Leonard Friedrich Bereska

Amsterdam, Netherlands
leonard.bereska@uva.nl | +31 683376135
PhD Candidate | AI Safety | UvA, Amsterdam

PROFILE

AI Safety enthusiast | Mechanistic Interpretability | JAX | Functional Programming

TECHNICAL SKILLS

- Python • JAX • PyTorch •
- Functional Programming
- Git • Bash
- Linux • LaTeX

LANGUAGE SKILLS

GERMAN native speaker
ENGLISH fluent
DUTCH conversational
MANDARIN conversational
FRENCH basic
ITALIAN basic
LATIN advanced Latinum
ANCIENT GREEK Graecum
OLD HEBREW Hebraicum

EDUCATION

UNIVERSITY OF AMSTERDAM PhD in Artificial Intelligence
Since October 2021. Expected Graduation: 2025 | Amsterdam, Netherlands
Pioneering transformer model interpretability through monosemanticity engineering for enhanced AI safety. Focused on AI Alignment strategies to ensure long-term value preservation.

UNIVERSITY OF HEIDELBERG MSc in Physics - Final grade 1.0
Graduated in February 2019 | Heidelberg, Germany
Visual Learning and Computer Vision (1.0), Machine Learning (1.0), Artificial Intelligence (1.0), Time Series Analysis (1.0). Thesis: 'Unsupervised Disentanglement of Geometric Shape and Visual Appearance' (1.0).

UNIVERSITY OF HEIDELBERG BSc in Physics - Final grade 1.7
Graduated in September 2016 | Heidelberg, Germany
Analysis 1 and 2 (2.3; 1.7), Linear Algebra 1 and 2 (1.3; 2.0), Theoretical Statistical Physics (1.3). Thesis: 'Optical Crosstalk in the Mu3e-Tile-Detector' (2.0).

NATIONAL TAIWAN UNIVERSITY Exchange Student
September 2014 - July 2015 | Taipei, Taiwan
Advanced-level Mandarin Chinese studies.

GYMNASIUM ERNESTINUM Abitur - Final grade 1.1
Graduated in July 2012 | Celle, Germany
Prized by German Mathematical, Physical, and Chemical Societies.

PUBLICATIONS

LORENZ, D., BERESKA, L., MILBICH, T., AND OMMER, B. (2019) Unsupervised part-based disentangling of object shape and appearance. CVPR, 2019 (oral, best paper finalist).
BRENNER, M., BERESKA, L., MIKHAEIL, J. M., HESS, F., MONFARED, Z., KUO, P.-C., & DURSTEWITZ, D. (2021) Tractable Dendritic RNNs for Identifying Unknown Nonlinear Dynamical Systems. ICML, 2021.
BERESKA, L., GAVVES, E. (2022). Continual Learning of Dynamical Systems with Competitive Federated Reservoir Computing. Conference on Lifelong Learning Agents, 2022. Published in PMLR.
BERESKA, L., GAVVES, E. (2023). Taming Simulators: Challenges, Pathways and Vision for the Alignment of Large Language Models. AAAI Inaugural Summer Symposium Series, 2023.

WORK EXPERIENCE

UNIVERSITY OF HEIDELBERG Research Assistant
February 2019 - today | Heidelberg, Germany
Infused dendritic computation principles into neural networks. Explored novel optimization criteria for dynamical systems.

CENTRAL INSTITUTE OF MENTAL HEALTH Research Intern
August 2017 - October 2017 | Mannheim, Germany
Investigated initialization schemes for a piecewise-linear recurrent neural network using expectation-maximization.