JEREMY TIEN

jeremyti@cs.cmu.edu • jeremy29tien.github.io

RESEARCH INTERESTS

AI Alignment and Safety, AI Agents, Human-AI Interaction, Reward Learning, Machine Learning

EDUCATION

Carnegie Mellon University
Doctor of Philosophy in Machine Learning Aug 2025 – Present
Advisors: Aran Nayebi and Zico Kolter
Coursework: Intermediate Statistics (in progress), Advanced Introduction to Machine Learning (in progress), Generative Artificial Intelligence (planned), Embodied Artificial Intelligence Safety (planned).

The Full-time Training in Anaheim Aug 2023 – July 2025
The Full-time Training in Anaheim (FTTA) is a two-year postgraduate program dedicated to preparing men and women for Christian service by fostering their spiritual growth and maturity, equipping them with a thorough and practical realization of Biblical truths through a comprehensive theological curriculum, and cultivating their skills in interpersonal relationships, collaborative teamwork, conflict resolution, counseling and mentorship, public speaking, community service, and domestic and international outreach.

University of California, Berkeley **Cumulative GPA: 4.00/4**
Bachelor of Science in Electrical Engineering and Computer Science, Highest Honors in General Scholarship Aug 2019 – May 2023
EECS Honors Program. Breadth Area: Cognitive Science (AI and the Brain)
Selected Coursework: Algorithmic Human-Robot Interaction, Deep Learning, Machine Learning, Artificial Intelligence, Basic Issues in Cognition, Computational Models of Cognition, Language Acquisition, Optimization Models, Algorithms, Data Structures, Operating Systems, Computer Architecture, Signals, Discrete Math, Multivariable Calculus, Computational Music, Elementary German (1 year).

PUBLICATIONS

Active Reward Learning and Iterative Trajectory Improvement from Comparative Language Feedback
Eisuke Hirota, Zhaojing Yang, Ayano Hiranaka, Miru Jun, **Jeremy Tien**, Stuart J. Russell, Anca Dragan, Erdem Biyik.
The International Journal of Robotics Research (IJRR), 2025.

Trajectory Improvement and Reward Learning from Comparative Language Feedback
Zhaojing Yang, Miru Jun, **Jeremy Tien**, Stuart J. Russell, Anca Dragan, Erdem Bıyık.
8th Annual Conference of Robot Learning (CoRL), 2024.

Optimizing Robot Behavior via Comparative Language Feedback
Jeremy Tien*, Zhaojing Yang*, Miru Jun, Stuart J. Russell, Anca Dragan, Erdem Bıyık.
HRI 2024 Workshop on Human-Interactive Robot Learning (HIRL).

Causal Confusion and Reward Misidentification in Preference-Based Reward Learning
Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, Daniel S. Brown.
International Conference on Learning Representations (ICLR), 2023.

A Study of Causal Confusion in Preference-Based Reward Learning (IL, SCIS)
Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, Daniel S. Brown.
RSS 2022 Workshop on Overlooked Aspects of Imitation Learning: Systems, Data, Tasks, and Beyond (IL); ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability (SCIS).

PROFESSIONAL EXPERIENCE

Carnegie Mellon University Machine Learning Department | Pittsburgh, PA
Graduate Student Researcher (Profs. Aran Nayebi and Zico Kolter) August 2025 – Present
- Leading a project to implement a provably corrigible agent. Rather than using a single learned reward for fine-tuning, we independently learn five utility heads corresponding to (1) deference to humans, (2) off-switch preservation, (3) truthfulness, (4) action reversibility, and (5) ordinary task usefulness, which are combined with lexicographic weights $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4 > \alpha_5 > 0$, following the framework proposed by Nayebi, A. 2025. This subsumes ordinary RLHF/RLAIF, which appears in the fifth reward head, and provably guarantees adherence to the aforementioned safety (corrigibility) values.

