

Kirill Bykov, Ph.D.

Machine Learning Post-Doctoral Researcher | Mechanistic Interpretability

Munich, Germany | kirill.bykov@tum.de | kirill-bykov.com | [Google Scholar](#) | [GitHub](#) | [LinkedIn](#)

PROFILE

Postdoctoral researcher in Prof. Akata's group at TUM, developing methods for Mechanistic Interpretability for LLMs. Ph.D. with summa cum laude from TU Berlin; research spans neuron- and representation-level analysis, multimodal / language-facing interpretability, model auditing, and reliable evaluation of explanations.

WORK EXPERIENCE

Postdoctoral Researcher

Aug 2025–Present

Technical University of Munich · Munich, Germany

Research in mechanistic interpretability for modern deep networks, with a focus on representations, feature reliability, and explanation methods for vision and language models.

Ph.D. Researcher in Machine Learning

Jan 2021–Jun 2025

TU Berlin and ATB Potsdam · Berlin and Potsdam, Germany

Research in mechanistic interpretability and explainable AI, developing methods to analyze neural representations, identify learned concepts, detect spurious correlations, and audit anomalous model behavior.

Student Research Assistant

May 2020–Jan 2021

Machine Learning Group, TU Berlin · Berlin, Germany

Research in explainable AI and Bayesian neural networks, contributing to methods for model explanations and uncertainty-aware analysis of deep networks.

Data Science Research Intern

Dec 2019–Mar 2020

TomTom, AI Geospatial Research · Berlin, Germany

Research and analytics in geospatial machine learning, developing anomaly detection models for large-scale spatial data.

Data Scientist

Jun 2018–Apr 2019

Skyeng · Remote

Research and analytics in recruitment and operations, building ML-driven candidate scoring, workload forecasting, process mining, and workflow optimization tools.

Data Analyst

Sep 2015–Dec 2017

MegaFon · Saint Petersburg, Russia

Analytics in trade marketing and business operations, focusing on sales forecasting, operational reporting, performance monitoring, and process optimization.

EDUCATION

Ph.D. in Machine Learning

2021–2025

Technische Universität Berlin · Berlin, Germany

Advisors: Prof. Marina Höhne and Prof. Klaus-Robert Müller; thesis: *Explaining Representations in Deep Neural Networks*; grade: summa cum laude, 1.0.

M.Sc. in Data Science / ICT Innovation

2018–2020

TU Berlin and TU Eindhoven · Berlin, Germany / Eindhoven, the Netherlands

Double-degree program between TU Berlin, Germany, and TU Eindhoven, the Netherlands; EIT Digital Excellence Scholarship; grade: cum laude, 1.3.

B.Sc. in Applied Mathematics and Computer Science

2014–2018

Saint Petersburg State University · Saint Petersburg, Russia

Faculty of Mathematics and Mechanics, Department of Statistical Modelling; grade: 4.47 / 5.

SELECTED PUBLICATIONS

NeurIPS 2025	Capturing Polysemanticity with PRISM: A Multi-Concept Feature Description Framework L. Kopf, N. Feldhus, K. Bykov , P. L. Bommer, A. Hedström, M. M.-C. Höhne, O. Eberle
NeurIPS 2025	Manipulating Feature Visualizations with Gradient Slingshots D. Bareeva, M. M.-C. Höhne, A. Warnecke, L. Pirch, K.-R. Müller, K. Rieck, S. Lapuschkin, K. Bykov
NeurIPS 2024	CoSy: Evaluating Textual Explanations of Neurons L. Kopf, P. L. Bommer, A. Hedström, S. Lapuschkin, M. M.-C. Höhne, K. Bykov
NeurIPS 2023	Labeling Neural Representations with Inverse Recognition K. Bykov , L. Kopf, S. Nakajima, M. Kloft, M. M.-C. Höhne
TMLR 2023	DORA: Exploring Outlier Representations in Deep Neural Networks K. Bykov , M. Deb, D. Grinwald, K.-R. Müller, M. M.-C. Höhne
AAAI 2022	NoiseGrad: Enhancing Explanations by Introducing Stochasticity to Model Weights K. Bykov , A. Hedström, S. Nakajima, M. M.-C. Höhne
TMLR 2025	Explaining Bayesian Neural Networks K. Bykov , M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, M. Kloft

LEADERSHIP, MENTORING, AND SERVICE

- **Research supervision:** Supervised 4 students; resulting work includes NeurIPS 2024, NeurIPS 2025, and an award-winning thesis recognized with the 2024 Rolf Niedermeier Prize. 3 students pursued PhDs afterwards.
- **Research organization:** Organized the “Global and Concept-Based Explainability” track at XAI 2024; moderated sessions at IEEE SaTML 2026, ECAI 2023.
- **Peer review and service:** Program Committee member for IEEE SaTML 2025, 2026; reviewer for NeurIPS, TMLR, IEEE TNNLS, and IEEE TRAMPI.
- **Grant writing:** Contributed to a successful BMBF-funded grant proposal (REFRAME, 01IS24073A-C) .

SELECTED INVITED TALKS

GLOBE: Global Explanation of Deep Neural Networks	Marburg University	Jun 2025
Explainable AI: From Local to Global	Potsdam Graduate School	Oct 2024
Labeling Neural Representations with Inverse Recognition	BLISS Berlin	Jan 2024
Panel discussion on Fair and Trustworthy AI	Helmholtz AI Conference	Jun 2022

TECHNICAL STRENGTHS

Technical	Python, PyTorch, vLLM, Hugging Face, SQL / NoSQL, Linux, Git, Bash, GCP / Azure, LLM coding agents, Codex
Languages	English C2, German B1, Russian native

SELECTED DISTINCTIONS

- Ph.D. defended with **summa cum laude**; thesis: *Explaining Representations in Deep Neural Networks*.
- EIT Digital Excellence Scholarship recipient, 2018–2020.
- Winner / prize-winner in hackathon, and mathematics competitions, including Data Natives 2019, HackDelft 2019, BioHack 2018, the SkolTech Statistical Learning Olympiad 2018, and the ITMO Open Mathematical Olympiad 2014.