(Last Updated: December 2025)

Gokul Swamy

Contact

gswamy@cmu.edu
https://gokul.dev

Education

Carnegie Mellon University

Ph.D. in Robotics (GPA: 4.04/4.00) **Sept. 2020 - Present**
- ▷ Thesis Committee: J. Andrew Bagnell, Zhiwei Steven Wu, Geoffrey J. Gordon, Arthur L. Gretton
- ▷ Publications: [Google Scholar]

University of California, Berkeley

M.S. in Computer Science **Aug. 2019 - May 2020**
- ▷ Thesis: [Learning with Humans in the Loop]

B.S. in Electrical Engineering and Computer Science, High Honors **Aug. 2016 - May 2019**

Research Experience

Robotics Institute @ CMU, Pittsburgh, PA
Graduate Student Researcher **Sep. 2020 - Present**
Collaborating with Profs. Drew Bagnell and Steven Wu on *efficient algorithms for interactive learning*. Currently researching imitation learning (IL) and reinforcement learning from human feedback (RLHF) with applications to robotics and large language models. Published at ICML 2021, 2022, 2023, 2024 (4x), NeurIPS 2022 (2x), 2023, 2024 (3x), 2025 (2x), ICLR 2025 (3x), RSS 2025, and RLC 2025.

Berkeley Artificial Intelligence Research Lab, Berkeley, CA
Graduate Student Researcher **Jan. 2018 - May 2020**
Collaborated with Prof. Anca Dragan on comparing increasingly structured models of human drivers as modeling assumptions are broken and with Anca Dragan and Sergey Levine on allowing one person to supervise and provide corrections to a fleet of learning robots. Published at HRI 2019, ICRA 2020.

Professional Experience

Google Research, Seattle, WA
Student Researcher **May 2023 - Nov. 2023**
Collaborated with Drs. Alekh Agarwal, Chris Dann, and Rahul Kidambi on game-theoretic algorithms for reinforcement learning from conflicting human feedback for preference fine-tuning of large language models.

Microsoft Research, Montreal, CA
Graduate Research Intern **June 2022 - Aug. 2022**
Collaborated with Prof. Geoff Gordon on learning factorized dynamics models from visual observations. Investigating game-theoretic methods for learning structured latent spaces and sequence model architectures.

Aurora, Pittsburgh, PA
Motion Planning ML Intern **May 2020 - Aug. 2020**
Collaborated with Prof. Sanjiban Choudhury on learning deep driving policies that respected safety constraints (e.g. avoiding cyclists). Built C++ data pipelines / simulation tools and implemented constrained training of deep networks.

NVIDIA, Santa Clara, CA
Autonomous Vehicles Perception Intern **May 2019 - Aug. 2019**
Collaborated with Dr. Trung Pham on single-image weakly-supervised 3D structure estimation of intersection entry/exit lines. Designed CNN to recover lines in 3D that learned from 2D key points and 3D geometric constraints. Produced significant improvement over existing predict-then-project method.

SpaceX, Hawthorne, CA
Data Engineering Intern **May 2018 - Aug. 2018**
Collaborated with Dr. Anthony Rose on estimating shop-floor operation durations. Used hierarchical navigable small world graphs on top of word2vec to build approximate nearest neighbors engine that significantly outperformed domain experts. Created classical computer vision algorithm to detect flight risks.

Intuit, San Diego, CA
Software Engineering Intern **May 2017 - Aug. 2017**
Worked on using CNNs to classify the sentiments of product reviews. Created themable iOS UI framework and Java/PHP/React-based internal tools.

Teaching Experience

Carnegie Mellon University, Pittsburgh, PA

Course Designer & Co-Instructor **Jan. 2025 - May. 2025**

Creator and co-instructor for new course: *17-740: Algorithmic Foundations of Interactive Learning* with Drew Bagnell and Steven Wu. Covering online learning, game-solving, RL, imitation, and RLHF. [Site]

Teaching Assistant **Sept. 2024 - Dec. 2024**

TA for *15-889: Advanced Topics in Reinforcement Learning* w/ Prof. Aviral Kumar. Covered model-free, model-based, inverse, and offline RL. Gave lectures on imitation learning / RL theory. [Site]

Teaching Assistant **Sept. 2022 - Dec. 2022**

TA for *10-732: Robustness and Adaptation in Shifting Environments* w/ Prof. Zachary Lipton. Covered causally-structured and feedback-driven distribution shifts, online adaptation, and adv. robustness. [Site]

University of California, Berkeley, Berkeley, CA

Teaching Assistant **Jan. 2019 - Dec. 2019**

Helped teach CS 188: Introduction to AI and CS 189: Introduction to Machine Learning. Gave guest lecture and created worksheet walkthrough videos.