Berkeley Artificial Intelligence Research | Berkeley, CA
Undergraduate Researcher, InterACT Lab (Prof. Anca Dragan) August 2021 – August 2024
- Led project to systematically analyze and explain the failure modes of preference-based reward learning (also known as reinforcement-learning from human feedback, or RLHF), the primary paradigm behind several popular AI models today (including ChatGPT) and found that PbRL is susceptible to causal confusion and reward misidentification. Performed sensitivity and ablation analyses on various (often overlooked) components of PbRL. Published at **ICLR 2023** as the **first author** of "Causal Confusion and Reward Misidentification in Preference-Based Reward Learning." Presented earlier iterations at RSS 2022 Imitation Learning and ICML 2022 Spurious Correlation workshops.
- Developed a novel way to learn reward functions from human feedback in the form of natural language, making use of T5 for language encoding. Language feedback was grounded in robot trajectory comparisons using a shared latent space. Formulated novel preference learning loss function

(accounting for explicit and implicit preferences), as well as a novel human response model based on feature-specific rewards (rather than trajectory rewards) for reward inference. Published as "Trajectory Improvement and Reward Learning from Comparative Language Feedback" at **CoRL 2024**.

Amazon | Seattle, WA

Software Development Engineer Intern, Fashion Tech May 2021 – August 2021

- Individually developed (full-stack) and tested an AWS cloud service to convert images of product size-charts into structured data using computer vision, which involved defining cloud infrastructure (Cloud Development Kit), backend development (Lambda), and frontend integration (Java Spring).
- Owned the project design process: gathered requirements from key stakeholders, compiled research on needed technologies, made decisions on project scope and priorities, created milestones and effort estimates, and compared the scalability and performance of candidate solutions. Drafted a Business Requirement Document and a Design Document and led review meetings to collect feedback from SDMs, PMs, and SDEs.

Tact.ai | Sunnyvale, CA

Artificial Intelligence Research Intern May 2020 – August 2020

- Spearheaded development of a Pointer Network machine-learning model addressing the task of text-to-query (generating a database query in SQL given a question in natural language)—enabling individuals without SQL knowledge to quickly access and manipulate large amounts of data using their voice.
- Trained and evaluated model on the challenging Spider dataset (*Tao Yu et al., 2018*), achieving an auROC of 0.86 in database column selection.
- Collaborated with Artificial Intelligence team to integrate project with Tact.ai's AI Sales Assistant. Presented work in a company-wide final deliverable.

Stanford School of Medicine | Stanford, CA

Research Intern, Montgomery Lab (Prof. Stephen Montgomery) June 2018 – August 2019

- Developed an ensemble machine-learning algorithm to identify GWAS-eQTL colocalizations given output data from several other colocalization tools. Achieved auROC of 0.86 and 90% accuracy. Integrated model into colocalization pipeline. Co-author on "Ensemble colocalization method improves causal gene prioritization in simulations and GWAS" (Gloudemans *et al.* ASHG 2019 Presentation https://doi.org/10.5281/zenodo.3625132).

SERVICE AND OUTREACH

The Full-Time Training in Anaheim (FTTA) | Anaheim, CA

Youth Outreach Coordinator and Mentor August 2023 – July 2025

- Holistically mentored middle- and high-school-aged young adults in the Church in Fullerton, with whom the FTTA is partnered.
- Organized weekly appointments with groups of 2 or 3 (generally 3 separate 1-3 hour appointments per week), spending time to understand their backgrounds, know them personally, and provide specific mentorship on a variety of topics, both human (e.g., college applications, sports) and spiritual.
- Played a major role in supporting weekly gatherings (15-20 young adults) and semesterly conferences (100+ young adults); frequently led singing sessions (while playing the guitar) and spoke about life, morality, peace, joy, and other topics of spiritual help by incorporating personal experiences from my life both on spontaneous occasions and after some preparation.

AI4ALL | Berkeley, CA

Teaching Assistant August 2022

- Supported AI4ALL, a weeklong summer program for high school students from backgrounds that are underrepresented in the field of AI.
- Provided technical (coding) and conceptual help as well as personal mentorship to students, which included walking them through and debugging code, explaining AI/ML concepts using illustrations, and giving personal anecdotes about my experience in AI/ML and college in general.
- Gave a 15 minute talk about my personal journey into AI. Also discussed my current project at a more elementary level by introducing a naive attempt at solving the problem of assistive healthcare, a more complex attempt involving machine learning, and the solution on which I was working.

