CHAN JUN SHERN

chanjunshern@gmail.com | github.com/JunShern | junshern.github.io | Google Scholar

Work & Education

2023-now ***Tech Lead (Contractor)*** at OpenAI, Dangerous Capability Evaluations

I was hired by OpenAI as a contract research engineer to build out Dangerous Capability Evaluations ("evals") and was promoted to Tech Lead for our 10-person team after half a year. On a day-to-day, I lead the development of our team's evals framework and provide advisory input on all our work, but I have also personally built evals studying social manipulation, deception, steganography, and self-augmentation capabilities. I am proud of my work taxonomizing target capabilities, have designed many of the evals being worked on by our team, and I literally "wrote the playbook" for our team to coordinate and scale our efforts. As our work matures, I have an increased focus on strategic direction, including our team's path-to-impact and ecosystem-building.

2023-now ***Research Engineer*** at Center for AI Safety

Building on collaborations from my CHAI internship, I also work part-time at the Center for AI Safety. I currently advise on a policymaker-facing report with RAND Corp to establish best practices for frontier model threat assessment. I've also worked on a project studying the robustness of learned proxy objectives under optimization pressure (e.g. optimizing language models toward helpful and harmless objectives) [1].

2022-23 ***Research Intern*** at Center for Human-Compatible Artificial Intelligence (with Dan Hendrycks)

Over 6 months, I worked on a variety of AI safety projects (including contributions to research on learning human values [6] and organizing an ML research competition on moral uncertainty [5]). My main focus was building a benchmark (MACHIAVELLI) to monitor the behavior of language agents within fictional text games [2]. I was heavily involved in conceptualizing the benchmark and led the development of our game environment, including an annotation pipeline to produce multiple label types for 100,000 text scenarios using large language models.

2022 ***Visiting Researcher*** at New York University (with Ethan Perez via Fund for Alignment Research)

This 6-month placement and ongoing collaborations led to one first-author paper [6] and three additional contributions [3,4,8]. My first-author paper explored the idea of extracting a large-scale dataset for language model few-shot learning (400k tasks) from internet tables. I led the majority of the work, including development, experimentation, and paper writing.

2018-21 ***Senior Research Engineer*** at Motional

As a founding member of the Calibration team, I prototyped our team's algorithms for camera, LiDAR, and radar sensor calibration (computer vision + multi-view geometry) and saw things through to deployment in operations. As our vehicle fleet grew, I led work on a tool enabling researchers to test new algorithms at scale: new algorithm code gets pushed to cloud infrastructure running a custom benchmark equivalent to testing on 100 vehicles across 5 locations worldwide.

2014-18 ***MEng Electrical & Electronic Engineering (1st Class)*** at Imperial College London

My thesis [10] explored autoencoder-learned music representations; sampling from this space enabled controllable music generation which I built into a live accompaniment system. Won the *Student Centenary Prize* (awarded to 1/150 students).

2018 ***Research Intern*** at nuTonomy

I joined the Perception team at this early-stage startup (<50 people), working on radar calibration among other things.

Other Experience

2020-23 **Writing & Side Projects**

ML Experiment Workflow	A guide and example code for structuring and managing ML experiments.
Present Tense, Future Tensor	My recent blog on AI and AI safety topics.
Paper Reading Group	I distill AI research publications into 10-slide summaries. (>1k followers)
Spinning Up in Deep RL	Blog post detailing my experience implementing key RL algorithms.
What is AI For Good?	A layperson's survey of AI applications for social good.

2021 **Prize Winner** at MineRL BASALT (NeurIPS 2021 Competition)

My team won 3rd place and the Most Creative Research prize in this Minecraft agent competition [8]. I suggested our approach mixing behavior cloning + RL from human feedback, built the feedback interface, and trained the reward model.

Community & Teaching

2023 **Subject-matter expert** at 80,000 Hours career-advising.
(Volunteer work) I take 1-2 calls monthly to advise promising candidates on entering and contributing to AI safety.

2023 **Facilitator** at BlueDot Impact's AGI Safety Fundamentals course.
(Volunteer work) I led a 5-person group in weekly sessions through a 9-week introductory course on AI safety.

2018 **Google Summer of Code participant** at Processing Foundation
Contributed to p5.js, culminating in the explorable web tutorial "Algorithmic Music Composition". [Write-up]

2018 **Founder & Instructor** at Creative Coding for Beginners
(Volunteer work) Developed a curriculum with p5.js and taught beginner coding workshops in Kuala Lumpur. [Blog post]

2018 **Instructor** at Fire Tech
Taught Python, Scratch, and Arduino at after-school clubs and holiday camps in London.

Research Highlights

[1] "Benchmarking Neural Network Proxy Robustness to Optimization Pressures" Upcoming 2024. (Available on request.)
*Andy Zou, Long Phan, Nathaniel Li, **Jun Shern Chan**, Mantas Mazeika, Aidan O'Gara, Steven Basart, Jonathan Ng, Scott Emmons, J Zico Kolter, Matt Fredrikson, Dan Hendrycks.*

[2] "Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark" ICML 2023, Oral Presentation.
Alexander Pan, **Chan Jun Shern***, Andy Zou*, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, Dan Hendrycks.*

[3] "Training Language Models with Language Feedback at Scale" Upcoming 2023. (Available on request.)
*Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, **Jun Shern Chan**, Angelica Chen, Kyunghyun Cho, Ethan Perez.*

[4] "Improving Code Generation by Training with Natural Language Feedback" Upcoming 2023. (Available on request.)
*Angelica Chen, Jérémy Scheurer, Tomasz Korbak, Jon Ander Campos, **Jun Shern Chan**, Samuel R. Bowman, Kyunghyun Cho, Ethan Perez.*

[5] "The Moral Uncertainty Research Competition" 2022.
***Jun Shern Chan**, Dan Hendrycks.*

[6] "How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios" NeurIPS 2022.
*Mantas Mazeika, Eric Tang, Andy Zou, Steven Basart, **Jun Shern Chan**, Dawn Song, David Forsyth, Jacob Steinhardt, Dan Hendrycks.*

[7] "Few-shot Adaptation Works with UnpredicTable Data" ACL Rolling Review 2022.
***Jun Shern Chan**, Michael Pieler, Jonathan Jao, Jérémy Scheurer, Ethan Perez.*

[8] "Training Language Models with Language Feedback" ACL 2022.
*Jérémy Scheurer, Jon Ander Campos, **Jun Shern Chan**, Angelica Chen, Kyunghyun Cho, Ethan Perez.*

[9] "Retrospective on the 2021 BASALT Competition on Learning from Human Feedback." NeurIPS 2021.
*Rohin Shah, Steven H. Wang, Cody Wild, Stephanie Milani, Anssi Kanervisto, Vinicius G. Goecks, Nicholas Waytowich, David Watkins-Valls, Bharat Prakash, Edmund Mills, Divyansh Garg, Alexander Fries, Alexandra Souly, **Chan Jun Shern**, Daniel del Castillo, Tom Lieberum.*

[10] "Comper: A Collaborative Musical Accompaniment System using Deep Latent Vector Models." 2018.
***Jun Shern Chan**, Yiannis Demiris.*