Georg Lange

Researcher working on mechanistic interpretability for LLMs

georglange.com | mail@georglange.com | Google Scholar | in/georg-lange | Goreg1234 | x.com/_georg_lange

EXPERIENCE

Independent Researcher

02/24 – Present

- Working on core infrastructure for mechanistic interpretability: SAE, Crosslayer Transcoder, generating feature explanations, circuit tracing (paper in prep.)
- Mentoring 7 scholars for the SPAR AI program on automating mechanistic interpretability with AI agents
- Created principled evaluations for Sparse Autoencoders (publication, arxiv, X)
- Research in dopamine and acetylcholine dynamics in mice (paper)

SERIMATS Scholar

Stanford Existential Risk Initiative

06/23 – 01/24 Berkeley, USA / London, UK

- Researcher working on mechanistic interpretability for LLMs with Alex Makelov, mentored by **Neel Nanda**
- Worked on Sparse Autoencoders and Distributed Alignment Search for feature detection and subspace activation patching (Paper)

Consultant for Data Science and Cloud Computing

Datametric, part-time

06/21 – 03/23 Amsterdam, Netherlands

- Developed Data Science solutions for major companies (Vattenfall, Ikea, T-mobile, Vesting-finance) with AWS, Sagemaker, and Azure ML

Student Consultant

SUGAR network, HPI, KIT

10/20 – 07/21 Potsdam, Germany

- Developed a new data-driven product from scratch for a major German insurance company; used Design Thinking and data to guide decisions; developed an Android App

ML Researcher

QiO Technologies Ltd., part-time

08/19 – 11/20 Potsdam, Germany

- Built an end-to-end Computer Vision pipeline for damage detection and annotation in sewers for a British wastewater company

Research Assistant

Digital Health Center, Hasso-Plattner-Institute, part-time

09/18 – 10/19 | 09/20 – 11/20 Potsdam, Germany

- Created a psychological test battery for epilepsy research
- Developed the frontend for a molecular tumor board

EDUCATION

M.Sc. Artificial Intelligence

University of Amsterdam

09/20 – 08/21 | 09/22 – 04/24 Amsterdam, Netherlands

- Courses on ML, DL, CV, IR, NLP, RL, interpretability, and proteomics
- Research in complex-valued neural networks for privacy protection and equivariant spatiotemporal CNNs for scene representation learning (paper)
- **Thesis** on brain-like interpretable spatiotemporal Computer Vision models with adaptation mechanisms, supervised by **Iris Groen** and Amber Brands

M.Sc. Cognitive Neuroscience

Graduate Center and Queens College, CUNY

Fulbright Scholar

08/21 – 01/23 New York, USA

- Investigating neural correlates of motivation and effort-based decision making in hyperdopaminergic (DAT-KD) and conditional D2-receptor KO (fDRD2 x Adora2a::Cre) mice using fiber photometry in Nucleus Accumbens
- **Thesis** on interaction between dopamine and acetylcholine during cocaine- or amphetamine-induced drug sensitization
- Developed a software package for fiber photometry analysis (fibermagic.org)
- Developed a platform to control operand boxes and neural data acquisition systems wirelessly (github)
- Hired and trained two undergraduate students on experimental neurobiology

B.Sc. IT-Systems-Engineering

Hasso-Plattner-Institute

10/17 – 09/20 Potsdam, Germany

- Courses on Math, Programming, TI, SWA, CP, OS, DBS, HCI, UML, CG
- Developed a Connected Health Care platform (Android App, Backend, Dashboard) for Unobtrusive Health Monitoring of Wearable Devices
- Thesis on ML for wireless EEG-based BCIs

TECHNICAL SKILLS

- Develop, debug and train NNs using **Pytorch**, Lightning, WandB, HF, **Tensorflow**
- Mechanistic Interpretability using **transformer-lens**, Pytorch, CircuitsVis, einops
- Python with **Pandas, NumPy, SciPy**, Sklearn, Matplotlib, Plotly, Seaborn, Statsmodels
- Cloud Computing with **AWS** using Sagemaker, EC2, Athena, EMR
- User-oriented product management using Design Thinking, rapid prototyping, interviews
- Basic knowledge of C (OS), C++ (CG), Kotlin, Java, Javascript, SQL, Arduino
- Building interactive devices using laser cutting, 3D printing, Arduino/Pi, electronics, CV

NEUROSCIENCE SKILLS

- Colony management (300 mice), **PCR**, genotyping, perfusion, cryostatic and vibratome slicing, **IHC**, IP injection, **fiber photometry**, electrophoresis, Confocal Microscopy
- **Stereotactic brain surgery** including viral injection, fiber fabrication, and implantation in mice
- **Neural and behavioral recordings in awake-behaving mice** (conditioning, operant, IP), setup development (3D-printing, computer vision, electronics, Raspberry Pi)

AWARDS AND ACHIEVEMENTS

Research Grants
Long-Term Future Fund
06/23 & 09/23
- Research grants from the Long-Term Future Fund for mechanistic interpretability research

Fulbright Scholar
German-American Fulbright Program
02/21 – 09/22
- Scholar of the German-American Fulbright Program

Cognitive Neuroscience Research Awards
GC, CUNY
04/22 & 11/22
- Research Awards of the Cognitive Neuroscience program, GC, CUNY

Konrad-Adenauer-Foundation Scholar
Political Foundation
01/18 – 01/23
- Scholar of Konrad-Adenauer-Foundation, Political Foundation

Online Course Creator
open.HPI
10/19 – 04/20
- Led and created a four-week **Online Course** about Deep Learning, Neural Nets, and image recognition for open.HPI
- 13,000+ participants

Additional Roles and Achievements
- Member of the **commission of studies, HPI**; developed M.Sc. Cybersecurity
- Head of Technology at **senkrechtstarter.org**, a buddy program for high school students
- Challenge winner of Hack Zurich, Europe's largest Hackathon with >1000 participants

PUBLICATIONS

- 04/25 — **Alex Makelov*** & **Georg Lange***, Neel Nanda. (2024). *Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control*. ***ICLR 2025*** (link)
- 04/25 — **Georg Lange**, Federico Gnazzo, Jeff Beeler (2025). *Accumbal Dopamine and Acetylcholine Dynamics during Psychostimulant Sensitization*. ***BioRxiv*** (link)
- 04/25 — Amber Brands, **Georg Lange**, Iris Groen (2025). *Temporal adaptation aids object recognition in deep convolutional neural networks in suboptimal viewing scenarios*. ***BioRxiv*** (link)
- 05/24 — **Alex Makelov*** & **Georg Lange***, Atticus Geiger, Neel Nanda. (2024). *Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching*. ***ICLR 2024*** (link)
- 05/21 — Arsen Sheverdin*, Alko Knijff*, Noud Corten* & **Georg Lange***. (2021). *[Re] of "Interpretable Complex-Valued NNs for Privacy Protection"*. ***Rescience*** (link)

Reviewer / Committee Service
- Served as a reviewer for **ICLR-2026, NeurIPS-2025, ICML-2024**, and the **Technical Program Committee of MATS 7 and 8**