

Maksym Andriushchenko – Curriculum Vitae

PERSONAL DATA

Site: <https://andriushchenko.me/>
Email: maksym@andriushchenko.me

Scholar: <https://scholar.google.com/citations?user=ZNtuJYoAAAAJ>
Github: <https://github.com/max-andr/>

EDUCATION

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland (*Sep 2019 - August 2024*)
PhD Degree in Computer Science advised by Nicolas Flammarion — *awarded with the Google PhD Fellowship, Open Phil AI PhD fellowship, G-Research PhD Thesis Prize*

Saarland University, Germany (*Oct 2016 – Aug 2019*)
Master's Degree in Computer Science advised by Matthias Hein from the University of Tübingen

Dnipro National University of Railway Transport, Ukraine (*Sep 2012 – June 2016*)
Bachelor's Degree in Software Engineering — *with honors*

EMPLOYMENT

Adobe Research, Media Intelligence Lab, remote **Time:** July 2021 – October 2021
Role: Research internship supervised by John Collomosse

PrivatBank, Dnipro, Ukraine **Time:** November 2015 – June 2016
Role: Part-time data scientist (predictive modeling, e-commerce personalization)

AWARDS

Grant and scholarship awards G-Research PhD Thesis Prize in Quantitative Research (\$3.3k award)
OpenAI Researcher Access Program (\$5k in API credits)
Google PhD fellowship 2022-2025 (\$80k per year for 3 years)
Open Philanthropy AI PhD Fellowship 2022-2024 (\$10k per year for travel/equipment)
Google Research Collab 2022-2023 (\$80k for one year + \$20k in cloud compute)
EDIC PhD fellowship from EPFL for the first year (\$60k)
DAAD MSc scholarship for 2 years to study at Saarland University (\$20k)

Awards for papers and competitions [Trojan detection challenge](#) at SatML'24: **first place**
[Swiss AI Safety Prize](#) (2024): **award for one of the top paper submissions**
Joint Conference of Korean AI Association (2023): **best paper award**
ICLR Workshop on Security & Safety in ML Systems (2021): **best paper honorable mention**
Swiss Machine Learning Day (2019): **best paper award**

ACADEMIC SERVICE

Participant Red teaming of the **OpenAI fine-tuning** service as an external expert (October 2023)
[Robust AI 4-day workshop](#) organized by Airbus AI Research and TNO (January 2021)

Reviewer NeurIPS'23, ICML'23, NeurIPS'22 (**top reviewer**), ICML'22, NeurIPS'21, ICML'21, CVPR'21, ICLR'21 (**outstanding reviewer**), NeurIPS'20 (**top 10% reviewers**)

Program committee in workshops **NeurIPS'23** “R0-FoMo Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models”, **NeurIPS'23** “Workshop on Distribution Shifts: New Frontiers with Foundation Models”, **ICML'23** “2nd ICML Workshop on New Frontiers in Adversarial ML”, **ICLR'23** “Workshop on Pitfalls of Limited Data and Computation for Trustworthy ML”, **NeurIPS'22** “Workshop on Distribution Shifts”, **NeurIPS'22** “ML Safety Workshop”, **ICML'22** “New Frontiers in Adversarial Machine Learning”, **ICML'22** “Principles of Distribution Shift”, **NeurIPS'21**: “Distribution Shifts: Connecting Methods and Applications”, **ICML'21** “Uncertainty and Robustness in Deep Learning”, **CVPR'21** “Adversarial ML in Real-World Computer Vision Systems”, **ICLR'21** “Robust and Reliable ML in the Real World”, “Security and Safety in ML Systems”, **ICML'20** “Uncertainty and Robustness in Deep Learning”, **CVPR'20** “Adversarial ML in Computer Vision”, **ICLR'20** “Towards Trustworthy ML” (**best reviewer award**)

Outreach activities National coordinator for Switzerland at [#ScienceForUkraine](#)
Coordinator for Switzerland and admission officer at the [Ukrainian Global University](#)
AI and STEM workshop at a [summer camp](#) for displaced Ukrainian children in Romania

STUDENT SUPERVISION

Joshua Freeman MSc project (2024): work in progress

Hao Zhao MSc thesis (2023): “Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning”

Hichem Hadhri MSc project (2023): “Understanding overfitting in large language models”

Tiberiu Musat BSc project (2023): “Investigating key components for fast optimization of deep networks”

Francesco d'Angelo PhD semester project (2023): “Understanding the role of weight decay in deep learning”

