Leon Lang

+49 1525 6152317
l.lang@uva.nl
https://langleon.github.io/

Education

PhD Student 09/2020 – 02/2025
AI Safety and Alignment, Abstract Information Theory — University of Amsterdam, Netherlands

Master of Science 09/2018 – 08/2020
Artificial Intelligence — University of Amsterdam, Netherlands

- GPA: 9.34/10 (second best GPA among 135 graduates)

Master of Science 10/2015 – 10/2017
Mathematics — University of Bonn, Germany

- GPA: 1.0, with distinction. German grading scale with 1.0 as the best grade

Bachelor of Science 10/2012 – 08/2015
Mathematics — University of Heidelberg, Germany

- GPA: 1.1. (1.0 best — 5.0 worst)

Experience

Reviewer 01/2021 – 04/2025
Conferences, Journals, and Workshops

- Conferences: ICML 2025, ICLR 2025, NeurIPS 2024, UAI 2024, ICCS 2022
- Journals: IEEE Transactions in Information Theory, Physical Review E, IEEE Transactions on Neural Networks and Learning Systems
- Workshop: Safe and Trustworthy AI Workshop (STAI), 2023

Research Visitor 01/2024 – 03/2024
Krueger AI Safety Lab (KASL) — Cambridge, UK

Research Intern 03/2023 – 06/2023
Center for Human-Compatible AI (CHAI) — Berkeley, USA

Supervisor 02/2023 – 05/2023
Supervised Program for Alignment Research (SPAR) — Berkeley, USA

Research Scholar 11/2022 – 02/2023
ML Alignment & Theory Scholars (MATS) Program — Berkeley, USA

Research Intern 11/2019 – 07/2020
Qualcomm-UvA Deep Vision Lab (QUVA Lab) — Amsterdam, Netherlands

Data Analytics Intern 03/2018 – 07/2018
Detecon International GmbH — Cologne, Germany

Software Assistant 11/2017 – 01/2018
QAware GmbH — Munich, Germany

Teaching Assistant for 9 courses in total 10/2013 – 06/2024
Universities of Heidelberg, Bonn, and Amsterdam

- Mathematics courses: linear algebra I, real analysis I, topology I, causality
- AI courses: deep learning I, FACT-AI, machine learning I and II, foundation models

Publications

Leon Lang, Patrick Forré. *Modeling Human Beliefs about AI Behavior for Scalable Oversight.* Under review, 2025

Long Phan et al. *Humanity's Last Exam.* arXiv preprint, 2025

Leon Lang, Pierre Baudot, Rick Quax, Patrick Forré. *Information Decomposition Diagrams Applied beyond Shannon Entropy: A Generalization of Hu's Theorem.* Compositionality, 2025.

Leon Lang, Clélia de Mulatier, Rick Quax, Patrick Forré. *Abstract Markov Random Fields.* Under review, 2024.

Scott Garrabrant*, Matthias Mayer*, Magdalena Wache*, **Leon Lang** et al. *Factored space models: Towards causality between levels of abstraction.* Under review, 2024

Lukas Fluri*, **Leon Lang*** et al. *The Perils of Optimizing Learned Reward Functions: Low Training Error Does Not Guarantee Low Regret.* Under review, 2024

Leon Lang*, Davis Foote* et al. *When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback.* NeurIPS, 2024

Teun van der Weij*, Simon Lermen*, **Leon Lang**. *Evaluating Shutdown Avoidance of Language Models in Textual Scenarios.* Safe and Trustworthy AI Workshop (STAI), 2023.

Gabriele Cesa, **Leon Lang**, Maurice Weiler. *A Program to Build E(n)-Equivariant Steerable CNNs.* ICLR, 2022.

Leon Lang, Maurice Weiler. *A Wigner-Eckart Theorem for Group Equivariant Convolution Kernels*. ICLR, 2021.

- Top 2% of submitted papers in terms of average initial review score.

Benjamin Kolb*, **Leon Lang*** et al. *Learning to Request Guidance in Emergent Communication.* EMNLP Workshop LANTERN, 2019

Achievements and Awards

Award for a top 550 question 01/2025
Humanity's Last Exam Benchmark

Discussion of paper on AXRP podcast 06/2024
My co-author Scott Emmons discussed our work on RLHF and deception

Grant over $76,000 from Open Philanthropy 09/2023 – 02/2025
Funding to transition my PhD from abstract information theory to AI safety and alignment

Honorable Mention, AI Alignment Awards 07/2023
For my post on a shutdownability experiment idea

Distillation Prize by Nate Soares 04/2023
For my contribution to a comprehensive review of the natural abstractions agenda

Lesswrong Review Prize 01/2023
Given for two reviews of John Wentworth's abstractions work

Patent Application by Qualcomm 10/2022
For our research contribution on equivariant convolutional neural networks

Nomination of my MSc Thesis on Steerable Kernels 2021
Nomination for the Dutch Thesis Prize in Computer Science and Information Sciences

Deutschlandstipendium 04/2017 – 09/2017
Scholarship awarded to approximately 0.9% of students at any given time

Bundeswettbewerb Mathematik 06/2012
1st prize in the first round of this mathematics competition

Conferences and Summer Schools

Conference on Neural Information Processing Systems (NeurIPS) 12/2024
Vancouver, Canada

Human-aligned AI Summer School 07/2024
Prague, Czech Republic

Singular Learning Theory & Alignment Conference 06/2023
Berkeley, USA

Center for Human-Compatible AI (CHAI) Workshop 06/2023
Asilomar Conference Grounds, USA

Conference on Neural Information Processing Systems (NeurIPS) 11/2022
New Orleans, USA

Information Universe Conference 06/2022
Groningen, Netherlands

International Conference on Learning Representations (ICLR) 05/2021
Virtual

LANTERN Workshop at EMNLP 11/2019
Hongkong, remote

Human-aligned AI Summer School 07/2019
Prague, Czech Republic

AI Safety: Further Engagement

Various blogposts and research articles on AI Safety 07/2022 – 10/2024

- An explanation of my paper on RLHF under partial observability
- Musings about the changing representation of academia in AI Safety
- A comprehensive review of the natural abstractions agenda
- A comprehensive distillation of the core claims of shard theory
- An experiment idea to show that model-based RL agents will try to circumvent learned shutdownability
- An introduction article for the importance of AI Safety based on the distribution shift problem
- I distilled my learnings from the Alignment Fundamentals Program into a document containing summaries of more than 60 articles

Main Organizer 09/2018 – 07/2020
AI Safety Student Reading Group University of Amsterdam, Netherlands

- Monthly discussion meetups with approximately 10 students on papers about technical AI safety research

Skills

Languages: German (Native), English (Fluent), Dutch (Intermediate)
Programming: Python (NumPy, Matplotlib, PyTorch); internship experience in Java, C#, SQL
Document Creation: LaTeX