

Kihyun Kim

+1-617-949-1247 | kihyun@mit.edu | kihyun.xyz | [Google Scholar](https://scholar.google.com/citations?user=...)

Research Interests

AI Alignment (RLHF, LLM post-training, social choice theory), **Reinforcement Learning** (offline RL, inverse RL, imitation learning), **Control Theory** (optimal control, robust control)

Education

Massachusetts Institute of Technology

Cambridge, United States

Ph.D. Program in Electrical Engineering & Computer Science

Sep. 2021 - Current

- M.S., 2024 — Thesis ([link](#))
- Advisor: Prof. Asuman Ozdaglar, Prof. Pablo Parrilo

Seoul National University

Seoul, Republic of Korea

B.S. in Electrical and Computer Engineering

Mar. 2014 - Aug. 2020

- Graduated with *Summa Cum Laude*
- Paused for two years to fulfill military duty (2016 - 2018)

Seoul Science High School

Seoul, Republic of Korea

High school for gifted students in science and mathematics

Mar. 2011 - Feb. 2014

Publications

[Preprint] Inverse Reinforcement Learning without an Optimal Demonstrator: A Feasible Reward Set Approach

Kihyun Kim, Shripad Deshmukh, Nikos Vlassis, Jiawei Zhang

arXiv preprint arXiv:2605.30903 (Under Review)

[1] Beyond RLHF and NLHF: Population-Proportional Alignment under an Axiomatic Framework

Kihyun Kim, Jiawei Zhang, Asuman Ozdaglar, Pablo Parrilo

International Conference on Learning Representations (ICLR), 2026

[2] A Unified Linear Programming Framework for Offline Reward Learning from Human Demonstrations and Feedback

Kihyun Kim, Jiawei Zhang, Asuman Ozdaglar, Pablo Parrilo

International Conference on Machine Learning (ICML), 2024

[3] Distributional robustness in minimax linear quadratic control with Wasserstein distance

Kihyun Kim, Insoon Yang

SIAM Journal on Control and Optimization (SICON), 2023

[4] Minimax control of ambiguous linear stochastic systems using the Wasserstein metric

Kihyun Kim, Insoon Yang

IEEE Conference on Decision and Control (CDC), 2020

[5] Optimizing large-scale fleet management on a road network using multi-agent deep reinforcement learning with graph neural network

Juhyeon Kim, **Kihyun Kim**

IEEE International Intelligent Transportation Systems Conference (ITSC), 2021

[6] Generative autoregressive networks for 3d dancing move synthesis from music

Hyemin Ahn, Jaehun Kim, **Kihyun Kim**, Songhwa Oh

IEEE Robotics and Automation Letters (RA-L), 2020

Research Experience

Laboratory for Information & Decision Systems (LIDS)

MIT

Advisor: Prof. Asuman Ozdaglar, Prof. Pablo Parrilo

Sep. 2021 - Present

- Research Focus: AI Alignment, Reward Learning
- Proposed a novel linear programming (LP) framework for offline reward learning (Inverse RL and RLHF) that estimates the reward function from expert demonstrations by effectively addressing the data coverage issue
- Developed a population-proportional preference learning algorithm inspired by social choice theory to improve fairness and robustness under diverse human preferences

Control and Optimization Research Lab

Seoul National University

Advisor: Prof. Insoon Yang

Sep. 2019 - Aug. 2021

- Research Focus: Stochastic optimal control, Distributionally robust optimization
- Developed a novel minimax linear-quadratic control method using the Wasserstein metric, which is robust to the unknown distribution of system parameters
- Suggested a theoretical connection between the classical H-infinity controller and the modern distributionally robust optimization technique with the Wasserstein ambiguity set

Robot Learning Lab

Seoul National University

Advisor: Prof. Songhwai Oh

Jun. 2019 - Aug. 2019

- Research Focus: Robot learning, Humanoid robot, Generative model
- Developed an experimental program for a real humanoid robot using ROS to evaluate motion sequences generated from deep neural network models

Work & Teaching Experience

Research Intern

Adobe Research

Summer Internship, Adobe Research

May. 2025 - Aug. 2025

- Developed a painting environment integrated with Adobe's AI agentic framework as a testbed for RL with LLM agents
- Developed an offline inverse RL algorithm addressing reward verification challenges in multi-step tool interactions

Research Intern

LG AI Research

Summer Internship, Advanced ML Lab at LG AI Research

Jun. 2024 - Sep. 2024

- Proposed and evaluated an RLHF framework that incorporates confidence levels into human feedback.

Teaching Assistant

MIT

6.7920: Reinforcement Learning: Foundations and Methods

Sep. 2024 - Dec. 2024

- Graduate-level reinforcement learning course (Instructors: Prof. Cathy Wu, Prof. Munther Dahleh)

Digital Signal Processing (DSP) Engineer

Republic of Korea

SEC Signals Laboratory, Republic of Korea Army

Dec. 2016 - Sep. 2018

- Specialized in detection and demodulation of digital signals

Mathematical Olympiad Instructor

Republic of Korea

Privatenote Co.

Aug. 2015 - Feb. 2016

- Led online courses for students preparing for the national Mathematical Olympiad
- Courses covered: Number Theory, Algebra, Geometry

Honors & Awards

2021 - 2026 **KFAS Doctoral Study Abroad Fellowship**, Korea Foundation for Advanced Studies

2021 - 2022 **Alan V. Oppenheim Fellowship**, MIT EECS

2014 - 2020 **Seoam Undergraduate Scholarship**, Seoam Yoon Se Young Foundation

2019 **Kwon Oh-hyun Scholarship**, Former CEO of Samsung Electronics & SNU ECE Alumni Association

2015 **6th Place (Special Prize)**, ACM International Collegiate Programming Contest Korea Regional