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

Textbooks Read

· Trefethen *et al.*'s *Numerical Linear Algebra*
· Jaynes's *Probability Theory*
· Braun's *Differential Equations and their Applications*
· Boyd's *Convex Optimization*
· Cowen & Tabarrok's *Modern Principles of Economics*
· Thurner, Hanel & Klimek's *Introduction to the Theory of Complex Systems*
· Russell & Norvig's *Artificial Intelligence: A Modern Approach*
· Sutton & Barto's *Reinforcement Learning*
· Suzuki's *WAIC and WBIC with R Stan*, Wantanabe's *Algebraic Geometry and Statistical Learning Theory* (current)
· Gerstner *et al.*'s *Neuronal Dynamics* (current)
· Grinfeld's *Introduction to Tensor Analysis and the Calculus of Moving Surfaces* (current)

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

SKILLS

- Programming - Python
- Supervised & reinforcement learning in Pytorch & Jax.

REFERENCES

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