Kamile Lukosiute

lukosiutekamile@gmail.com

Education

- **University of Amsterdam** — Amsterdam, NL
  Master of Science in Physics and Astronomy — *Sep 2019 - Jul 2021*
  Thesis: Analysis of black hole and neutron star mergers using machine learning kilonova models
- **Wellesley College** — Wellesley, MA, USA
  Bachelor of Arts in Physics, cum laude — *Sep 2015 – May 2019*

Experience

- **Anthropic** — San Francisco, CA
  Resident AI Alignment Researcher — *Oct 2022 - Jul 2023*
- **University of Amsterdam** — Amsterdam, NL
  Instructor for Machine Learning Courses, PhD Student in Dark Matter Theory — *Jan 2021 - Apr 2022*
- **Fermi National Accelerator Laboratory** — Batavia, IL, USA
  Astrophysics Research Intern — *June 2019 - Jan 2020*

Awards

- AI Alignment Awards, 2023: For work on measuring corrigibility
- Future Fund, FTX, 2022: Financial support for career switch to focus on long-term future
- Amsterdam Science Talent Scholarship, University of Amsterdam, 2020: For top 10% of EU students for MSc study
- Wellesley College Graduate Fellowship, Wellesley College, 2020: Financial support for future graduate study
- Trustee Scholarship, Wellesley College, 2019: Financial support for future graduate study
- Levitt Fellowship, Wellesley College, 2018: Financial support the honor's thesis work

Publications

- D. Gaunguli et al. [including K. Lukosiute], "The Capacity for Moral Self-Correction in Large Language Models," arXiv:2302.07459., February 2023
- Y. Bai et al. [including K. Lukosiute], "Constitutional AI: Harmlessness from AI Feedback," preprint: arXiv:2212.08073., December 2022
- E. Perez, K. Lukosiute, S. Ringer, et al.. "Discovering Language Model Behaviors with Model-Written Evaluations," Findings of ACL, July 2023.
- S.R. Bowman et al. [including K. Lukosiute], "Measuring Progress on Scalable Oversight for Large Language Models," preprint: arXiv:2211.03540. , November 2022.
- B. Dorsman et al. [including K. Lukosiute], "Prospects of Gravitational Wave Follow-up Through a Wide-field Ultra-violet Satellite: a Dorado Case Study," The Astrophysical Journal, Volume 944, Number 2, Febraury 2023.
- K.Lukosiute, G. Raaijmakers, Z. Doctor, M. Soares-Santos, B. Nord, "`KilonovaNet`: Surrogate Models of Kilonova Spectra with Conditional Variational Autoencoders," Monthly Notices of the Royal Astronomical Society, Volume 516, Issue 1, October 2022
- K.Lukosiute, et. al., "Error Analysis of Kilonova Surrogate Models," NeurIPS Machine Learning for Physical Sciences Workshop, Dec 2021. Accompanying Code: github.com/klukosiute/kilonovanet
- G. Raaijmakers et al. [including K. Lukosiute], "The Challenges Ahead for Multimessenger Analyses of Gravitational Waves and Kilonova: A Case Study on GW190425," The Astrophysical Journal, 2021

Presentations

- K.Lukosiute, E. Perez "Discovering Language Model Behaviors with Model-Written Evaluations," DeepMind, invited talk, Feb 2023.
- K.Lukosiute, "Neutron star mergers and surrogate modelling," AI Association of Lithuania, Vilnius, Mar 2022.
- K.Lukosiute, et. al., "Surrogate modelling of kilonova spectra using conditional variational autoencoders," Fermi National Accelerator Laboratory New Perspectives Conference, Aug 2021.