Course Instructor **Aug. 2018 - May 2019**

Created material for and taught student course twice on societal impacts and ethical considerations of AI, covering topics like automation, bias in AI, data privacy, synthetic data, and human-compatible AI.

Activities

Graduate Application Mentor, Queer in AI 2024–

Graduate Application Aid Program Reviewer, Queer in AI 2024–

Communications Chair, Reinforcement Learning Conference (RLC) 2024

Workshop Organizer, Reinforcement Learning Beyond Rewards @ RLC 2024 2024

Future Faculty Program Participant, Eberly Center @ CMU 2022–

Teaching & Learning Summit Advisory Board, Eberly Center @ CMU 2022

Undergraduate Research Engagement Working Group, CMU SCS Dean's Advisory Committee 2021–

▷ Put together interactive workshop for graduate students on best practices for mentoring undergraduates on research projects. Conducted IRB-approved study to measure training effects. Results selected for presentation at Eberly Teaching Summit. [Workshop Slides and Results Poster]

Graduate Application Support Program, CMU Robotics Institute, Mentor 2021–2022

Graduate Application Support Program, CMU Robotics Institute, Organizer 2021

Undergraduate AI Mentoring Program, CMU 2021

Journal Reviewer: IEEE Transactions on Robotics (IEEE T-RO), IEEE Transactions on Information Theory (IEEE T-IT), International Journal of Robotics Research (IJRR), Transactions on ML Research (TMLR)

Conference Reviewer: NeurIPS (2021, 2022, 2023), ICRA (2022, 2024), ICML (2023, 2024, 2025), ICLR (2024, 2026), IROS 2024, COLT 2025

Workshop Reviewer: Strategic ML @ NeurIPS 2021, Real World RL @ NeurIPS 2022, Interactive Learning with Implicit Human Feedback @ ICML 2023, Frontiers for Learning, Dynamics, and Control @ ICML 2023, Robot Learning Workshop @ NeurIPS 2023, Aligning RL Experimentalists and Theorists @ NeurIPS 2025

Mentorship

Andrew Z. Li (CMU, MS), Compute-optimal learning to search for VLAs. 2025–

Hyun Joe Jeong (CMU, PhD), Learning to search in language space for VLAs. 2025–

Dhruv Sreenivas (MILA, PhD), Steering diffusion policies via latent-space inverse RL. 2025–

Lujing Zhang (CMU, PhD), Self-play algorithms for multi-criteria LLM alignment. 2025–

Jiahao Zhang (CMU, PhD), Self-play algorithms for multi-criteria LLM alignment. 2025–

Xiaoyuan Zhu (USC, UG), Optimizing justifications provided for LLM reasoning. 2025–

Lintang Sutawika (CMU, PhD), Self-play algorithms for low-resource language reasoning. 2025–

Daman Arora (CMU, PhD), Learning to parallelize search algorithms for efficient LLM reasoning. 2025–

Nicolas Espinosa Dice (Cornell, PhD), Efficient imitation under misspecification. 2024–

▷ 2025 NSF GRF.

Yilin Wu (CMU, PhD), Learning to search with VLM verifiers. 2024–2025

Arnav Kumar Jain (MILA, PhD), Learning to search from demonstrations. 2024–2025

Vibhakar Mohta (CMU, MS), Learning to search from demonstrations. (next: Nuro) 2023–2025

Abigail DeFranco (CMU, UG), Confidence-aware classifier-based guidance. (next: Startup) 2024–2025

Juntao Ren (Cornell, UG), Hybrid algorithms for inverse RL. (next: Stanford PhD) 2023–2025

▷ Knight-Hennessy Scholar, 2025 NSF GRF, Runner Up for 2025 CRA OURA.

