Jack Koch

2223 S Bendelow Trl, Tampa, FL 33629
+1-561-789-6860 • jack@jbkjr.com • jbkjr.com

EDUCATION

Yale University — **New Haven, CT**
B.S., Applied Mathematics — *August 2015 - May 2019*
GPA: 3.7

Coursework includes: Data Analysis, Deep Learning Applications, Data Mining & Machine Learning, Natural Language Processing (NLP), Artificial Intelligence (AI), Data Structures, Python Programming, Probability and Statistics, Discrete Mathematics, Optimization, Linear Algebra.

PUBLICATIONS

- **Jack Koch***, Lauro Langosco*, Jacob Pfau, James Le and Lee Sharkey. "Objective Robustness in Deep Reinforcement Learning". In *Uncertainty & Robustness in Deep Reinforcement Learning Workshop at ICML* and *RL4RealLife Workshop at ICML*, 2021.

RESEARCH-RELATED BLOG POSTS

- Integrating Three Models of (Human) Cognition. Alignment Forum, 2021. Tying together different frameworks for human cognition, so that we can better understand how the brain implements goal-directed behavior, with the motivation of constraining expectations about how advanced ANNs might do the same.
- Grokking the Intentional Stance. Alignment Forum, 2021. On understanding takeaways from Dennett's "intentional stance" for modeling agency in the context of AI safety and alignment.
- Empirical Observations of Objective Robustness Failures. Alignment Forum, 2021. Associated blog post discussion for my paper "Objective Robustness in Deep Reinforcement Learning".
- Discussion: Objective Robustness and Inner Alignment Terminology. Alignment Forum, 2021. Auxiliary post for the "Objective Robustness" project, discussing differences in framings of the "objective robustness" or "inner alignment" problem in the AI alignment community.
- Mapping the Conceptual Territory in AI Existential Safety and Alignment. Alignment Forum, 2021. A post distilling what I had learned about different areas within beneficial AI research and how they relate while becoming acquainted with the field.

RESEARCH AND PROFESSIONAL EXPERIENCE

Language, Information, and Learning at Yale (LILY) — **New Haven, CT**
Research Assistant — *November 2017 - December 2018*

- Bachelor's Thesis: Tree-to-Tree Neural Semantic Parsing. Compared effectiveness of sequence-to-sequence, sequence-to-tree, and new tree-to-tree architecture for semantic parsing.
- Project: Comparing Pre-Trained Language Models with Semantic Parsing. Applied pre-trained language models to semantic parsing, resulting in significant performance improvements. Developed first application of (then-)recent ELMo, GPT, and BERT language models to language generation.

*equal contribution

- Developed and tested new "Temporal Capsule Net" architecture on Visual QA tasks.
- Mentored other researchers in applying Transformer architecture to text-to-SQL tasks.
- Scraped and analyzed initial data for the lab's new "BioNLP" project in conjunction with the Yale School of Medicine.

Center on Long-term Risk **London, U.K.**
Summer Research Fellow *July 2021 - October 2021*

- Investigated how the human brain implements goal-directed behavior by attempting to unify perspectives from predictive processing and global neuronal workspace frameworks, with the motivation of eventually being able to use such a perspective to constrain expectations about how advanced artificial neural networks might implement similar behaviors.
- Inquired into how we can model agency (in the context of AI) using Dennett's "intentional stance."

AI Safety Camp **Remote**
Program Participant, Project Lead *March 2021 - June 2021*

- Proposed a research project that was selected as a camp project by other participants.
- Co-led the aforementioned project, which produced the first empirical demonstrations of objective robustness failures, a novel kind of robustness failure in which a reinforcement learning agent retains a general level of capability under distributional shift but uses it to pursue an objective other than the one for which it was trained.

Machine Intelligence Research Institute **Remote**
AI Safety Retraining Program *March 2021 - June 2021*

- Awarded grant from Open Philanthropy to self-study topics in technical AI safety and alignment for three months.
- Self-taught reinforcement learning with OpenAI's "Spinning Up in Deep RL" and Berkeley's CS285 courses. Implemented a few classic RL algorithms in Gym environments, gaining familiarity with the various kinds (and relative advantages in different problems or contexts) of different reinforcement learning algorithms.
- Examined recent literature on inverse reinforcement learning and learning from human feedback.
- Investigated cases for catastrophic risk from advanced AI systems, timelines, takeoff speeds, etc.

Custom Social **Remote**
Part-time Machine Learning Intern *October 2020 - June 2021*

- Assisted in management of the outsourcing of the implementation of an AI recommendation system for targeted ads. Facilitated understanding between company management (who did not have technical backgrounds) and the data scientists at the contracted company.

AWARDS

- **Open Philanthropy Grant** *March 2021*
- **George J. Schultz Summer Fellowship in the Physical Sciences** *Summer 2018*

SERVICE AND LEADERSHIP

Yale Undergraduate Think Tank **New Haven, CT**
Founder & President *August 2016 - May 2018*

- Created and managed club. Discussed/debated world trends with potential to radically change society.
- Secured $21k in funding.
- Organized speaker events with artists, researchers, and thought leaders. Recruited members.

Bringing Youth Technology Everywhere (BYTE) **Florida**
Founder & President *Fall 2012 - Spring 2015*

- Founded non-profit dedicated to providing underprivileged children with technology to assist them in their studies.
- Fundraised to provide a local youth center with over 10 laptops.
- Raised funds to secure more than 20 laptops for an under-resourced school in my area.