GARRETT BAKER

GarretteBaker@outlook.com, Github, LessWrong

PREVIOUS JOB EXPERIENCE

- ML Alignment Theory Scholars scholar, Daniel Murfet & Jesse Hoogland (January, 2024 – present)
- Independent alignment researcher (October, 2022 – January, 2024)
- SERI ML Alignment Theory Scholars scholar, John Wentworth (June, 2022 – September, 2022)

EDUCATION

University of Maryland, College Park

· Formal Linear Algebra
· Differential Equations
· Multivariable Calculus
· Spectral Graph Theory
· *Less relevant classes available upon request*

Textbooks Read

· Axler's *Linear Algebra Done Right*
· Jaynes's *Probability Theory*
· Boyd's *Convex Optimization*
· Yudkowsky's *Rationality: AI to Zombies*
· Friedman's *Price Theory: An Intermediate Text*
· (Currently reading) Dayan & Abbott's *Theoretical Neurosicence: Computational and Mathematical Modeling of Neural Systems*
· (Currently reading) Wantanabe's *Algebraic Geometry and Statistical Learning Theory*

Online Courses

· Murfet, et al. *Singular Learning Theory and Alignment Primer*
· Goodman, et al. *Neuroscience for Machine Learners*
· Coursera Ng's "Machine Learning"

PUBLICATIONS

- Coauthored the paper "Generalization Analogies (GENIES): A Testbed for Generalizing AI Oversight to Hard-To-Measure Domains"
- Wrote "My hopes for alignment: Singular learning theory and whole brain emulation" on LessWrong
- Wrote "Singular learning theory and bridging from ML to brain emulations" on LessWrong
- Coauthored "Don't design agents which exploit adversarial inputs" on LessWrong

REFERENCES

- Alex Turner, alexmturner@google.com
- John Wentworth, jwentworth@g.hmc.edu
- Daniel Murfet, d.murfet@unimelb.edu.au