Jinwgu Tang (CMU, PhD), Multi-agent imitation learning. 2023–2024

Silvia Sapora (Oxford, PhD), Faster transfer in inverse reinforcement learning via evolution. 2023–2024

Konwoo Kim (CMU, UG), Learning safety constraints from demos. (next: Stanford PhD) 2022–2023

▷ 2025 NSF GRF, 2023 CMU SCS Allen Newell Award for Excellence in Undergraduate Research.

Matthew Peng (Berkeley, UG), Minimax-optimal inverse RL. (next: Applied Intuition) 2021–2022

Honors

(Inaugural) Outstanding Graduate Teaching Assistant Award, CMU Robotics Institute — 2025
- ▷ Awarded for "*co-creation of a new course ... and dedication to student mentoring.*"

Rising Star, Northeast Robotics Colloquium @ Cornell University — 2025

Rising Star in Data Science, Stanford University — 2025

Outstanding Paper Award, ICLR 2025 World Models Workshop — 2025

Invited to Jane Street Graduate Research Fellowship Research Workshop — 2024, 2025

Best Reviewer Award, ICML — 2024

Top Reviewer, NeurIPS — 2022

Finalist for NVIDIA Graduate Research Fellowship, received GPU award — 2021

NSF GRFP, Honorable Mention — 2020

Press Coverage

A Smooth Sea Never Made a Skilled SAILOR
- ▷ RoboPapers Podcast, 2025. [Link]

Inverse Reinforcement Learning Without Reinforcement Learning
- ▷ TWiML AI Podcast, 2023. [Link]

Conference Publications

Arnav Kumar Jain*, Vibhakar Mohta*, Subin Kim, Atiksh Bhardwaj, Juntao Ren, Yunhai Feng, Sanjiban Choudhury, **Gokul Swamy**, *A Smooth Sea Never Made a Skilled SAILOR: Robust Imitation via Learning to Search*. **Spotlight Presentation (3.2%)**, Neural Info. Processing Symposium (NeurIPS), 2025. [Site]

Nicolas Espinosa-Dice, Yiyi Zhang, Yiding Chen, Bradley Guo, Owen Oertell, **Gokul Swamy**, Kiante Brantley, Wen Sun, *Scaling Offline RL via Efficient and Expressive Shortcut Models*. Neural Information Processing Symposium (NeurIPS), 2025. [PDF]

Mohamad Qadri, **Gokul Swamy**, Jonathan Francis, Michael Kaess, Andrea Bajcsy, *Your Learned Constraint is Secretly a Backwards Reachable Tube*, Reinforcement Learning Conference (RLC), 2025. [PDF]

Yilin Wu, Ran Tian, **Gokul Swamy**, Andrea Bajcsy, *From Foresight to Forethought: VLM-In-the-Loop Policy Steering via Latent Alignment*. Robotics Science & Systems (RSS), 2025. [Site]

Runzhe Wu, Yiding Chen, **Gokul Swamy**, Kiante Brantley, Wen Sun, *Diffusing States and Matching Scores: A New Framework for Imitation Learning*. International Conference on Learning Representations (ICLR), 2025. [PDF]

Zhaolin Gao, Wenhao Zhan, Jonathan D. Chang, **Gokul Swamy**, Kianté Brantley, Jason D. Lee, Wen Sun, *Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLHF*. International Conference on Learning Representations (ICLR), 2025. [PDF]

Nicolas Espinosa Dice, Sanjiban Choudhury, Wen Sun, **Gokul Swamy**, *Efficient Imitation Under Misspecification*. International Conference on Learning Representations (ICLR), 2025. [PDF]

Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, **Gokul Swamy**, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, Wen Sun, *REBEL: RL via Regressing Relative Rewards*. Neural Information Processing Symposium (NeurIPS), 2024. [PDF]

Yuda Song, **Gokul Swamy**, Aarti Singh, J. Andrew Bagnell, Wen Sun, *The Importance of Online Data: Understanding Preference Fine-tuning via Coverage*. Neural Information Processing Symposium (NeurIPS), 2024. [PDF]

Jingwu Tang, **Gokul Swamy**, Fei Fang, Zhiwei Steven Wu, *Multi-Agent Imitation Learning: Value is Easy, Regret is Hard*. Neural Information Processing Symposium (NeurIPS), 2024. [PDF]