Théau Vannier MSc project (2023): “Understanding the training instability of transformers”

Joshua Freeman BSc project (2022, unofficial): “Automatic recognition of unexploded ordnance using transfer learning”

Jana Vuckovic MSc project (2022): “Rethinking the relationship between sharpness and generalization” (follow-up work is published at ICML'23)

Mehrdad Saberi Summer internship (2021): “Wasserstein adversarial training and perceptual robustness”

Edoardo Debenedetti MSc project (2021): “RobustBench: a standardized adversarial robustness benchmark” (published at NeurIPS'21 Datasets and Benchmarks Track)

Klim Kireev PhD semester project (2020): “On the effectiveness of adversarial training against common corruptions” (published at UAI'22)

Etienne Bonvin MSc project (2020): “Adversarial robustness of kernel methods”

Oriol Barbany MSc project (2019): “Affine-invariant robust training”

TEACHING EXPERIENCE

EPFL Probability & Statistics 2021, 2022 (by E. Abbé), Machine Learning 2020, 2021, 2022, 2023 (by M. Jaggi, N. Flammarion), Advanced Algorithms 2020 (by M. Kapralov)

MPI for Informatics Machine Learning 2018-2019 (lecturer: B. Schiele)

Saarland University Neural Networks: Implementation and Application 2017 (lecturer: D. Klakow)

SELECTED PUBLICATIONS

M. Andriushchenko, F. Croce, N. Flammarion. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks (arXiv, April 2024; [an earlier version](#) was recognized as **one of the top submissions for Swiss AI Safety Prize**) [[paper](#)]

H. Zhao, **M. Andriushchenko**, F. Croce, N. Flammarion. Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning (ICML 2024) [[paper](#)]

M. Andriushchenko, F. Croce, M. Müller, M. Hein, N. Flammarion. A Modern Look at the Relationship between Sharpness and Generalization (ICML 2023) [[paper](#)]

M. Andriushchenko, A. Varre, L. Pillaud-Vivien, N. Flammarion. SGD with Large Step Sizes Learns Sparse Features (ICML 2023) [[paper](#)]

F. Croce*, **M. Andriushchenko***, V. Sehwag*, E. Debenedetti*, N. Flammarion, M. Chiang, P. Mittal, M. Hein. RobustBench: a standardized adversarial robustness benchmark (NeurIPS 2021 Datasets and Benchmarks Track, **Best Paper Honorable Mention Prize** at [ICLR 2021 Workshop on Security and Safety in Machine Learning Systems](#)) [[paper](#)]

M. Andriushchenko, N. Flammarion. Understanding and Improving Fast Adversarial Training (NeurIPS 2020) [[paper](#)]

M. Andriushchenko*, F. Croce*, N. Flammarion, M. Hein. Square Attack: a Query-Efficient Black-Box Adversarial Attack via Random Search (ECCV 2020) [[paper](#)]

