VLADIMIR IVANOV

EDUCATION

École Normale Supérieure Paris

- Graduate Computer Science *2023-2024*
- Undergraduate Computer Science & Mathematics *2022-2023*

Lycée Louis-le-Grand

- Undergraduate Mathematics & Computer Science & Physics *2021-2022*
- (Classe Préparatoire MP2I)

INTERNSHIPS

ERA Fellowship affiliated with Cambridge University, July-August 2024 (ongoing)

- Empirical study of deceptive and scheming tendencies in frontier LLMs.

SatisfIA Feb-Jun 2024

- Full-time technical AI safety internship.
- Practical and theoretical aspects of apsiration learning in deep reinforcement learning settings and known world model planning settings.

Realizability Interpretation of the Countable Axiom Of Choice Jun-Jul 2023

- Theoretical computer science research.
- Formalization in Coq.

TRAINING

Silver and Bronze International Mathematical Olympiad (IMO) Medals

- Also participated in other international mathematical olympiads (RMM, JBMO, MYMC).
- Participated in national computer science competition (Prologin).

ML4Good Camp by EffiSciences

- One week long hands on technical AI safety camp.
- PyTorch, LLMs, mechanistic interpretability, conceptual safety.

2022 Xena Project Undergraduate Workshop Imperial College, London

- One week long hands on Lean theorem prover workshop.
- My team worked on the Shannon-Lovász capacity theorem.

SELECT PROJECTS

Replication of Towards Monosemanticity with Dictionary Learning

- Replication of Anthropic's sparse autoencoder paper on the first layer of TinyStories-1M.
- Reimplementation of the transformer architecture.
- GitHub

Investigating the Influence of Dropout on Xor Probes in Neural Networks

- Small original research project, somewhat conclusive results.
- GitHub

Toy Self Replicating GPT4 Agent

- Toy scaffolding of GPT4 with a terminal which managed to self replicate to a remote machine.
- Note that my definition of self replication is much weaker than what METR failed to elicit.
- GitHub

PetitC Compiler

- Compiler of a subset of C.
- Semantics formalized and proven to be deterministic in Coq.
- Work in progress towards fully certifying the compiler.
- GitHub

TEACHING

Olympiad math classes at Animath and Lyon Discrete Math Club.

Assistance with AGI Safety Fundamentals at École Normale Supérieure Paris

COMMUNITY ENGAGEMENT

Member of the AI Unit at EffiSciences

- EffiSciences is a nonprofit effective altruism organization funded by Open Philantropy.