

# Aleksandar Makelov

---

## Experience

- July 2024 – Present **Visiting Scientist**, *Guide Labs*, LLM interpretability.
- Feb 2024 – July 2024 **Independent Researcher**, Research in mechanistic interpretability, Supported by the Long Term Future Fund.
- May 2023 – Jan 2024 **Researcher**, *SERI MATS*, Research in mechanistic interpretability, mentored by Neel Nanda, Supported by the Long Term Future Fund.

## Education

- Sep 2016– Sep 2022 **PhD**, *MIT EECS*, Mądry lab.  
Robust machine learning, spectral graph theory, mathematical optimization
- Oct 2015–June 2016 **Part III Mathematical Tripos**, *Emmanuel College*, University of Cambridge, with distinction.  
Coursework in combinatorics and algebra. Part III Essay: 'The graph isomorphism problem', supervised by Prof. Timothy Gowers
- Sep 2011–May 2015 **BA in Honors Mathematics and Computer Science**, *Harvard University*, summa cum laude.  
Undergraduate thesis: 'Expansion in lifts of graphs', supervised by Prof. Salil Vadhan

## Publications

- 2024 **Sparse Autoencoders Match Supervised Features for Model Steering on the IOI Task**, *A Makelov*, ICML 2024 Workshop on Mechanistic Interpretability (**Spotlight**).
- 2024 **Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control**, *A Makelov\**, *G Lange\**, *N Nanda*, Set LLM Workshop at ICLR 2024.
- 2024 **mandala: Compositional Memoization for Simple & Powerful Scientific Data Management**, *A Makelov*, SciPy 2024.
- 2023 **Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching**, *A Makelov\**, *G Lange\**, *N Nanda*, ICLR 2024.
- 2023 **Backdoor or Feature? A New Perspective On Data Poisoning**, *A Khaddaj\**, *G Leclerc\**, *A Makelov\**, *K Georgiev*, *A Ilyas*, *H Salman*, *A Mądry*, ICML 2023.
- 2018 **Towards Deep Learning Models Resistant to Adversarial Attacks**, *A Madry*, *A Makelov*, *L Schmidt*, *D Tsipras*, *A Vladu.*, ICLR 2018.
- 2015 **Expansion in Lifts of Graphs**, *A. Makelov*, Undergraduate thesis.

## Open source software projects

- 2023 **mandala**.  
A Python framework for data management of computational experiments.

---

## Teaching and Service

- June 2024 **NeurIPS 2024**, Reviewer.
- June 2024 **Workshop on Mechanistic Interpretability at ICML 2024**, Reviewer.  
**Top Reviewer Prize**
- April 2024 **Balkan Mathematical Olympiad**, *Bulgaria*, Coordinator.
- Fall 2019 **6.854: Advanced Algorithms**, *MIT*, Teaching Assistant.
- Spring 2019 **6.046: Design and Analysis of Algorithms**, *MIT*, Teaching Assistant.
- July 2017 **International Mathematical Olympiad**, *Brazil*, Observer A for Bulgaria, With support from 'American Foundation for Bulgaria'.
- July 2016 **International Mathematical Olympiad**, *Hong Kong*, Observer A for Bulgaria, With support from 'American Foundation for Bulgaria'.
- Fall 2014 **CS 125: Algorithms and Complexity**, *Harvard University*, Teaching Fellow.
- Fall 2013 **Math 131: Topology**, *Harvard University*, Teaching Fellow.
- 2010-2017 **International Mathematics Olympiad Preparation**, *With Bulgarian national team*, Delivered lectures on topics in olympiad mathematics.

---

## Awards and Honors

- 2015 **Akamai fellowship for first-year graduate students**, *MIT*.
- 2015 **Thomas Temple Hoopes Prize**, *Harvard University*.  
For undergraduate thesis 'Expansion in lifts of graphs'
- 2015 **Herchel Smith fellowship**, *Harvard University*.  
To support graduate studies at the University of Cambridge
- 2015 **Certificate of Teaching Excellence**, *Harvard University*.  
For 'Algorithms and complexity', Fall 2014
- 2014 **Phi Beta Kappa Junior 24**, *Harvard University*.
- 2012 **Honorable mention**, *William Lowell Putnam Mathematical Competition*.
- 2010 **AMC Medal**, *Australian Mathematics Competition*.
- 2010 **Silver medal**, *International Mathematical Olympiad*, Kazakhstan.  
Representing Bulgaria
- 2010 **Gold Medal**, *Balkan Mathematical Olympiad*, Moldova.  
Representing Bulgaria
- 2009, 2010 **Bronze & Silver medal**, *International Physics Olympiad*, Mexico & Croatia.  
Representing Bulgaria

---

## Open source software contributions

- 2017 **CIFAR10 Adversarial Examples Challenge**.  
A benchmark for training neural networks on the CIFAR10 dataset robust to adversarial examples

2012 **sympy**, *Google summer of code*.  
Contributed algorithms for computational group theory, advised by Prof. David Joyner,  
United States Naval Academy

---

## Coursework

**Advanced Algorithms**, *MIT*.

**Math 231a&b: Algebraic Topology**, *Harvard University*.

**Graduate courses in CS Theory**, *Harvard University*.

CS221 (Complexity), CS225 (Pseudorandomness), CS228 (Learning Theory), 2xCS229r  
(Topics in the Theory of Computation)

**Physics 16**, *Harvard University*.

**Math 55a&b**, *Harvard University*, with Prof. Yum-Tong Siu.

---

## Technical skills

**Programming Languages.**

Proficient in Python. Extensive experience with the PyData stack (numpy, pandas, scikit-learn, dask, matplotlib), Pytorch

**Databases.**

SQL (Postgres, sqlite) and ORMs (SQLAlchemy)

**OS.**

Linux/Unix

---

## Personal

In my free time I enjoy cycling, playing guitar/singing, hiking, and reading sci-fi.