

Maximilian Dreyer

Interpretability for safe and reliable AI

CEO, Layerwise · Postdoc, Fraunhofer HHI · Berlin
max.dreyer@outlook.com · +49 176 47364632 · maxdreyer.de · [Google Scholar](#)



Current safeguards for AI systems operate on outputs: judge models and content filters add latency and cost, yet miss failures rooted in the model’s internal computation. My research targets that computation directly. I develop mechanistic interpretability methods that attribute decisions to the concepts a model encodes, surface hidden failure modes such as shortcuts and biases, and correct them through targeted interventions in latent space. With Layerwise, I am bringing this into practice by monitoring and steering open-weight models at inference time. The goal is AI built like any other critical system: understood, monitored and correctable.

PhD	1,000+	2	11	3
summa cum laude	citations	Nature Machine Intelligence papers	talks & tutorials	patent families

POSITIONS

Postdoctoral Researcher , Explainable AI Group, Fraunhofer HHI	since 2026
Department of Artificial Intelligence (Prof. W. Samek), group of Dr. S. Lapuschkin	
Co-founder and CEO , Layerwise	2026
Berlin startup building an inference engine that reads and steers a model’s internal signals to detect and correct failures such as hallucinations and prompt injection	
Doctoral Researcher , Fraunhofer HHI	2022 – 2026
Concept-based explanations, model auditing and bias correction; medical imaging applications	
Student Research Assistant , Fraunhofer HHI	2020 – 2022
Student positions , IAV Berlin, Baumer Hübner, DESY Zeuthen	2016 – 2020
Autonomous-driving simulation, sensor metrology, teaching	

EDUCATION

Dr. rer. nat. , Technische Universität Berlin, <i>summa cum laude</i>	Mar 2026
From Local Explanations to Comprehensive Mechanistic Understanding of Deep Vision Models. Examiners: W. Samek, P. Biecek (Warsaw University of Technology), J. Choi (KAIST)	
M.Sc. Computational Science , University of Potsdam	2022
Thesis on concept-based explanations for segmentation and object detection, awarded by the Friends of Fraunhofer HHI	
B.Sc. Physics , Humboldt-Universität zu Berlin	2019
Exchange semester at Uppsala University, 2018	

SELECTED PAPERS

- Mechanistic understanding and validation of large AI models with SemanticLens.** [M. Dreyer*](#), J. Berend*, T. Labarta, J. Vielhaben, T. Wiegand, S. Lapuschkin, W. Samek. *Nature Machine Intelligence* 2025.
- From attribution maps to human-understandable explanations through Concept Relevance Propagation.** R. Achibat*, [M. Dreyer*](#), I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin. *Nature Machine Intelligence* 2023.
- Dyslexify: A mechanistic defense against typographic attacks in CLIP.** L. Hufe, C. Venhoff, [M. Dreyer](#), E. Pürelku, S. Lapuschkin, W. Samek. *ICLR* 2026.

From hope to safety: Unlearning biases of deep models via gradient penalization in latent space. [M. Dreyer*](#), F. Pahde*, C. J. Anders, W. Samek, S. Lapuschkin. *AAAI* 2024.

AttnLRP: Attention-aware layer-wise relevance propagation for transformers. R. Achibat, S. M. V. Hatefi, [M. Dreyer](#), A. Jain, T. Wiegand, S. Lapuschkin, W. Samek. *ICML* 2024.

Navigating neural space: Revisiting concept activation vectors to overcome directional divergence. F. Pahde, [M. Dreyer](#), M. Weckbecker, L. Weber, C. J. Anders, T. Wiegand, W. Samek, S. Lapuschkin. *ICLR* 2025.

Reveal to Revise: An explainable AI life cycle for iterative bias correction of deep models. F. Pahde*, [M. Dreyer*](#), W. Samek, S. Lapuschkin. *MICCAI* 2023.

Understanding the (extra-)ordinary: Validating deep model decisions with prototypical concept-based explanations. [M. Dreyer](#), R. Achibat, W. Samek, S. Lapuschkin. *CVPR Workshops* 2024.

PURE: Turning polysemantic neurons into pure features by identifying relevant circuits. [M. Dreyer](#), E. Purlku, J. Vielhaben, W. Samek, S. Lapuschkin. *CVPR Workshops* 2024, spotlight.

* equal contribution. 1,024 citations, h-index 12 (Google Scholar, October 2026).

INVITED TALKS

Right answer, wrong reason? From pixel-level grounding to concept-level auditing of VLMs · Amazon, Berlin	Sep 2026
Mechanistic understanding and validation of large AI models with SemanticLens · Interpretable Deep Learning Seminar Series	Jul 2026
From concepts to control: Diagnosing and steering vision foundation models · Tutorial, XAI4CV, CVPR 2026	Jun 2026
From explanation to unlearning: Concept-based diagnosis and control · Machine Unlearning for Vision Workshop, CVPR 2026	Jun 2026
Inspecting AI like engineers: From explanation to validation with SemanticLens · Keynote, Human-Centric AI Workshop, CIKM 2025	Nov 2025
Probing for the unknown: Mechanistic concept discovery beyond human expectation · Keynote, XAI4Science, ICLR 2025	Apr 2025
SemanticLens: A search engine to find bugs in large AI models · Fraunhofer AI Days	Jan 2025
Concept-based monitoring and debugging of model behavior · KAIST XAI Tutorial Series	Nov 2024
Pruning vision transformers using layer-wise relevance propagation · All Hands Meeting of the German AI Community	Oct 2024
From attribution maps to human-understandable explanations with CRP · Visual Intelligence Workshop on Concept-based XAI	May 2024
Debugging biases of deep medical models · Stanford MedAI	Feb 2024

MORE

PATENTS	Inventor on three Fraunhofer patent families, pending in Europe and the US or via PCT: SemanticLens (lead inventor), Concept Relevance Propagation, and AttnLRP.
SOFTWARE	Open-source implementations of all methods above, including <i>zennit-crp</i> , <i>LXT</i> (LRP for transformers) and <i>semanticlens</i> ; about 530 GitHub stars across repositories. Public web demo of SemanticLens.
MENTORING	Supervised nine Master's and doctoral students, including as senior author on their work (World Conference on XAI 2026; CVPR Workshops 2026).
AWARD	Young Talent Award of the Friends of Fraunhofer HHI, for the Master's thesis, 2022
LANGUAGES	German (native), English (fluent)