Dr. rer. nat. Sebastian Lapuschkin

* December 16, 1986 in Würzburg

Fraunhofer Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, HHI
Einsteinufer 37, 10587 Berlin
http://www.hhi.fraunhofer.de
sebastian.lapuschkin@hhi.fraunhofer.de
+49 (30) 31002-371 • +49 (177) 483-2754

github.com/sebastian-lapuschkin • linkedin.com/in/sebastian-lapuschkin
scholar.google.com/citations?user=wpLQuroAAAAJ

Short Bio

Sebastian Lapuschkin is the Head of the Explainable Artificial Intelligence research group at Fraunhofer Heinrich Hertz Institute (HHI) in Berlin.

He received his Ph.D. degree with distinction from the Technische Universität Berlin in 2018 for his pioneering contributions to the field of Explainable Artificial Intelligence (XAI) and interpretable machine learning. From 2007 to 2013 he studied computer science (B. Sc. and M. Sc.) at the Technische Universität Berlin, with a focus on software engineering and machine learning.

Sebastian is the recipient of multiple awards, including the Hugo-Geiger-Prize for outstanding doctoral achievement and the 2020 Pattern Recognition Best Paper Award.

His research has shaped the field of XAI from the very beginning, with contributions to the first wave of XAI such as the popular and widely-used Layer-wise Relevance Propagation method, as well as timely works influencing the second wave of XAI with additions to the sub-fields of Mechanistic Interpretability, Data Attribution and XAI-based model- and data improvement.

Sebastian is an avid advocate for Open Science, demonstrated by numerous Free Open Source Software toolboxes published with the intent to warrant and facilitate reproducibility in AI research.

Since 2024 he is co-organizing The World Conference on eXplainable Artificial Intelligence and serves as a Topic Editor on "Opportunities and Challenges in Explainable Artificial Intelligence" for the MDPI Open Access Journals.

Further research interests include efficient machine learning and data analysis, as well as data and algorithm visualization.

Professional Experience

Technological University Dublin — Dublin, Ireland

External Scholar *2025 -*
at the Centre of eXplainable Artificial Intelligence. The Centre is the first of its kind in the Republic of Ireland and it aims to increase further and synergise cross-centres, college and external research collaboration.

Multidisciplinary Digital Publishing Institute (MDPI)

Topic Editor *2024 - 2026*
for "Opportunities and Challenges in Explainable Artificial Intelligence".
Submission pre-screening, review management and decision handling.

XAI4Science

Organizer *2024 - 2025*
of the Workshop "XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge (2025)", co-located with ICLR 2025 at Singapore EXPO, Singapore.

World Conference on eXplainable Artificial Intelligence

Steering Committee Member *2024 -*
for the 3rd XAI World Conference (2025) in Istanbul, Turkey.
Conference and Special Track co-organization.

Programme Committee Chair *2023 - 2024*
for the 2nd XAI World Conference (2024) in Valetta, Malta.
Conference and Special Track co-organization.

Fraunhofer Heinrich-Hertz-Institute — Berlin, Germany

Contact Person *2025 -*
for the Erasmus Mundus Joint Master in Intelligent Field Robotic Systems (IFRoS) associate partnership of Fraunhofer HHI.

Ethics Committee Member *2023 -*
Founding member of the first ethics committee at Fraunhofer HHI.

Head of Explainable Artificial Intelligence *2021 -*
Research Group Leadership and direction of XAI research & applications.

Tenured Researcher *2019 - 2020*
PostDoc research position in the Machine Learning Group at Fraunhofer HHI.

Research Associate *2014 - 2018*
Founding member of the Machine Learning Group at Fraunhofer HHI.

Technische Universität Berlin — Berlin, Germany

Research Associate *2013 - 2014*
Supervision by Prof. Dr. Klaus-Robert Müller.

Student Research- & Teaching Assistant *2011 - 2013*
Research & Teaching assistant to Prof. Dr. Klaus-Robert Müller.

