



LEON LANG

+49 1525 6152317

l.lang@uva.nl

<https://langleon.github.io/>

EDUCATION

PhD Student AI Safety and Alignment, Abstract Information Theory	09/2020 – 02/2025 University of Amsterdam, Netherlands
Master of Science Artificial Intelligence • GPA: 9.34/10 (second best GPA among 135 graduates)	09/2018 – 08/2020 University of Amsterdam, Netherlands
Master of Science Mathematics • GPA: 1.0, with distinction. German grading scale with 1.0 as the best grade	10/2015 – 10/2017 University of Bonn, Germany
Bachelor of Science Mathematics	10/2012 – 08/2015 University of Heidelberg, Germany
Prestudies Mathematics, Real Analysis I • A program that allows talented high school students to already study at university	10/2011 – 03/2012 University of Würzburg, Germany
High School Balthasar-Neumann-Gymnasium	09/2004 – 06/2012 Marktheidenfeld, Germany

EXPERIENCE

Reviewer 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	01/2021 – 04/2025
Research Visitor Krueger AI Safety Lab (KASL)	01/2024 – 03/2024 Cambridge, UK
Research Intern Center for Human-Compatible AI (CHAI)	03/2023 – 06/2023 Berkeley, USA
Supervisor Supervised Program for Alignment Research (SPAR)	02/2023 – 05/2023 Berkeley, USA
Research Scholar ML Alignment & Theory Scholars (MATS) Program	11/2022 – 02/2023 Berkeley, USA
Research Intern Qualcomm-UvA Deep Vision Lab (QUVA Lab)	11/2019 – 07/2020 Amsterdam, Netherlands
Data Analytics Intern Detecon International GmbH	03/2018 – 07/2018 Cologne, Germany
Software Assistant QAware GmbH	11/2017 – 01/2018 Munich, Germany
Teaching Assistant for 9 courses in total 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	10/2013 – 06/2024

PUBLICATIONS

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

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

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

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 Humanity's Last Exam Benchmark	01/2025
Discussion of paper on AXRP podcast My co-author Scott Emmons discussed our work on RLHF and deception	06/2024
Grant over \$76,000 from Open Philanthropy Funding to transition my PhD from abstract information theory to AI safety and alignment	09/2023 – 02/2025
Honorable Mention, AI Alignment Awards For my post on a shutdownability experiment idea	07/2023
Distillation Prize by Nate Soares For my contribution to a comprehensive review of the natural abstractions agenda	04/2023
Lesswrong Review Prize Given for two reviews of John Wentworth's abstractions work	01/2023
Patent Application by Qualcomm For our research contribution on equivariant convolutional neural networks	10/2022
Nomination of my MSc Thesis on Steerable Kernels Nomination for the Dutch Thesis Prize in Computer Science and Information Sciences	2021
Deutschlandstipendium Scholarship awarded to approximately 0.9% of students at any given time	04/2017 – 09/2017
Bundeswettbewerb Mathematik 1st prize in the first round of this prestigious mathematics competition	06/2012

CONFERENCES AND SUMMER SCHOOLS

Conference on Neural Information Processing Systems (NeurIPS) Vancouver, Canada	12/2024
Human-aligned AI Summer School Prague, Czech Republic	07/2024
Singular Learning Theory & Alignment Conference Berkeley, USA	06/2023
Center for Human-Compatible AI (CHAI) Workshop Asilomar Conference Grounds, USA	06/2023
Conference on Neural Information Processing Systems (NeurIPS) New Orleans, USA	11/2022
Information Universe Conference Groningen, Netherlands	06/2022
International Conference on Learning Representations (ICLR) Virtual	05/2021
LANTERN Workshop at EMNLP Hongkong, remote	11/2019
Human-aligned AI Summer School Prague, Czech Republic	07/2019

AI SAFETY: FURTHER ENGAGEMENT

Various blogposts on AI Safety	07/2022 – 10/2024
<ul style="list-style-type: none">• 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
<ul style="list-style-type: none">• 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