JACOB PFAU

jacob.pfau@gmail.com scholar.google.com/citations?user=rl1IMgMAAAAJ jacobpfau

EDUCATION

New York University, Bowman Lab Fall 2022 - Present

PhD Student NYU Alignment Research Group, Center for Data Science

Current work includes: Red-teaming and latent adversarial training for LMs, Filler tokens and chain-of-thought in LMs

University of Edinburgh Sept. 2021 - Aug. 2022

MSc Mind, Language and Embodied Cognition (Philosophy Department). First class honors

Dissertation: 'When Decision Theories and AI Bluff in Games with Perfect Information'

Ecole Polytechnique (Université Paris-Saclay), Paris Sept. 2018 - Aug. 2019

M1, Masters Computer Science, Data Science track. GPA 3.91

Amherst College Sept. 2013 - May 2017

BA Mathematics & 5 College Logic Certificate. GPA (Math/CS/Logic) 3.73, Overall GPA 3.64

ETH Zurich (Math) Visiting student Jan-Aug 2016. GPA 5.34/6.0

EMPLOYMENT

Anthropic, *Contractor* Dec. 2023 - Present

Collaborative project writing report on 'Responsible Scaling for AI Moral Patienthood' assessing the implications of philosophy literature on LLMs

AI Safety Camp, *Mentor* Mar. 2023 - July 2023

Mentored four researchers transitioning into AI safety. Developed a benchmark for evaluating LM self-consistency on an arithmetic task

University of California, San Francisco. Keiser Lab, *Bold and Basic Fellow* Mar. 2019 - Aug. 2020

Developed interpretability methods and benchmarking datasets for clinical imaging. Mentored rotating students

PUBLICATIONS

Eliciting Language Model Behaviors using Reverse Language Models. http://tinyurl.com/rlmpdf,
Pfau J, Infanger A, Sheshadri A, Panda A, Huebner C, Michael J. Spotlight SoLaR at NeurIPS (2023)

Open problems and fundamental limitations of reinforcement learning from human feedback,
Casper S, ..., Pfau J, et al. TMLR (2023)

Self-Consistency of Large Language Models under Ambiguity, Bartsch H,... Pfau J. BlackboxNLP (2023)

Objective Robustness in Deep Reinforcement Learning, Koch J, Langosco L, Pfau J. Le J, Sharkey L. ICML (2022)

Robust Semantic Interpretability: Revisiting Concept Activation Vectors,
Pfau J, Young A, Wei J, Wei M, Keiser M. ICML Workshop on Human Interpretability (2020)

Stress Testing Reveals Gaps in Clinic Readiness of Image-Based Diagnostic AI Models,
Young A, Fernandez K, Pfau J, et al. npj Digital Medicine (2020)

Artificial Intelligence in Dermatology: A Primer, Young A, ... Pfau J, et al. Journal of Investigative Dermatology (2020)

Artificial Intelligence in Teledermatology, Xiong M, Pfau J, et al. Curr Derm Rep (2019)

Global Saliency: Aggregating Saliency Maps to Assess Dataset Artefact Bias,
Pfau J, Young A, Wei M, Keiser M. NeurIPS ML4Health (2019)

FELLOWSHIPS AND AWARDS

Bold and Basic Fellowship, *University of California, San Francisco* Sept. 2019 - Aug. 2020

Awarded yearlong research fellowship for applying CNNs to bridge skin cancer image data and vectorized genomic data

Research Internship Prize, *Ecole Polytechnique, Paris (for work conducted at UCSF)* Oct. 2019

Labex DigiCosme Fellowship, *Labex DigiCosme, Paris* Sept. 2018 - Mar. 2019

SKILLS AND MISCELLANEOUS

LANGUAGES: English (native), German (fluent), French (conversant), Mandarin (basic)