Student Teaching Assistant *2009 - 2011*
Teaching assistant to Prof. Dr. Marc Alexa, Prof. Dr. Odej Kao and Prof. Dr. Oliver Brock.

Education

Technische Universität Berlin — Berlin, Germany

PhD in Machine Learning (***summa cum laude***) *2018*
Date of oral defense: December 19th, 2018.
Dean's signature on Doctorate Certificate dated January 23rd, 2019.

Thesis: "Opening the machine learning black box with Layer-wise Relevance Propagation"
Supervision headed by Prof. Dr. Klaus-Robert Müller.

Master of Science in Computer Science *2013*
Focus on machine learning, computer vision and large scale data analysis.

Bachelor of Science in Computer Science *2010*
Focus on algorithms and software development

Deutschhaus-Gymnasium — Würzburg, Germany

Abitur (pre-university secondary education) *2007*

Teaching

Teaching of and teaching support for 19 university courses since 2009, including

"Responsible Artificial Intelligence 1", Technische Universität Berlin *2023 -*

"(An Introduction to) Explainable AI", Universitat de Girona / IFRoS *2021 -*

Talks & Lectures

Over 32 invited talks and individual lectures held since 2017, including

Decoding AI: How Explainability Unlocks Actionable Insights *2025*
dida conference 2025, Berlin, Germany

XAI as a Tool Beyond Model Understanding – From Heatmaps to Concepts and XAI Automation *2024*
CBS CoCoNUT | Max Planck Institute for Human Cognitive and Brain Sciences

Explaining AI with Concept Relevance Propagation *2023*
4th Japanese-American-German Frontiers of Science (JAGFOS) Symposium

Towards Human-understandable Explanations with XAI 2.0 *2022*
AI4Good webinar series of the International Telecommunication Union (ITU)

Towards Actionable XAI *2022*
International Artificial Intelligence Doctoral Academy

Third-Party Funded Research Projects

7 third-party funded research projects acquired and managed since 2018, including

ACHILLES *2024 - 2028*
Human-Centred Machine Learning: Lighter, Clearer, Safer
Funded with 8.2MM€ by the European Union

TEMA *2022 - 2026*
Trusted Extremely Precise Mapping and Prediction for Emergency Management
Funded with 11.3MM€ by the European Union

iToBoS *2021 - 2025*
Intelligent Total Body Scanner
Funded with 11.7MM€ by the European Union

Honors & Awards

Machine Learning and the Physical Sciences Reproducibility Badge *2024*

Stanford Top 2% Scientist Worldwide* *2021 - 2023*

Best Short Paper Award *2023*

Pattern Recognition Best Paper Award and Pattern Recognition Medal *2020*

Hugo-Geiger-Prize (1st place) *2019*

Freunde des HHI Nachwuchspreis *2019*

ERCIM Cor van Baayen Award (finalist) *2019*

Best Paper Award *2016*

Patents

Attention Head Functionalities in Machine Learning Models *2025*

A Concept Representation of a Machine Learning Model *2024*

Relevance Score Assignment dealing with an Attention Module and Applications thereof *2024*

Analyzing an Inference of a Machine Learning Predictor *2023*

Method and System for Simulating an Optical Image of a Photonic and/or Electronic Device *2022*

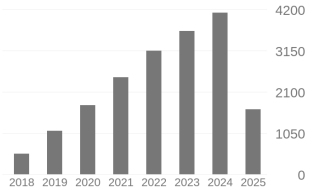
Pruning and/or Quantizing Machine Learning Predictors *2020*

Relevance Score Assignment for Artificial Neural Networks *2016*

Publications

Summary of Scientific Impact

	All	Since 2020
# Publications	89	89
# Citations	19064	16906
h-index	35	34
i10-index	56	55



per Google Scholar, retreived on June 10th, 2025.

