Nathan Lambert

nathan@natolambert.com • natolambert.com • natolambert
natolambert • natolambert • Last updated on June 19, 2025

I research **data-driven decision making**, including progressing **reinforcement learning** algorithms, applying them to real-world problems such as **large language models** and **robotics**, and planning for the **societal implications** therein.

Education

Ph.D. in Electrical Engineering and Computer Science, University of California, Berkeley (4.0/4.0) 2017 – 2022
Synergy of Prediction and Control in Model-based Reinforcement Learning
Advisors: Kristofer S.J. Pister, Roberto Calandra
Committee: Sergey Levine, Claire Tomlin

B.S. in Electrical and Computer Engineering, Cornell University (4.0/4.0) 2013 – 2017

Industry Experience

Allen Institute for AI, Seattle, WA — Research Scientist 2023 – cont.
HuggingFace, Remote — Research Scientist and RLHF Team Lead 2022 – 2023
DeepMind, London (*Virtual*) — Research Intern (Host: Martin Riedmiller) 2021
Facebook AI, Menlo Park — Research Intern and Student Researcher (Host: Roberto Calandra) 2019 – 2020
Tesla, Palo Alto Test Engineering Intern 2015

Honors & Awards

CVPR Best Paper Honorable Mention (Molmo & Pixmo) 2025
Best Reasoning & RL Talk, AI Engineer World's Fair 2025
Ai2 The Cat Herder 2024
ACL Best Resource Paper Award (Dolma) 2024
ACL Best Theme Paper Award (OLMo) 2024
GeekWire Innovation of the year (OLMo) 2024
Reward Reports - Auditing AI Mozilla Technology Fund Cohort 2023
Best Oral Presentation; Berkeley Sensor and Actuator Sensor Spring Review 2022
Best Student Paper Finalist; IEEE Symposium on Multi-Robot and Multi-Agent Systems 2021
Berkeley EECS Demetri Angelakos Memorial Achievement Award 2021
Heart to Humanity Eternal (H2H8) Pioneer 2021
NDSEG Graduate Research Fellowship Program Top 200 2018
NSF Graduate Research Fellowship Program Honorable Mention 2017, 2018
Berkeley EECS Department Fellowship 2017
Eight undergraduate scholarships 2013 – 2017
Cornell Rowing Charles E. Courtney Award, Tau Beta Pi Scholarship, Southeastern New England Defense Industry Alliance STEM Scholarship 2016, 2017, Cornell Athletics 400 Club Induction, Beta Pi Induction, Eta Kappu Nu Induction, American Society of Engineering Education SMART Scholar Award

Publications [Google Scholar: 5.7k+ citations and an h-index of 32, Semantic Scholar]

Representative publications that I am a primary author on are highlighted.

2025

1. *Spurious Rewards: Rethinking Training Signals in RLVR*
Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, **Nathan Lambert**, Sewon Min, Ranjay Krishna, and others
arXiv Preprint 2025

2. *RewardBench 2: Advancing Reward Model Evaluation*
Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and **Nathan Lambert**
arXiv Preprint 2025

3. *Reinforcement Learning from Human Feedback*
Nathan Lambert
Book 2025

4. *RewardBench: Evaluating Reward Models for Language Modeling* [code]
Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi
NAACL 2025

2024

5. *2 OLMo 2 Furious* [code]
Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, **Nathan Lambert**, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi
arXiv Preprint 2024

6. *Tülu 3: Pushing Frontiers in Open Language Model Post-Training* [code]
Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi
Technical Report 2024

7. *Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models* [code]
Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Christopher Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, **Nathan Lambert**, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Christopher Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jennifer Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Marie Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hanna Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi
arXiv Preprint 2024

8. *M-RewardBench: Evaluating Reward Models in Multilingual Settings* [code]
Srishti Gureja, Lester James V Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, **Nathan Lambert**, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee
arXiv Preprint 2024

9. *OLMoE: Open Mixture-of-Experts Language Models* [code]
Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Daniel Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, **Nathan Lambert**, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hanna Hajishirzi
arXiv Preprint 2024

10. *Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence*
Adrien Basdevant, Camille François, Victor Storchan, Kevin Bankston, Ayah Bdeir, Brian Behlendorf, Merouane Debbah, Sayash Kapoor, Yann LeCun, Mark Surman, Helen King-Turvey, **Nathan Lambert**, Stefano Maffulli, Nik Marda, Govind Shivkumar, and Justine Tunney
Workshop Proceedings 2024

