Jacek Karwowski

jacek.karwowski@cs.ox.ac.uk Academic website

EDUCATION

Ph.D. in Computer Science, University of Oxford *2022 - ongoing*
Specialization: Information geometry, programming languages, AI safety

M.Sc. in Mathematics, University of Warsaw *2018 - 2021*
Thesis: Formal semantics of a reversible language

B.Sc. in Computer Science, University of Warsaw *2018 - 2021*
Thesis: Low-latency Apache Parquet library (blogpost)

B.Sc. in Mathematics, University of Warsaw *2015 - 2018*
Thesis: Products in positive opetopic sets

PUBLICATIONS

V. Choudhury, J. Karwowski, and A. Sabry. Symmetries in reversible programming: From symmetric rig groupoids to reversible programming languages. *Proc. ACM Program. Lang.*, 6(POPL), Jan. 2022

N. Ackerman, C. E. Freer, Y. Kaddar, J. Karwowski, S. Moss, D. Roy, S. Staton, and H. Yang. Probabilistic programming interfaces for random graphs: Markov categories, graphons, and nominal sets. *Proc. ACM Program. Lang.*, 8(POPL), Jan. 2024

J. Karwowski, O. Hayman, X. Bai, K. Kiendlhofer, C. Griffin, and J. M. V. Skalse. Goodhart's law in reinforcement learning. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024

R. Douglas, J. Karwowski, C. Bae, A. Draguns, and V. Krakovna. Limitations of agents simulated by predictive models. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024

J. Janiak, J. Karwowski, C. S. Mangat, G. Giglemiani, N. Petrova, and S. Heimersheim. Characterizing stable regions in the residual stream of LLMs. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, Nov. 2024

J. Karwowski and F. Nielsen. Hilbert geometry of the symmetric positive-definite bicone. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, Oct. 2025

J. Karwowski and R. Douglas. Incoherence in goal-conditioned autoregressive models. In *The 29th International Conference on Artificial Intelligence and Statistics*, Feb. 2026

J. Karwowski and F. Nielsen. Geometric structures and deviations on James' symmetric positive-definite matrix bicone domain, Mar. 2026

J. Karwowski, Y. Kaddar, Z. Ye, N. Malkin, and S. Staton. Likelihood hacking in probabilistic program synthesis, Mar. 2026

TEACHING & ACADEMIC

Teaching

Teaching Assistant - Probability Theory, University of Oxford 2024
Teaching Assistant - Introduction to AI Alignment, Oxford AI Safety 2024
Teaching Assistant - Bayesian Statistical Probabilistic Programming, University of Oxford 2024
Teaching Assistant - Lambda Calculus and Types, University of Oxford 2023
Grader - Linear Algebra, University of Warsaw 2021

Academic Service

Reviewer - AISTATS 2026, NeurIPS 2025, ICLR 2025, NeurIPS 2024 (top reviewer), CAV 2023
Program Officer - PL in ML Conference 2017

Scholarships & Fellowships

Research Scholarship - SERI ML Alignment & Theory Scholars 2023
PhD Scholarship - Improving the Long-Term Future, Open Philanthropy 2022 - 2025
Fellowship - Global Talent Attraction Program, Indiana University 2019

WORK EXPERIENCE

Researcher - University of Oxford 2025 - ongoing

I'm working on an Opportunity Seed project "SynthStats: GFlowNet-steered probabilistic program synthesis for safer AI", funded by UK's Advanced Research and Invention Agency.

Researcher - National Institute of Informatics 2025

I worked on the information-geometric aspects of probabilistic programming languages theory at Sugiyama Lab, National Institute of Informatics in Tokyo, Japan.

Research lead - Supervised Program for Alignment Research 2025 - ongoing

I supervised the project "Understanding and verifying the autoregressive conditioning hypothesis". The team consisted of six people with backgrounds in computer science and mathematics.

Researcher - Stanford Existential Risks Initiative 2023

As part of SERI ML Alignment Theory Scholars program, I worked with Victoria Krakovna and others on understanding how can powerseeking arise in predictive models.

Researcher - Oxford AI Safety Labs 2022 - 2023

I worked with Joar Skalske at the Future of Humanity Institute and others on understanding Goodhart's law in reinforcement learning.

Quantitative Developer - Xantium 2021 - 2022

Xantium is the algorithmic trading division of Tudor Corporation. I worked on volatility trading, where I built models, implemented execution and managed datasets. Work details under NDA.

SWE Intern - G-Research 2020

G-Research is a quantitative research company, developing algorithms for predicting movements in financial markets. I worked on programmers tools for the simulations platform. Work details under NDA.

Research Intern - Indiana University 2019

As part of Global Talent Attraction Program, I conducted research on understanding the semantics of reversible programming languages, in the context of Homotopy Type Theory.

SWE Intern - Microsoft 2019

I worked on Office365 back-end, adding new features and optimizing the existing ones, and deploying code to production. Work details under NDA.

Research Intern - Golem Factory GmbH 2017 - 2018

Golem is a blockchain startup, creating a platform for sharing computational resources on Ethereum. I developed a novel probabilistic algorithm for verification of results in an untrusted environment (whitepaper, blogpost), and implemented it for distributed hyper-parameter optimization (blogpost).

Data Scientist - Applica.ai 2016 - 2017

Applica is a startup working on ML tools for NLP. I was responsible for creating new models: from designing experiments to prototyping models in Python and implementing them on the production cluster.

SWE Intern - University College London 2014 - 2015

Over the academic year, I have worked on developing a platform for testing high-frequency trading algorithms, and on an open-source medical data platform.