B Plaut*, JP Dickerson, T Sandholm†

AAAI Conference on Artificial Intelligence, 2016.

Physical Chemistry

1. Direct Observation of Folding Energy Landscape of RNA Hairpin at Mechanical Loading Rates
   H Xu*, **B Plaut**, X Zhu, M Chen, U Mavinkurve, A Maiti, G Song, K Murari, M Mandal†
   The Journal of Physical Chemistry B, 2017.

Other Experience

President, Co-founder, and primary instructor, Cardinal West Coast Swing *Nov 2018 – Aug 2020*

- Co-founded a west coast swing club at Stanford University.
- Designed a curriculum for beginner lessons
- Taught weekly lessons
- Trained and supervised new teachers

Teaching Assistant, Stanford University *Mar 2018 – June 2018*

- Duties included office hours, advising students on a final project, grading, and overall course structuring

Teaching Assistant, Carnegie Mellon University *Jan 2014 – Dec 2015*

- Taught recitation, ran review sessions, mentored students on a final project
- Also was "Head of Staff Morale" for a staff of 40 TAs. Organized social events, coordinated purchases, etc.

Languages

Programming Languages	Python, C++, SQL, Unix, LaTeX, Matlab
Natural Languages	English (native), Spanish (advanced proficiency)

Personal projects

- Music composition – I write music under the name Melancholy Flower (pun on "melon cauliflower"). Check out my Spotify!
- RAMbrandt – An algorithmic art-generator based on Markov chains, computer vision, and vector calculus.
- Ballroom Dance Video Tutorial – Database containing demonostration videos for ballroom dance moves.