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, Alekh Agarwal, *A Minimaximalist Approach to Reinforcement Learning from Human Feedback*. Internat. Conf. on ML (ICML), 2024. [Site]

Juntao Ren*, **Gokul Swamy***, Zhiwei Steven Wu, J. Andrew Bagnell, Sanjiban Choudhury, *Hybrid Inverse Reinforcement Learning*. International Conference on ML (ICML), 2024. [Site]

Silvia Sapora, **Gokul Swamy**, Chris Lu, Yee Whye Teh, Jakob Nicolaus Foerster, *EvIL: Evolution Strategies for Generalisable Imitation Learning*, Internat. Conf. on ML (ICML), 2024. [PDF]

My Phan, Kianté Brantley*, Stephanie Milani*, Soroush Mehri*, **Gokul Swamy***, Geoffrey J. Gordon, *When is Transfer Learning Possible?*, Internat. Conf. on ML (ICML), 2024. [PDF]

Konwoo Kim*, **Gokul Swamy***, Zuxin Liu, Ding Zhao, Sanjiban Choudhury, Zhiwei Steven Wu, *Learning Shared Safety Constraints from Multi-task Demonstrations*, Neural Information Processing Symposium (NeurIPS), 2023. [Site]

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, *Inverse Reinforcement Learning without Reinforcement Learning*, Internat. Conf. on ML (ICML), 2023. [Site]

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, *Sequence Model Imitation Learning with Unobserved Contexts*, Neural Information Processing Symposium (NeurIPS), 2022. [Site]

Gokul Swamy*, Nived Rajaraman*, Matt Peng, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, Jiantao Jiao, Kannan Ramchandran, *Mimimax Optimal Imitation Learning via Replay Estimation*, Neural Information Processing Symposium (NeurIPS), 2022. [Site]

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, *Causal Imitation Learning under Temporally Correlated Noise*, **Oral Presentation (2.1%)**, Int. Conf. on ML (ICML), 2022. [Site]

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, *Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap*, Internat. Conf. on ML (ICML), 2021. [Site]

Gokul Swamy, Siddharth Reddy, Sergey Levine, Anca D. Dragan, *Scaled Autonomy: Enabling Human Operators to Control Robot Fleets*, International Conf. on Robotics and Automation (ICRA), 2020. [PDF]

Gokul Swamy, Jens Schulz, Rohan Choudhury, Dylan Hadfield-Menell, Anca D. Dragan, *On the Utility of Model Learning in HRI*, International Conf. on Human-Robot Interaction (HRI), 2019. [PDF]

Workshop Papers

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, J. Andrew Bagnell, *All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning*, Aligning RL Theorists and Experimentalists Workshop @ NeurIPS 2025. [PDF]

Owen Oertell, Wenhao Zhan, **Gokul Swamy**, Zhiwei Steven Wu, Kianté Brantley, Jason D. Lee, Wen Sun, *Heuristics Considered Harmful: RL With Random Rewards Should Not Make LLMs Reason*, NY RL Workshop 2025. [Link]

Jiahao Zhang*, Lujing Zhang*, Keltin Grimes, **Gokul Swamy**, Zhiwei Steven Wu. *Back to Blackwell: Closing the Loop on Intransitivity in Multi-Objective Preference Fine-Tuning.* NY RL Workshop 2025.

Naicheng He*, Arjun Prakash*, **Gokul Swamy**, Eugene Vinitsky, Amy Greenwald. *Inverse Reinforcement Learning on GPUDrive.* NY RL Workshop 2025.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, *Complementing a Policy with a Different Observation Space*, Interactive Learning with Implicit Human Feedback, Spurious Correlations, Invariances, and Stability Workshops @ ICML 2023. [PDF]

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Zhiwei Steven Wu, *Game Theoretic Algorithms for Conditional Moment Matching*, Neglected Assumptions in Causal Inference @ ICML 2021. [PDF]

Pre-prints

Denizalp Goktas, Amy Greenwald, Takayuki Osogami, Roma Patel, Kevin Leyton-Brown, Grant Schoenebeck, ..., **Gokul Swamy**, ..., Quanyan Zhu, *Strategic Foundation Models.* [PDF]

Talks

Beyond Behavioral Cloning: Robust Imitation via Learning to Search
- ▷ Sidd Srinivasa's Group @ UW, 2025

