Leonard F. Bereska

Amsterdam, Netherlands
leonard.bereska@uva.nl | leonardbereska.github.io/ | +31 683376135
PhD Candidate | AI Safety | University of Amsterdam

PROFILE

AI Safety Researcher | Mechanistic Interpretability | Transformer Models

TECHNICAL SKILLS

• Python • JAX • PyTorch • Transformer Models • Adversarial Training • Circuit Analysis • Causal Interventions • Git • Bash • Linux • Vim • LaTeX

LANGUAGE SKILLS

GERMAN native speaker
ENGLISH fluent
DUTCH conversational
MANDARIN conversational

AWARDS

AI SAFETY HACKATHON 2nd Place
December 2023 | Delft, Netherlands

CERTIFICATES

ML SAFETY COURSE Dan Hendrycks, Center for AI Safety
August 2023

EDUCATION

UNIVERSITY OF AMSTERDAM PhD in Artificial Intelligence
Since October 2021. Expected Graduation: 2025 | Amsterdam, Netherlands
Thesis: "Mechanistic Interpretability for AI Safety"

UNIVERSITY OF HEIDELBERG MSc in Physics - Final grade 1.0
Graduated in February 2019 | Heidelberg, Germany
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

RESEARCH EXPERIENCE

UNIVERSITY OF AMSTERDAM PhD Candidate
October 2021 - Present | Amsterdam, Netherlands

- Reviewing the field of mechanistic interpretability
- Developing techniques for engineering monosemanticity in transformer models

UNIVERSITY OF HEIDELBERG Research Assistant
February 2019 - September 2021 | Heidelberg, Germany

- Integrated dendritic computation principles into neural networks
- Explored novel optimization criteria for dynamical systems

PUBLICATIONS

BERESKA, L., GAVVES, E. (2024). Mechanistic Interpretability for AI Safety - A Review. CoRR, Apr 2024.
BERESKA, L., GAVVES, E. (2023). Taming Simulators: Challenges, Pathways and Vision for the Alignment of Large Language Models. AAAI Inaugural Summer Symposium Series, 2023.
BERESKA, L., GAVVES, E. (2022). Continual Learning of Dynamical Systems with Competitive Federated Reservoir Computing. Conference on Lifelong Learning Agents, 2022. Published in PMLR.
BRENNER, M., BERESKA, L., ET AL. (2022) Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems. ICML, 2022.
LORENZ, D., BERESKA, L., ET AL. (2019) Unsupervised Part-Based Disentangling of Object Shape and Appearance. CVPR, 2019 (oral, best paper finalist).

LEADERSHIP & OUTREACH

AI SAFETY INITIATIVE AMSTERDAM Co-founder and Core Team Member
September 2023 - Present | Amsterdam, Netherlands

- Organized OpenAI Talk and Q&A on AI and Existential Risk
- Coordinated Panel Discussion on AI Risks: From Today to Doomsday
- Facilitated reading groups on AGI Safety Fundamentals