11. *Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback* [code]
Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, **Nathan Lambert**, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi
NeurIPS 2024

12. *Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms* [code]
Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, **Nathan Lambert**, Yejin Choi, and Nouha Dziri
NeurIPS Dataset and Benchmarks 2024

13. *Self-directed synthetic dialogues and revisions technical report*
Nathan Lambert, Hailey Schoelkopf, Aaron Gokaslan, Luca Soldaini, Valentina Pyatkin, and Louis Castricato
Technical Report 2024

14. *Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research* [code]
Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, **Nathan Lambert**, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo
ACL 2024 (Best Paper Award)

15. *OLMo: Accelerating the Science of Language Models* [code]
Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, **Nathan Lambert**, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi
ACL 2024 (Best Paper Award)

16. *A Survey on Data Selection for Language Models*
Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, **Nathan Lambert**, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang
TMLR 2024

17. *Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback*
Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, **Nathan Lambert**, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker
ICML Position Paper 2024

18. *D2PO: Discriminator-Guided DPO with Response Evaluation Models*
Prasann Singhal, **Nathan Lambert**, Scott Niekum, Tanya Goyal, and Greg Durrett
COLM 2024

19. *Zephyr: Direct Distillation of LM Alignment* [code]
Lewis Tunstall, Edward Beeching, **Nathan Lambert**, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf
COLM 2024

20. *A Unified View on Solving Objective Mismatch in Model-Based Reinforcement Learning*
Ran Wei, **Nathan Lambert**, Anthony McDonald, Alfredo Garcia, and Roberto Calandra
Transactions on Machine Learning Research 2024

21. *BLISS: Interplanetary Exploration with Swarms of Low-Cost Spacecraft*
Alexander N Alvara, Lydia Lee, Emmanuel Sin, **Nathan Lambert**, Andrew J Westphal, and Kristofer SJ Pister
Acta Astraunica 2024

2023

22. *Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2* [code]
Hamish Ivison, Yizhong Wang, Valentina Pyatkin, **Nathan Lambert**, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi
arXiv Preprint 2023

23. *The Alignment Ceiling: Objective Mismatch in Reinforcement Learning from Human Feedback*
Nathan Lambert and Roberto Calandra
arXiv Preprint 2023

24. *Entangled Preferences: The History and Risks of Reinforcement Learning and Human Feedback*
Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick
arXiv Preprint 2023

25. *Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings*
Sarah Shoker, Andrew Reddie, Sarah Barrington, Miles Brundage, Husanjot Chahal, Michael Depp, Bill Drexel, Ritwik Gupta, Marina Favaro, Jake Hecla, Alan Hickey, Margarita Konaev, Kirthi Kumar, **Nathan Lambert**, Andrew Lohn, Cullen O'Keefe, Nazneen Rajani, Michael Sellitto, Robert Trager, Leah Walker, Alexa Wehsener, and Jessica Young
arXiv Preprint 2023

26. *Reward Reports for Reinforcement Learning* [code]
Thomas Gilbert, Sarah Dean, **Nathan Lambert**, Tom Zick, and Aaron Snoswell
AAAI/ACM Conference on AI, Ethics, and Society 2023

2022

27. *Measuring Data*
Margaret Mitchell, Alexandra Sasha Luccioni, **Nathan Lambert**, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela
arXiv Preprint 2022

28. *Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems*
Thomas Gilbert, Sarah Dean, Tom Zick, and **Nathan Lambert**
Center for Long-Term Cybersecurity Whitepaper Series 2022

29. *Investigating Compounding Prediction Errors in One-step Dynamics Models* [code]
Nathan Lambert, Roberto Calandra, and Kristofer Pister
arXiv Preprint 2022

30. *Understanding the Challenges of Exploration for Offline Reinforcement Learning*
Nathan Lambert, Markus Wulfmeier, Arunkumar Byravan, Michael Bloesch, William Whitney, Vibhavari Dasagi, Tim Hertweck, and Martin Riedmiller
arXiv Preprint 2022

2021

31. *MBRL-Lib: A Modular Library for Model-based Reinforcement Learning* [code]
Luis Pineda, Brandon Amos, Amy Zhang, **Nathan Lambert**, and Roberto Calandra
arXiv Preprint 2021

32. *BotNet: A Simulator for Studying the Effects of Accurate Communication Models on High-agent-count Multi-agent Control* [code]
Mark Selden, Felipe Campos, Jason Zhou, **Nathan Lambert**, Daniel Drew, and Kristofer Pister
Symposium on Multi-Agent and Multi-Robot Systems 2021 (Best Student Paper Finalist)

