Roman Kniazev

roman@knzv.me | roman.knzv.me | github.com/kameronton | Paris, France

Research interests

Mechanistic interpretability, representations of reasoning, AI alignment and safety

Research experience

Incoming Fellow | *Iliad* Dec 2026 – Mar 2027

- AI alignment research fellowship

Postdoctoral Researcher | *LaBRI, University of Bordeaux* Sep 2024 – Sep 2026

- **Mechanistic interpretability.** Trained a transformer that solves Sudoku puzzles, and showed that its linear world model is organized around rows, columns and boxes instead of cells, reflecting the task's constraint structure. Identified a sparse, monosemantic circuit in the final MLP layer responsible for decisions. Analyzed attention heads role and organization around puzzle structures. Published at *NeurIPS 2026* and the *ICML 2026 Mechanistic Interpretability Workshop*
- **Research software.** Built BordAX, a modular JAX deep-RL framework with fully JIT-compiled training pipelines; measured 2–3× speed-ups over Stable-Baselines3. Published in *Journal of Open Source Software*

Teaching and Research Assistant | *IRIF, Université Paris Cité* Sep 2023 – Aug 2024

- Researched multi-player game semantics of distributed protocols; taught undergraduate and graduate mathematics and computer science

Doctoral Researcher | *LIX, École Polytechnique* Sep 2020 – Aug 2023

- Developed topological and logical models for concurrent and distributed computing; taught courses including multithreaded CPU programming and CUDA

Education

École Polytechnique PhD in Computer Science 2023
Thesis: *On geometric models of epistemic logic*
Supervisors: Eric Goubault and Jean Goubault-Larrecq

ENS Paris-Saclay Master in Theoretical CS (MPRI), *magna cum laude* 2020

Novosibirsk State University Bachelor in Mathematics and CS, *summa cum laude* 2018

Selected publications

Transformers Linearly Represent Highly Structured World Models
R. Kniazev, N. Fijalkow
NeurIPS 2026; ICML 2026 Mechanistic Interpretability Workshop | Paper | Code | Blog

BordAX: A High-Performance JAX Framework for Programmatic Reinforcement Learning
R. Kniazev, N. Fijalkow
Journal of Open Source Software, 2026 | Paper | Code

Approximating the Reachability Value of Posterior-Deterministic POMDPs
N. Fijalkow, A. Ghosh, **R. Kniazev**, G. A. Pérez, P. Vandenhove
Preprint, 2026 | arXiv

Training

BlueDot Technical AI Safety course Aug 2026
Iliad Intensive AI alignment program, LISA, London Sep 2026

Technical skills

Tools: Python, JAX/Flax, PyTorch, Hugging Face Transformers
Interpretability: Linear probing, activation patching, circuit analysis
Training: Reinforcement learning; familiarity with SFT, PPO/GRPO and reward modelling
Evaluation: Model benchmarking, ablation studies
Formal methods: Epistemic logic, topological models, distributed computing