JOE KWON

(+1) 203-809-9460 ⋄ joekwon@csail.mit.edu

EDUCATION

Yale University *August 2016 - May 2021*
B.A. in Computer Science and Psychology.

Canada/USA Mathcamp *July 2016 - August 2016*
Among 65 students accepted from global applicant pool. Received full program scholarship.

PAPERS

Schoenegger et al (large project) (Under Review). Large Language Models Are More Persuasive Than Incentivized Human Persuaders.

Kwon, J.,Wong, L., & Levine, S. (2025). What Do Moral Rules Mean? *Society for Philosophy and Psychology*

Brumley, M., Kwon, J., Krueger, D., Krasheninnikov & Anwar, U. (2024). Comparing Bottom-Up and Top-Down Steering Approaches on In-Context Learning Tasks. *NeurIPS MINT Workshop on Foundation Model Interventions*

Kwon, J., Levine, S., & Tenenbaum, J. (2024). Neuro-Symbolic Models of Human Moral Judgment *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Huang, B., Kwon, J. (2023). Does It Know?: Probing for Uncertainty in Language Model Latent Beliefs. *NeurIPS Attributing Model Behavior at Scale*

Casper, S., Lin, J., Kwon, J., Culp, G., & Hadfield-Menell, D. (2023). Explore, Establish, Exploit: Red Teaming Language Models from Scratch.

Kwon, J., Tan, Z-X., Tenenbaum, J., & Levine, S. (2023) When it's not out of line to get out of line: Principles of universalizability, welfare, and harm. *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Kwon, J. (2022) Values Shape Optimizers Shape Values. *NeurIPS Human-Centered AI Workshop*

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., & Hendrycks, D. (2022). Forecasting Future World Events With Neural Networks. *NeurIPS.*

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., & Steinhardt, J. (2022). Scaling Out-of-Distribution Detection for Real-World Settings. *ICML*

Kwon, J., Tenenbaum, J., & Levine, S. (2022). Flexibility in Moral Cognition: When is it okay to break the rules? *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2022). Social inferences from physical evidence via Bayesian event reconstruction. *Journal of Experimental Psychology: General.*

Lopez-Brau, M., Kwon, J., McBean, B., Yildirim, I., & Jara-Ettinger, J. (2021). Detecting the involvement of agents through physical reasoning. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.*

Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2020). Mental state inference from indirect evidence through Bayesian event reconstruction. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Zhang, M., Tseng, C., Montejo, K., Kwon, J., Kreiman, G. (2019) Lift-the-flap: what, where and when for context reasoning

RESEARCH EXPERIENCE

Constellation *Starting Jan 2026*
Astra Fellow

· Threat modeling for secretly loyal AI and generating shovel-ready empirical ML research advised by Tom Davidson and Fabien Roger. Generally interested in threat modeling for risks from internal deployment (sabotage, concentration of power, ...).

Centre for the Governance of AI *September 2025 - December 2025*
Fall DC Fellow

· Research investigating regulatory gaps in addressing risks from internal AI deployment with Stephen Casper and metrics for tracking the state of automated AI R&D with Ranay Padarath, Hilary Greaves, Alan Chan, and Markus Anderljung

Center for AI Policy *January 2025 - June 2025*
Technical Policy Analyst

· Supporting legislative drafting, distilling literature on AI Safety and Security, policy research with eye on model evaluations and AI Agents.

Krueger Lab *May 2024 - November 2024*
Research Intern

· Stress-Testing Representation Engineering and evaluating steering vectors. Advised by Usman Anwar, Dima Krasheninnikov, David Krueger

LG AI Research *June 2023 - May 2024*
Research Engineer Intern

· Working on improving cross-lingual capabilities in LLMs. Work involved synthetic data generation, extended pretraining, finetuning, and model evaluations. Advised by Lajanugen Logeswaran, Dongsub Shim, Honglak Lee

Computational Cognitive Science Lab *September 2021 - Present*
Research Assistant

· Examining flexible generalizations of moral rules and judgments in humans, and engineering moral machines. Advisors: Sydney Levine, Yejin Choi, and Josh Tenenbaum.
· Constructing an automated pipeline for red-teaming large language models from scratch. Work with Stephen Casper; project advised by Dylan Hadfield-Menell.

Aligned AI *June 2022 - September 2022*
Researcher

· ACE mitigates simplicity bias. Ran experiments on image classification, exploring key obstacles in learning diverse features. Emphasis on how the relative complexity of the various generalizations affects diversification quality. Advised by Dr. Stuart Armstrong

Berkeley AI *June 2021 - May 2022*
Research Assistant

· Built a benchmark for ML forecasting and ran empirical experiments for baseline demonstration of transformer based algorithm.
· Built a new dataset for OOD detection. Advised by Jacob Steinhardt and Dan Hendrycks.

Computational Social Cognition Lab *January 2019 - May 2021*
Research Assistant

· Worked on computational models of social inferences from physical events: 1) How do we know if an agent was previously present? 2) How do we infer their mental states?). Advised by Julian Jara-Ettinger and Michael Lopez-Brau.

AI Safety Camp *February 2021 - June 2021*
Research Collaborator

· Worked with my team to learn a generative model (a VAE) on an RL agent and environment states, for visualization of internal states. Team led by Lee Sharkey.

Computational Affective and Social Cognition Lab *May 2020 - November 2020*
Summer Research Intern

· Studied skill specialization and planning adaptation via Bayesian planning and inference in a multi-agent game environment. Advised by Desmond Ong.

Scale AI *February 2020 - November 2021*
Contractor

· Contracted to work on OpenAI's first RLHF project on summarization. Analyzing human evaluations and identifying GPT-3 weaknesses/edge cases in instruction following. Project supervised by Long Ouyang, Jeff Wu.

Kreiman Lab *June 2018 - August 2018*
Summer Research Intern

· Worked on object recognition using a recurrent attentional DL model for contextual reasoning. Advised by Gabriel Kreiman and Dr. Mengmi Zhang.

SCHOLARSHIPS & AWARDS

Open Philanthropy Grant for Model Interpretability and Behavior Steering Research (2024)

Future of Life Institute AI Existential Safety Community Membership (2023)

Open Philanthropy Early-Career Funding to Improve the Long-Term Future (2022)

Stanford Existential Risk Initiative ML Alignment Theory Scholars Program (2022)

Stanford Existential Risk Initiative Summer Research Fellowship (Summer 2022)

NSF GRFP Honorable Mention (2021)

Stanford Existential Risk Initiative Fellow (Summer 2021)

Ezra Stiles Richter Fellowship (Summer 2020)

Yale Domestic Summer Award (Summer 2019)

Rotary International Scholarship (2016)

QuestBridge Scholar (2016-2021)

TEACHING EXPERIENCE

Machine Learning Safety Scholars *June 2022 - August 2022*
Teaching Assistant

· Held office hours and graded problem sets and projects for ML Safety Scholars (https://course.mlsafety.org/) which covers topics in ML, DL, and AI Safety.

CPSC 202: Mathematical Tools for Computer Science *August 2019 - December 2019*
Undergraduate TA (ULA)

- Graded problem sets on introductory discrete math, probability, proof and logic, linear algebra content.

Math 1 *August 2015 - May 2016*
John W. North High School Teaching Assistant