33. *Axes for Sociotechnical Inquiry in AI Research*
Sarah Dean, Thomas Krendl Gilbert, **Nathan Lambert**, and Tom Zick
Transactions on Technology and Society (TTS) 2021 (Authors arranged alphabetically)

34. *On the Importance of Hyperparameter Optimization for Model-based Reinforcement Learning*
Baohe Zhang, Raghu Rajan, Luis Pineda, **Nathan Lambert**, André Biedenkapp, Kurtland Chua, Frank Hutter, and Roberto Calandra
International Conference on Artificial Intelligence and Statistics (AISTATS) 2021

35. *Learning Accurate Long-term Dynamics for Model-based Reinforcement Learning* [code]
Nathan Lambert, Albert Wilcox, Howard Zhang, Kristofer SJ Pister, and Roberto Calandra
International Conference on Decision and Control (CDC) 2021

36. *Nonholonomic Yaw Control of an Underactuated Flying Robot With Model-Based Reinforcement Learning*
Nathan Lambert, Craig Schindler, Daniel Drew, and Kristofer Pister
Robotics and Automation Letters (RAL) 2021

2020

37. *Objective Mismatch in Model-based Reinforcement Learning*
Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra
Conference on Learning for Decision and Control (L4DC) 2020

38. *AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks*
McKane Andrus, Sarah Dean, Thomas Gilbert, **Nathan Lambert**, and Tom Zick
International Symposium on Technology and Society (ISTATS) 2020 (Authors arranged alphabetically)

39. *Learning for Microrobot Exploration: Model-based Locomotion, Robust Navigation, and Low-Power Deep Classification*
Nathan Lambert, Fahran Toddywala, Brian Liao, Eric Zhu, Lydia Lee, and Kristofer Pister
International Conference on Manipulation, Automation and Robotics at Small Scales (MARSS) 2020

40. *Learning Generalizable Locomotion Skills with HierarchicalReinforcement Learning*
Tianyu Li, **Nathan Lambert**, Roberto Calandra, Akshara Rai, and Franziska Meier
International Conference on Robotics and Automation (ICRA) 2020

2019

41. *Low-Level Control of a Quadrotor With Deep Model-Based Reinforcement Learning* [code]
Nathan Lambert, Daniel Drew, Joseph Yaconelli, Sergey Levine, Roberto Calandra, and Kristofer Pister
Robotics and Automation Letters (RAL) 2019

2018

42. *Toward Controlled Flight of the Ionocraft: A Flying Microrobot Using Electrohydrodynamic Thrust With Onboard Sensing and No Moving Parts*
Daniel S Drew, **Nathan Lambert**, Craig B Schindler, and Kristofer Pister
Robotics and Automation Letters (RAL) 2018

2017

43. *Enhanced lithium niobate pyroelectric ionizer for chip-scale ion mobility-based gas sensing*
K. B. Vinayakumar, V. Gund, **Nathan Lambert**, S. Lodha, and A. Lal
Sensors 2017

Repositories

natolambert/rlhf-book ★814 — *Book on RLHF* 2025
allenai/awesome-open-source-lms ★274 — *A curated list of open-source language models* 2024
allenai/reward-bench ★213 — *The first evaluation tool for reward models* 2024
allenai/open-instruct ★2.9k — *Post-training language models* 2023
natolambert/blogcaster ★189 — *AI tools for multimodal blogging* 2023
huggingface/alignment-handbook ★4k — *RLHF model lessons and recipes* 2023
huggingface/trl ★8.5k — *Lean library for RLHF* 2023
huggingface/simulate ★185 — *Tool for building embodied AI environments* 2022
huggingface/diffusers ★23k — *Diffusion models library* 2022
facebookresearch/mbrl-lib ★918 — *Model-based reinforcement learning library* 2021
natolambert/dynamicslearn ★54 — *Model-based RL for mixed sim. and real* 2020

Machine Learning Artifacts

Key – `M: Model, D: Dataset, S: Space`
allenai/OLMo-2-1124-13B-Instruct ★46 — `M` — *Round 2!* 2024
allenai/Llama-3.1-Tulu-3-8B ★163 — `M` — *Our SOTA post-training on Llama 3.1* 2024
allenai/tulu-3-sft-mixture ★152 — `D` — *Our new SFT dataset* 2024
allenai/OLMoE-1B-7B-0924-Instruct ★89 — `M` — *Open-source MoE model* 2024
allenai/tulu-2.5-preference-data ★17 — `D` — *Collection of preference datasets* 2024
allenai/reward-bench ★379 — `S` — *Benchmark for reward modeling* 2024
allenai/OLMo-7B-Instruct ★54 — `M` — *A truly open-source chat model* 2024
allenai/OLMo-7B ★642 — `M` — *Our first open-source language model* 2024
allenai/tulu-2-dpo-70b ★157 — `M` — *First model scaling DPO* 2023

