MARTIN KRUTSKÝ

AI Researcher — Mechanistic Interpretability • Neurosymbolic AI • Responsible AI

martin.krutsky@cvut.cz • Prague, Czechia • martin-krutsky.github.io

linkedin.com/in/martin-krutsky • github.com/martin-krutsky • Google Scholar • rai.fel.cvut.cz

SUMMARY

PhD researcher at the Czech Technical University in Prague, working on mechanistic interpretability with neurosymbolic methods (sparse autoencoders, logic-guided circuit discovery, provable GNN explanations) and on responsible AI, assessing explainability for AI regulation. First-author papers in IEEE TNNLS and ECAI; 5+ years of industry ML engineering, from NLP models to shipped LLM applications.

RESEARCH EXPERIENCE

Applied Research Intern, Superhuman (formerly Grammarly) Sep 2026 – Dec 2026
- Natural language autoencoders to explain AI-text detectors and AI humanizer models.

Research Participant, EleutherAI Summer of Open AI Research (SOAR) Jul 2026 – Oct 2026
- Identified SAE pathologies that mimic feature hierarchy; proposed metrics to detect them.
- Evaluated SAEs from toy models to Gemma 2; manuscript in preparation.

AI Researcher & PhD Student, Intelligent Data Analysis Lab, FEE CTU Apr 2024 – Present
- Neurosymbolic mechanistic interpretability: logic-guided circuit discovery and sparse feature learning.
- Provably sufficient GNN explanations by pruning logical formulas matched to GNN expressivity.
- Binarized physics-inspired GNNs for combinatorial optimization on dense problems.
- Symmetry-invariant Rubik's Cube architecture, generalizing better than the state of the art.

Co-founder & Researcher, Responsible AI in Prague Sep 2023 – Present
- Assessing explainability methods for AI governance; derived XAI desiderata from the EU AI Act.
- Organized the **First Prague Workshop on Responsible AI**; talks and debates for students and the public.

Student Researcher, Center for Machine Perception, FEE CTU Aug 2018 – Aug 2019
- CNN- and SVM-based recognition of nationality and first impressions from faces.

ENGINEERING EXPERIENCE

ML Researcher & AI Engineer, Leeaf / McEDISON Aug 2023 – Apr 2026
- Shipped two LLM apps (RAG, tools, guardrails): an IVF patient assistant and a psychological-support app.
- Researched tabular ML models on IVF treatment data, in collaboration with medical experts.

Software Engineering Intern, Microsoft Jul 2022 – Sep 2022
- Modeled Microsoft Teams middleware load for Azure resource planning.

Machine Learning Developer, Blindspot Solutions Oct 2020 – May 2022
- Developed a smart inbox for manufacturing orders, customer-facing chatbots, and dynamic pricing models.
- NER and text classification with Transformer models; intent recognition.

Part-Time Developer, Barclays Sep 2019 – Sep 2020
- Server activity prediction (scikit-learn) and a monitoring dashboard.

PUBLICATIONS (PEER-REVIEWED)

Geometric Deep Learning for the Rubik's Cube Group IEEE TNNLS 2025
Krutský, Šír. *IEEE Transactions on Neural Networks and Learning Systems, 2025*

Binarizing Physics-Inspired GNNs for Combinatorial Optimization ECAI 2025
Krutský, Šír, Kungurtsev, Korpas. *European Conference on Artificial Intelligence, 2025*

XAI Desiderata for Trustworthy AI: Insights from the AI Act ECAI 2025 Workshop
Krutský et al. *European Workshop on Trustworthy AI at ECAI, 2025*

MANUSCRIPTS (UNDER REVIEW)

Towards Neurosymbolic Mechanistic Interpretability: The Case of GNNs ICLR 2027
Krutský, Šír. *International Conference on Learning Representations*

Sufficient Reasons for Explaining GNNs ICLR 2027
Krutský, Horčík, Šír. *International Conference on Learning Representations*

Mechanistic Interpretability Through Neurosymbolic Lens: Methods and Opportunities AIR (Springer)
Krutský, Šír. *Artificial Intelligence Review*

EDUCATION

Ph.D., Computer Science, Czech Technical University in Prague 2024 – Present
- Explainable neurosymbolic AI; advisor Gustav Šír.

M.Sc., Artificial Intelligence, Czech Technical University in Prague 2021 – 2024
- Graduated with honors; GPA 3.74/4.0.
- Thesis: Interpretable Symmetry-Aware Deep Learning for Planning.
- Erasmus+ exchange at Aalto University, Finland (2023).

B.Sc., Computer Science and Artificial Intelligence, Czech Technical University in Prague 2018 – 2021
- prg.ai Minor, a selective interdisciplinary AI program (2019 – 2021).

TEACHING

Teaching Assistant, Machine Learning Methods, FEE CTU Feb 2026 – Jun 2026
- New course; taught explainable AI, neurosymbolic AI, and LLMs for structured data.

SKILLS

Interpretability & XAI: sparse autoencoders, circuit discovery, GNN explainers, XAI for AI regulation
ML & LLMs: PyTorch, PyTorch Geometric, Hugging Face Transformers, scikit-learn, RAG, agents
Engineering: Python, SQL, Docker, FastAPI, Git, CI/CD, MLflow, AWS
Languages: Czech (native), English (proficient), German (intermediate)

SERVICE & LEADERSHIP

Reviewer, NeurIPS, ICANN, and the Neurosymbolic Artificial Intelligence journal 2026
Ethics Reviewer, NeurIPS; **Member**, CTU Ethics Committee 2026
President, Czech Alpha Chapter of Upsilon Pi Epsilon 2024 – Present

TALKS & PRESENTATIONS

Communicating Responsible AI to Different Audiences, *talk, CTU SciComm Meetup* 2026
Assessing XAI Methods for AI Safety Governance, *oral, Conf. on Large-Scale AI Risks* 2025
Dimensions of Explainability in AI Alignment, *poster, POSTER Conference* 2025
When AI Slanders Us & Evaluating Explanations, *public talks, AI Days* 2025
Ethics in AI Development, *public talk, AI Days* 2024
Responsible AI in Prague, *public talk, AI Days* 2023

AWARDS

UPE Academic Achievement Award, Upsilon Pi Epsilon 2022
Scholarship for Talented Students, Czech Technical University in Prague 2018

SUMMER SCHOOLS

European Summer School on Artificial Intelligence (Vienna, AT) 2026
Eastern European Machine Learning Summer School (Sarajevo, BA) 2025
Human-aligned AI Summer School (Prague, CZ) 2024
AI & Ethics Summer School (Bamberg, DE) 2022