Selected Publications

Theoretical & Methodological Contributions to XAI

1. Achtibat, Hatefi, Dreyer, Jain, Wiegand, **Lapuschkin**, Samek (2024).
"AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers".
In: *Proceedings of the 41st International Conference on Machine Learning (ICML)* 135–168.
In this paper we adapt the popular LRP method to contemporary Transformer architectures, yielding state-of-the-art explanation quality at exceptional computational efficiency. This acievement enables the analysis of the reasoning processes of Vision Transformers and LLMs in real time, eg. in chat bot applications.

2. Achtibat, Dreyer, Eisenbraun, Bosse, Wiegand, Samek and **Lapuschkin** (2023).
"From attribution maps to human-understandable explanations through Concept Relevance Propagation".
In: *Nature Machine Intelligence* 5(9):1006–1019.
A paper introducing the second generation of Explainable Artificial Intelligence with human-readable and abstract concept-based explanations.

3. Pahde, Dreyer, Samek and **Lapuschkin** (2023).
"Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models".
In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* 596–606.
This paper is dedicated to the incorporation of XAI as a standard component into the life cycle of Artificial Intelligence systems, with the intent to improve performance, reliability, and safety of AI.

4. Hedström, Weber, Krakowczyk, Bareeva, Motzkus, Samek, **Lapuschkin** and Höhne (2023).
"Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond".
In: *Journal of Machine Learning Research* 24(34):1–11.
In this paper we present the Quantus toolkit, the first-ever comprehensive XAI evaluation toolkit, constituting a well-organized collection of metrics and tutorials for evaluating explainable models, driven by community contributions.

5. Montavon, **Lapuschkin**, Binder, Samek and Müller (2017).
"Explaining NonLinear Classification Decisions with Deep Taylor Decomposition".
In: *Pattern Recognition* 65:211–222.
A paper discussing the mathematical foundation of LRP and its properties. Pattern Recognition Best Paper Award and Pattern Recognition Medal winner of 2020.

6. **Bach**, Binder, Montavon, Klauschen, Müller and Samek (2015).
"On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation".
In: *PLoS ONE* 10(7):e0130140.
A very influential and early work on local XAI, introducing the widely used Layer-wise Relevance Propagation method. This work has so far received over 5400 citations as counted by Google Scholar.

Applications of XAI

7. Anders, Weber, Neumann, Samek, Müller and **Lapuschkin** (2022).
"Finding and Removing Clever Hans: Using Explanation Methods to Debug and Improve Deep Models".
In: *Information Fusion* 77:261–295.
The authors' first work in a series dedicated to the exploitation of knowledge derived from XAI for the improvement of performance and robustness of AI systems.

8. Yeom, Seegerer, **Lapuschkin**, Binder, Wiedemann, Müller and Samek (2021).
"Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning".
In: *Pattern Recognition* 115:107899.
In this paper we leverage information about the importance of latent neural network structures obtained through XAI, in order to drastically reduce over-parameterization by identifying and removing non-critical components, resulting in up to 95% smaller models without loss of performance, and thus strong gains in energy and run time efficiency.

9. Horst, **Lapuschkin**, Samek, Müller and Schöllhorn (2019).
"Explaining the Unique Nature of Individual Gait Patterns with Deep Learning".
In: *Scientific Reports* 9:2391.
This paper is representative of many, in which we leverage techniques from XAI in domains where model transparency is critical, enabling for the first time the application of more powerful non-linear predictors beyond traditional linear systems in a feasible manner.

10. **Lapuschkin**, Wäldchen, Binder, Montavon, Samek and Müller (2019).
"Unmasking Clever Hans Predictors and Assessing what Machines Really Learn".
In: *Nature Communications* 10:1069.
One of the first papers to rigorously perform model- and data analysis through the lens of XAI, adding a voice of caution to the ongoing excitement about machine intelligence.

11. **Lapuschkin**, Binder, Montavon, Müller and Samek (2016).
"Analyzing Classifiers: Fisher Vectors and Deep Neural Networks".
In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016:2912-2920.
This is the first paper to use XAI to analyze and document differences in the behavior of state of the art predictors from several epochs of AI, in turn illuminating their use of yet unknown confounding features embedded in widely used computer vision benchmark datasets, critically scrutinizing previous key results from AI research.