Efficient Algorithms for Interactive Learning: Teaching Robots to Focus
- ▷ Rising Star Talk, Northeast Robotics Colloquium @ Cornell, 2025
- ▷ Anirudha Majumdar's Group @ Princeton, 2025

All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning
- ▷ Rising Stars in Data Science @ Stanford, 2025
- ▷ NY RL Day @ Amazon, 2025
- ▷ Kempner Institute @ Harvard, 2025
- ▷ Andreea Bobu's Group @ MIT, 2025
- ▷ AI Seminar @ Cornell, 2025
- ▷ Andrea Zanette's Group @ CMU, 2025
- ▷ Ameet Talwalkar / Virginia Smith's Group @ CMU, 2025
- ▷ Google DeepMind Robotics Reading Group, 2025

Efficient Reductions for Reinforcement Learning from Human Feedback
- ▷ INFORMS Annual Meeting, 2024
- ▷ Jakob Foerster's Group @ Oxford, 2024
- ▷ Foundations Team @ Google DeepMind, 2024
- ▷ Adaptive Learning in Complex Environments @ TTIC, 2024
- ▷ Emma Brunskill's Group @ Stanford, 2024
- ▷ Learning Theory Team @ Google Research, 2024
- ▷ NLP Lunch @ Google Research, 2024

Efficient Reductions for Interactive Learning from Implicit Feedback
- ▷ Control and Learning Seminar @ CMU, 2024
- ▷ Dylan Hadfield-Menell's Group @ MIT, 2024
- ▷ Northeast Systems and Control Workshop, 2024

Efficient Reductions for Inverse Reinforcement Learning
- ▷ Jaime Fisac's Group @ Princeton, 2024
- ▷ Lerrel Pinto's Group @ NYU, 2024
- ▷ Coordinated Science Laboratory Student Conference @ UIUC, 2024
- ▷ Robotics Seminar @ Cornell, 2024

REBEL: Reinforcement Learning by Regressing Relative Rewards
- ▷ Oral Presentation at Aligning RL Experimentalists and Theorists Workshop @ ICML 2024

Causal Confounds in Imitation Learning
- ▷ Causality for Robotics Workshop @ IROS, 2023

Efficient Algorithms for Interactive Imitation Learning
- ▷ Jonathan Shock's Group @ University of Cape Town, 2023

Learning Shared Safety Constraints from Multi-task Demonstrations
- ▷ Oral Presentation at the Adversarial ML Workshop @ ICML 2023

An Interactive Workshop on Undergraduate Research Mentorship for Graduate Students
- ▷ Eberly Center Teaching & Learning Summit, 2022

Learning Modular World Models
- ▷ MSR Montreal, 2022

On Interaction, Imitation and Causation
- ▷ Zachary Lipton's Group @ CMU, 2023
- ▷ Scott Niekum's Group @ UT Austin, 2022
- ▷ Reinforcement Learning Discussion Group @ MSR-NYC, 2022
- ▷ Dave Held's Group @ CMU, 2022

Causal Imitation Learning under Temporally Correlated Noise
- ▷ Long Talk at ICML 2022
- ▷ Oral Presentations at Causal Sequential Decision Making, Offline RL, and Safe and Robust Control of Uncertain Systems Workshops @ NeurIPS 2021

Of Moments and Matching: A Game Theoretic Framework for Closing the Imitation Gap
- ▷ Dave Held's Group @ CMU, 2021

Guest Lectures

Beyond Behavioral Cloning: Robust Imitation via Learning to Search
▷ Reinforcement Learning @ NTU, 2025

An Introduction to RLHF
▷ Foundations of Autonomous Decision Making under Uncertainty @ CMU, 2024

Efficient Algorithms for Inverse RL
▷ Reinforcement Learning @ NTU, 2024

Policy and Q-value Iteration with Function Approximation
▷ 2x, Robot Learning @ Cornell, 2024

An Introduction to Imitation
▷ Advanced RL @ CMU, 2024

On Interaction, Imitation and Causation
▷ Learning for Robot Decision Making @ Cornell, 2022

Leveraging Human Input for Training Self-Driving Cars
▷ Human-AI Interaction @ CMU, 2020, 2022