UC Berkeley EECS Department | Berkeley, CA

Lab Assistant Jan 2020 – May 2020

- Taught in weekly lab sections of COMPSCI61A, UC Berkeley's largest computer science class. Lab sections consisted of 20 to 30 students.
- Presented lectures on class topics, fielded questions, debugged student code, walked through different solution strategies on a student-by-student basis.

TALKS AND PRESENTATIONS

Lightning Talk: "My AI Journey: AI Everywhere" and Project Presentation, Berkeley Artificial Intelligence Research (BAIR) AI4ALL, August 2022.

LEADERSHIP AND EXTRACURRICULARS

Christians on Campus at Berkeley | Berkeley, CA

Student Lead August 2019 – May 2023

- Led and facilitated various club gatherings—Bible studies (~10 people), Fellowship Nights (~50 people), and College Conferences (100+ people).
- Arranged and engaged in weekly one-on-one appointments with younger students in the club to provide mentorship, counseling, and support.
- Worked with 1-2 other teachers to teach and take care of elementary school-aged children in a weekly Bible lesson for our community church.

Berkeley ABA | Berkeley, CA

Marketing Strategy Project Manager (SoFi) May 2020 – Dec 2020

- Led project team of 7 (senior analysts and analysts) in a semester-long consulting project with the Strategic Finance Team at SoFi.
- Planned and executed a marketing strategy project with in-depth market research, competitor and SWOT analyses, primary research via Qualtrics and focus groups, data analysis in Numpy and Pandas, market segmentation and customer personas, and mockups for SoFi Money's smart-budget offering.

Strategy Consulting Senior Analyst (Facebook) Dec 2019 – May 2020

- Headed analyst team of 5 in a semester-long consulting project with Facebook; collaborated with Facebook via weekly updates and deliverables.
- Conducted market intelligence analysis in 12 industries across 4 geographical regions, researching trends and key players via secondary sources (Mintel and Capital IQ databases) and primary sources (interviews with Haas School of Business professors and industry professionals).

TECHNICAL SKILLS

- **Programming/Data Analysis:** Python (numpy, pandas, matplotlib), C/C++, Java, Scheme, RISC-V, R, SQL, Excel, HTML, CSS, JavaScript, Git

- **Machine-learning/AI:** Pytorch, Tensorflow, Keras, Scikit-learn, OpenAI Gym, Assistive Gym, Robosuite, Ray RLlib, Stable Baselines

AWARDS AND HONORS

Graduated from UC Berkeley with Highest Honors in Electrical Engineering and Computer Sciences; a cumulative 4.0 GPA (May 2023).
University Medal Candidate. *(Invitation-only) candidate for the award given to the most distinguished graduating undergraduate senior at UC Berkeley* (Feb 2023).
Eta Kappa Nu (HKN). *International EECS Honor Society; top 25% of students with junior class standing. Was initiated the spring semester of freshman year* (Jan 2020).
Edward Frank Kraft Award. *One of Berkeley's most prestigious awards to freshmen on the campus who attain the highest scholastic records in their first semester* (Mar 2020).
Regents' and Chancellor's Scholarship Award Finalist (University of California, Berkeley). *The most prestigious scholarship offered by UC Berkeley to entering undergraduate students. The top 10% of admitted students (the top 1% of applicants) are invited to interview* (Feb 2019).
National Merit Scholarship Recipient & Nvidia Corporate Scholarship Recipient (2019).
Jessica Lynn Saal Fellowship (Stanford Institutes of Medicine Research Program (SIMR)). *The top award for "exemplary performance, enthusiasm, and leadership," awarded to 2 interns out of 50 (SIMR's acceptance rate is less than ~3%). Granted the opportunity to return the following summer with a stipend* (Aug 2018).

LINKS

Google Scholar • arXiv • LinkedIn • Github • Twitter