HuggingFaceH4/zephyr-7b-beta ★1.7k+ — M — *Small and powerful chat model v2* 2023
HuggingFaceH4/zephyr-7b-alpha ★1.1k+ — M — *Small and powerful chat model* 2023
HuggingFaceH4/starchat-playground ★800 — S — *Playground for coding assistant models* 2023
HuggingFaceH4/starchat-alpha ★232 — M — *Coding assistant language model* 2023
HuggingFaceH4/open-llm-leaderboard ★13.1k+ — S — *Leaderboard for open instruction-tuned LLMs* 2023
HuggingFaceH4/stack-exchange-preferences ★132 — D — *Preference dataset for RLHF* 2023

Invited Talks

2025

1. Reasoning; What comes next? – VentureBeat Transform
2. The art of a good (reasoning) model – Enterprise AI Agent Summit, Seattle WA (slides, recording)
3. A taxonomy for next-generation reasoning models – AI Engineer World's Fair (slides, recording)
4. Experiments with Reinforcement Learning with Verifiable Rewards – USC Information Sciences Institute (slides, recording)
5. The RL Era of Language Models – UC Santa Cruz, Silicon Valley Extension (slides, recording)
6. An unexpected RL renaissance – Minds and Machines Seminar, Seattle (slides, recording)

2024

7. (Tutorial) Language Modeling – Neural Information Processing Systems (slides, recording)
8. How to approach post-training for AI applications – Infer AI Engineer Vancouver (slides, recording)
9. The state of reasoning – Latent Space @ NeurIPS (slides, recording)
10. Tülu 3 Preview – Princeton AI Alignment and Safety Seminar (PASS) (slides, recording)
11. RewardBench, Evaluating Reward Models for Language Modeling – Deep Learning Classics and Trends (slides)
12. RewardBench – Anthropic Societal Impacts Team

2023

13. History and Risks of RLHF – Workshop of Social Choice for AI Ethics and Safety
14. History and Risks of RLHF – Workshop on Sociotechnical AI Safety (slides)
15. Bridging RLHF from LLMs back to control – CoRL LangRob Workshop (slides,recording)
16. Objective Mismatch in Reinforcement Learning from Human Feedback – Deep Learning Classics and Trends (Slides Available, Recording Available)
17. (Tutorial) Reinforcement Learning from Human Feedback – International Conference on Machine Learning (Slides Available, Recording Available)
18. (Tutorial) Steering language models with reinforcement learning from human feedback and constitutional AI – ACM Conference on Fairness, Accountability, and Transparency (Slides Available)
19. Reinforcement Learning from Human Feedback; Open and Academic Progress – UCL Dark Lab (Recording Available, Slides)
20. Reinforcement Learning from Human Feedback; Pathways to Open Reproduction of ChatGPT – Microsoft Data Science Gems

2022

21. Reward Reports for Reinforcement Learning – ICML Workshop on Responsible Decision Making in Dynamic Environments (Recording Available, Slides)
22. Synthesizing Robotic Controllers with Model-based Reinforcement Learning – Lead The Future
23. Planning through Exploration and Exploitation in Model-based Reinforcement Learning – University of Pennsylvania – Perception, Action, and Learning Group (Recording Available, Slides)
24. (Job Talk) Legible Reinforcement Learning via Dynamics Models – Microsoft Research (Slides)
25. The Challenges of Exploration for Offline Reinforcement Learning – DeepMind Robotics All Hands
26. (Job Talk) Synergy of Prediction and Control in Model-based Reinforcement Learning – Amazon Robotics & AI (Slides)

2021

27. (Job Talk) Control-oriented Predictions in Model-based Reinforcement Learning – Cruise AI
28. Improving Model Predictive Control in Model-based Reinforcement Learning – Cornell Robotics Seminar (Recording Available, Slides)

