Algon

Algon33.ar@gmail.com

Education

MPhys. Physics with Theoretical Physics, First Class, University of Manchester
Thesis: "The Polaron Transform Applied to a Landau-Zener Transition"

Experience

AI Evaluation Specialist, Equistamp.com

- Developed and implemented evaluation frameworks for assessing AI capabilities across diverse models and architectures
- Created automation scripts to standardize and integrate evaluations into the company's database system
- Contributed to the development of novel alignment evaluations while enhancing existing evaluations
- Conducted comprehensive analysis of AI research papers and codebases to identify and adapt high-quality evaluation methods

Distillation Fellow and Project Lead, AIsafety.info

- Authored articles on AI safety concepts, including Natural Abstraction Hypothesis and AI alignment fundamentals
- Led project transformation to improve focus on feedback and organizational culture
- Implemented creation of new content section for people first encountering AI safety

Independent Consulting

- Authored "Ambitious Methods in Education" report for Atlas Fellowship, analyzing educational methodologies surpassing common practices in generality, effectiveness, cost, and scalability
- Conducted research on "Detecting Strategic Reasoning in AI Systems," examining methods to evaluate strategic capabilities in near-term AI systems
- Developed automated Metaculus forecasting system for a superforecaster

Community Involvement

AI Safety Reading Group

- Long-term participant in Soren Elverlin's AI Safety reading group

Technical Skills

Programming: Python, Mathematica, TypeScript
Machine Learning: Deep Learning, Classical Machine Learning, Evals
Theoretical: Statistical Mechanics, AI Alignment Theory
Other: Technical Writing, Forecasting