Jakub Kryś

Nationality: Polish DoB: 31/10/1996

jakkrys@gmail.com github.com/kryjak linkedin.com/in/jakub-krys

PhD graduate in theoretical particle physics working on AI safety. Strong research, mathematical and analytical skills. Adept at carrying out fast-paced projects in a team, as well as teaching, public speaking and science communication.

Experience

Sep 2025 — present

Research Scientist, SaferAI (London)

- Working as a research scientist at SaferAI, with a focus on risk modelling of AI-driven cyberattacks and Loss of Control scenarios
- Co-author of Toward Quantitative Modeling of Cybersecurity Risks Due to AI Misuse
- Created bespoke cyber risk models for external customers (private, sadly cannot share)
- Co-supervised 8 mentees across 3 SPAR projects (Spring 2026 cohort)

Jan 2025 — Jul 2025

ML Alignment and Theory Scholars (Berkeley and London)

- Worked with Janet Egan from CNAS on a technical AI governance project – studying the feasibility, incentives and implications of decentralised training runs
- Accepted for the Extension Phase in London, funded by Open Philanthropy for 6 months
- First-author paper accepted to TAIG ICML as an oral presentation

Jul 2024 — Aug 2024

Pivotal Research Fellowship (London)

- Worked on a technical AI safety project on adversarial robustness of neural networks. Researched transferability of image-based jailbreaks between Vision-Language Models
- Wrote PyTorch code to allow for the backpropagation of gradients to the image inputs of open-source models. This enabled training adversarial attacks on SOTA models.
- Mentored by Stanislav Fort from DeepMind and Arush Tagade from Leap Labs

Feb 2024 — Mar 2024

Faculty Fellowship (London)

- Completed a highly competitive 2-month programme organised by Faculty
- Received intense training in various topics in machine learning and data science
- Carried out a 6-week project exploring AI safety, in particular new jailbreak attacks on latest Vision-Language Models
- Exploited a combination of known techniques from prompt engineering, as well as novel ideas, to elicit undesired behaviour by targeting the visual modality

Oct 2022 — present

Other AI safety training and self-study

- Won (with a team) 4th prize in an Apart Research hackathon on technical governance (February 2026). Project on hashing data centre traffic with network taps.
- Participated in a 4-week AI Security Bootcamp (August 2025) and worked on creating hash collisions on AI-relevant file formats

- Participated in the training phase of the Talos Fellowship (Spring 2025). Produced a research piece on improving whistleblower protections in the EU, published by the AI Whistleblower Initiative
- Completed the Biosecurity Fundamentals course (Winter 2024) and created a Streamlit app for visualising correlations and forecasting of Covid-19 predictors (code and app)
- Participated in the S-Risk Introductory Fellowship (September 2024)
- Attended ML4Good Germany, an AI safety bootcamp (September 2024), with a mini project on AGI benefit-sharing
- Completed the AI Safety Fundamentals course (Summer 2024), with a project on sycophancy and bias in fine-tuned LLMs (code and report)
- Attended two summer schools in 2023: Oxford Machine Learning Summer School and Advanced Artificial Intelligence for Precision High Energy Physics

Oct 2019 — Nov 2023

PhD student in Theoretical Particle Physics (Durham University)

- Developed advanced tools for precision calculations in Quantum Chromodynamics using a combination of mathematical and computational methods
- Built a framework which enabled the calculation of cutting-edge scattering amplitudes for processes studied at the Large Hadron Collider
- Became proficient in symbolic programming (>3000hrs in Wolfram Mathematica)
- Used Linux, Git and LaTeX routinely
- Co-authored 3 peer-reviewed articles, gave seminars in the UK, Italy and South Korea

Education

Oct 2019 — Nov 2023

PhD in Theoretical Particle Physics (Durham University)

Full time at the University of Turin since Oct 2020

- Thesis title: 'Techniques for high-multiplicity scattering amplitudes and applications to precision collider physics'. Available online.
- Supervised by Prof. Simon Badger in the JetDynamics group

Oct 2015 — May 2019

MPhys in Theoretical Physics (Durham University)

- Graduated with First Class Honours
- Durham Physics Award for Outstanding Achievement in Year 2 and 3
- Programming projects in Wolfram Mathematica and Python

Computer skills

- Python: standard libraries such as numpy, pandas, seaborn, scipy, scikit-learn, pillow; OOP, Streamlit, PyTorch, einops, wandb, TransformerLens
- Other: Wolfram Mathematica, SQL, Docker, Unix, Git, LaTeX
- My work can be found at github.com/kryjak. Note that all code from my PhD and the Faculty Fellowship is not public.

Languages

Polish (native), English (fluent), Italian (working proficiency)