FULL PUBLICATION LIST

- J. Rando, F. Croce, K. Mitka, S. Shabalin, **M. Andriushchenko**, N. Flammarion, F. Tramèr. Competition Report: Finding Universal Jailbreak Backdoors in Aligned LLMs (arXiv, April 2024) [[paper](#)]
- M. Andriushchenko**, F. Croce, N. Flammarion. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks (arXiv, April 2024; [an earlier version](#) was recognized as **one of the top submissions for Swiss AI Safety Prize**) [[paper](#)]
- P. Chao*, E. Debenedetti*, A. Robey*, **M. Andriushchenko***, F. Croce, V. Schwag, E. Dobriban, N. Flammarion, G.J. Pappas, F. Tramer, H. Hassani, E. Wong. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models (arXiv, March 2024) [[paper](#)]
- H. Zhao, **M. Andriushchenko**, F. Croce, N. Flammarion. Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning (ICML 2024) [[paper](#)]
- L. Adilova, **M. Andriushchenko**, M. Kamp, A. Fischer, M. Jaggi. Layer-Wise Linear Mode Connectivity (ICLR 2024) [[paper](#)]
- M. Andriushchenko**. Adversarial Attacks on GPT-4 via Simple Random Search (December 2023) [[paper](#)]
- E. Debenedetti, Z. Wan, **M. Andriushchenko**, V. Schwag, K. Bhardwaj, B. Kailkhura. Scaling Compute Is Not All You Need for Adversarial Robustness (ICLR 2024 Workshop on Reliable and Responsible Foundation Models) [[paper](#)]
- M. Andriushchenko***, F. D'Angelo*, A. Varre, N. Flammarion. Why Do We Need Weight Decay in Modern Deep Learning? (NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning) [[paper](#)]
- S. Shin, D. Lee, **M. Andriushchenko**, N. Lee. The Effects of Overparameterization on Sharpness-Aware Minimization: An Empirical and Theoretical Analysis (September 2023, **best paper award** at the Joint Conference of Korean Artificial Intelligence Association (2023) [[paper](#)])
- M. Andriushchenko**, D. Bahri, H. Mobahi, N. Flammarion. Sharpness-Aware Minimization Leads to Low-Rank Features (NeurIPS 2023) [[paper](#)]
- K. Kireev, **M. Andriushchenko**, C. Troncoso, N. Flammarion. Transferable Adversarial Robustness for Categorical Data via Universal Robust Embeddings (NeurIPS 2023) [[paper](#)]
- M. Andriushchenko**, F. Croce, M. Müller, M. Hein, N. Flammarion. A modern look at the relationship between sharpness and generalization. (ICML 2023) [[paper](#)]
- M. Andriushchenko**, A. Varre, L. Pillaud-Vivien, N. Flammarion. SGD with large step sizes learns sparse features (ICML 2023) [[paper](#)]
- K. Kireev*, **M. Andriushchenko***, N. Flammarion. On the effectiveness of adversarial training against common corruptions (UAI 2022, [ICLR'21 Workshop on Robust and Reliable Machine Learning in the Real World](#)) [[paper](#)]
- Michael Rose, Sanita Reinson, **Maksym Andriushchenko**, Marcin Bartosiak, Anna Bobak et al. #ScienceForUkraine: an Initiative to Support the Ukrainian Academic Community. “3 Months Since Russia’s Invasion in Ukraine”, February 26 – May 31, 2022 (SSRN, 2022) [[paper](#)]
- M. Andriushchenko**, N. Flammarion. Towards Understanding Sharpness-Aware Minimization (ICML 2022) [[paper](#)]
- M. Andriushchenko**, X. Rebecca Li, Geoffrey Oxholm, Thomas Gittings, Tu Bui, Nicolas Flammarion, John Collomosse. ARIA: Adversarially Robust Image Attribution for Content Provenance ([CVPR 2022 Workshop on Media Forensics](#)) [[paper](#)]
- F. Croce, **M. Andriushchenko**, N. Singh, N. Flammarion, M. Hein. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks (AAAI 2022) [[paper](#)]
- F. Croce*, **M. Andriushchenko***, V. Schwag*, E. Debenedetti*, N. Flammarion, M. Chiang, P. Mittal, M. Hein. RobustBench: a standardized adversarial robustness benchmark (NeurIPS 2021 Datasets and Benchmarks Track, **Best Paper Honorable Mention Prize** at [ICLR 2021 Workshop on Security and Safety in Machine Learning Systems](#)) [[paper](#)]
- M. Mosbach, **M. Andriushchenko**, D. Klakow. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines (ICLR 2021) [[paper](#)]
- M. Andriushchenko***, F. Croce*, N. Flammarion, M. Hein. Square Attack: a query-efficient black-box adversarial attack via random search (ECCV 2020) [[paper](#)]
- M. Andriushchenko**, N. Flammarion. Understanding and Improving Fast Adversarial Training (NeurIPS 2020) [[paper](#)]
- M. Andriushchenko**, M. Hein. Provably Robust Boosted Decision Stumps and Trees against Adversarial Attacks (NeurIPS 2019, contributed talk at [Workshop on Machine Learning with Guarantees](#); **best paper award** at Swiss Machine Learning Day (2019)) [[paper](#)]
- M. Hein, **M. Andriushchenko**, J. Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem (oral at CVPR 2019, 5.6% acceptance rate, contributed talk at [ICML 2019 Uncertainty and Robustness in Deep Learning Workshop](#)) [[paper](#)]

F. Croce*, **M. Andriushchenko***, M. Hein. Provable Robustness of ReLU Networks via Maximization of Linear Regions (AISTATS 2019) [[paper](#)]

M. Mosbach*, **M. Andriushchenko***, T. Trost, M. Hein, D. Klakow. Logit Pairing Methods Can Fool Gradient-Based Attacks ([NeurIPS 2018 Workshop on Security in ML](#)) [[paper](#)]

M. Hein and **M. Andriushchenko**. Formal Guarantees on the Robustness of a Classifier Against Adversarial Manipulation (NeurIPS 2017) [[paper](#)]