2020

29. Model Learning for Low-level Control in Robotics – UC Berkeley Semiautonomous Seminar (Recording Available, Slides)

2019

30. – UC Berkeley Semiautonomous Seminar

Mentorship

Daniel Marta (Thesis Committee Member)	2025
Saumya Malik (Ai2 Pre-Doctoral Researcher '25)	2024
Jacob Morrison (Ai2 Pre-Doctoral Researcher '25)	2024
Mark Selden (UC Berkeley BS '22)	2020
Albert Wilcox (UC Berkeley BS '22, Ph.D. student at Georgia Tech)	2019
Jason Zhou (UC Berkeley BS, MS '21 to Matician)	2019
Felipe Campos (UC Berkeley BS '20 to Armstrong Robotics)	2018
Howard Zhang (UC Berkeley BS, MS'21 to UCLA PhD)	2018

Peer Review

Association for Computational Linguistics (ACL) ARR	2025 (1)
Empirical Methods in Natural Language Processing, Demo Track	2024
Transactions on Machine Learning Research (TMLR)	2024
Conference on Machine Learning (ICML) (count is 1 unless labelled)	2020, 2022 (3), 2023, 2024 (3)
Conference on Neural Information Processing Systems (NeurIPs)	2022 (2), 2023, 2024, 2025 (3)
Conference on Learning Representations (ICLR) (*Outstanding Reviewer)	2021* (3), 2022 (3)
Conference on Robot Learning (CORL)	2020
Conference on Robotics and Automation (ICRA)	2020, 2021, 2022 (2)
Conference on Intelligent Robots and Systems (IROS)	2021, 2022 (2)
Robotics - Science and Systems (RSS)	2022
Conference on Decision and Control (CDC)	2021
Robotics and Automation Letters (RA-L)	2019, 2020, 2022
Transactions on Cybernetics	2019, 2020

Professional Activities

Area Chair (AC), Conference on Language Modeling (COLM)	2025
Associate Editor (AE), Conference on Intelligent Robots and Systems (IROS)	2023
Farama Foundation Board of Technical Advisors	2022 – 2024
NeurIPs Workshop on Robot Learning Organizer	2021, 2022
Member of Well-Being in Machine Learning	2021 – 2024
RLDM Workshop on Building Accountable and Transparent RL Organizer	2022
Berkeley AI Research Audio & Video Team	2021 – 2022
Machine Learning Collective Office Hours	2021 – cont.
Tapia Panel on Student Mental Health Organizor	2021
Founder of UC Berkeley EECS Equal Access to Application Assistance (EAAA) Program	2020 – 2022
Wellness Coordinator for UC Berkeley Electrical Engineering Graduate Student Assembly (EEGSA)	2020 – 2022
Bay Area Teachers in Schools	2017

Policy Engagement

Partnership on AI Partner Forum, Panel on AI Agents	2024
Columbia Convening on Openness and AI	2024
NAIRR Panel at AI Expo for National Competitiveness	2024
IAS AI Policy and Governance Working Group	2024

Teaching

Advances and Challenges in Language Models, Reasoning, and AI Agents (U.W. CSE599H), Guest Lecture, An introduction to post-training *(slides, course)* Sp2025
NLP with Deep Learning (Stanford University CS224N), Guest Lecture, Life after DPO *(slides, course, recording)* Sp2024
Transformers United (Stanford University CS25N), Guest Lecture, hi story of open alignment *(slides, recording)* Sp2024
Policy Challenges of AI (Harvard University), Guest lecture on open-source AI *(slides, course)* Sp2024
Machine Learning from Human Preferences (Stanford University CS329H), Guest lecture on RLHF *(slides, course)* Fa2023
AI History (Seoul National University, School of Law), Guest lecture on RL History *(course)* Fa2023
HuggingFace Deep RL Course (RLHF, from 0 to ChatGPT), Online (¿160k views) *(watch)* Win.2022
Introduction to Artificial Intelligence (UCB CS188), TA Su2020, Fa2020
Introduction to Artificial Intelligence (UCB CS188), Instructor *lectured to 800+ students* Sp2020
Designing Information Devices and Systems II (UCB EE16B), TA Fa2019
Integrated Micro Sensors and Actuators (Cornell ECE4320), Grader Sp2017
Mathematics of Signal and System Analysis (Cornell ECE 3250), TA Fa2016

Extracurriculars

The Retort AI Podcast 2023 – cont.
Interconnects AI Blog 2020 – cont.
Cornell Varsity Lightweight Rowing 2013 – 2017
Novice Rowing Coach 2017 – 2018
Graduates for Engaged and Extended Scholarship in Computing and Engineering President 2021 – 2022