Ensuring alignment of AI with societal interests is a key challenge of our time. I believe education can play a vital role in achieving this goal—**teaching technical subjects in AI must be combined with promoting active thinking of AI's broader impact**. Moreover, it is important to prepare students for real-world challenges by **adapting the educational process to account for the availability of generative AI**. Previously, I have **helped teach nine courses** in multiple institutions including *Machine Learning*, *Probability and Statistics*, and *Advanced Algorithms*. Additionally, I have **supervised thirteen students** at bachelor's, master's, and PhD level, whose work has led to academic distinctions and multiple top-tier publications [NeurIPS'21], [NeurIPS'24c], [ICML'24]. Looking ahead, I am eager to develop and **teach courses spanning responsible AI, AI safety, and LLM agents**, as well as more general courses related to LLMs, machine learning, introductory mathematics, and computer science.

## Teaching Philosophy

Rapid advances in generative AI are challenging both research and educational processes. I believe we need to adapt courses and design assignments to take this into account.

◇ **Teaching a data-centric view on AI.** Given the broad implications of AI development, I would actively encourage students to think critically about its long-term effects on society and explore ways to improve its outcomes by combining both algorithmic *and* data-driven approaches. The challenges of *AI alignment*—determining which values we should optimize for—becomes evident when examining various biases in training data: from outright discriminatory stereotypes to very subtle biases in human preferences that can lead to *sycophancy*, i.e., when the model learns to simply tell a user what they want to hear. Ensuring responsible development of AI systems thus requires examining large amounts of data at all stages of AI development, from collecting pretraining data and pairwise preferences to generating synthetic training data and conducting post-deployment monitoring. Historically, many AI courses have been focusing on the role of learning algorithms. However, with the rise of foundation models, it has become clear that the way we provide input-output specifications for these models is equally important. Thus, in my courses, I intend to promote a data-centric view of AI development that emphasizes both the role of data and learning algorithms.

◇ **Teaching to build *with* LLMs.** I think that deep technical expertise will remain crucial even with advanced AI models. While getting *a* solution becomes easier with each new model release, verifying these solutions will remain necessary due to AI's hallucinations and inaccuracies. Taking this into account, I believe we should train students to focus on correctness verification: if a student can accurately verify if a solution is correct, it should not matter whether it is AI-generated or not. This would also better reflect the real-world conditions, where generative AI is always available to help with any task. For example, more and more code is nowadays being written with AI assistance, which should be totally acceptable as long as students can understand it and verify its correctness. I believe it is important to adapt assignments by ensuring that they cannot be completely solved by LLMs, while still allowing LLMs to be used as helpful tools in the learning process. One promising approach would be to focus more on large project-based assignments where LLMs can be used as assistants but where they cannot replace students' scientific thinking and high-level design decisions.

◇ **General approach to teaching.** I believe it is crucial to make sure that students are both up-to-date with the latest advances in AI and have strong theoretical foundations that allow them to reason from first principles. My own experience of completing a PhD in the Theory of Machine Learning group at EPFL has shown me that strong theoretical foundations are essential for developing a holistic view of the field and being able to question existing assumptions—an approach that has been central to my work [ICML'22], [ICML'23b], [NeurIPS'24c]. On the practical side, I intend to promote *direct experimentation*, encouraging students to build proofs of concept and prototypes as quickly as possible and then continuously iterate over them. This is particularly easy now with widely available LLM APIs that are extremely easy to use and allow us to solve most traditional AI tasks with a single API call. To implement these principles effectively, I strive to create an engaging and inclusive learning environment that accommodates students from diverse backgrounds and disciplines, ensuring that study groups and class discussions remain accessible and productive for all participants.

## Teaching Experience

I have had the opportunity to serve as a teaching assistant at multiple institutions. At Saarland University, I assisted in teaching the *Neural Networks: Implementation and Application* course led by Prof. Dietrich Klakow in 2017. I also assisted with the *Machine Learning* course taught by Prof. Bernt Schiele at the MPI for Informatics in 2018. At EPFL, I served as a teaching assistant for three courses: *Advanced Algorithms* (2020) taught by Prof. Michael Kapralov, which had over 200 students; *Probability and Statistics* (2021, 2022) taught by Prof. Emmanuel Abbé; and *Machine Learning* (2020, 2021, 2022, 2023) taught by Prof. Martin Jaggi and Prof. Nicolas Flammarion. The *Machine Learning* course was among EPFL's largest, with more than 600 enrolled students. My responsibilities included developing lecture materials, creating assignments, and supervising ML for Science projects. In particular, to keep the course material up-to-date, I designed lectures on adversarial robustness and transformers from scratch. Outside academia, I taught AI lectures to children (ages 7-12) displaced by the war in Ukraine who were temporarily residing in Romania. Since I grew up in Ukraine, I felt that I could meaningfully contribute in this way by teaching basic AI concepts and promoting STEM. More generally, I believe it is especially important to begin outreach activities from a very young age to motivate children to study STEM-related subjects.

## Teaching Interests

I would be excited to teach courses that range from the latest advances in AI safety and LLM agents to basic mathematics for undergraduates. For example, I look forward to teaching the following courses:

- ◇ A course on responsible AI focusing on AI safety and security, adversarial robustness, privacy, science of evaluations, interpretability, and societal implications of AI. I would also be excited to give a seminar in this area, focusing on the most influential recent papers.
- ◇ A course on LLM agents covering LLM basics, teaching LLMs to use external tools, retrieval-augmented generation, reasoning and planning, software engineering agents, multi-agent systems, alignment techniques, and ethical considerations of using autonomous agents.
- ◇ A broad course on LLMs focusing on architectures, tokenization, pretraining, instruction tuning, in-context learning, advanced reasoning, and multimodal extensions of LLMs.
- ◇ Diverse introductory courses in computer science (e.g., machine learning, algorithms, computer vision, information theory) and mathematics (e.g., probability, statistics, analysis).

## Advising Approach

Mentoring the next generation of AI researchers is one of my key goals. I have worked closely with a number of undergraduate (Mehrdad Saberi, Tiberiu Musat), master's (Oriol Barbany, Etienne Bonvin, Edoardo Debenedetti, Jana Vuckovic, Théau Vannier, Hichem Hadhri, Hao Zhao, Joshua Freeman), and PhD students (Klim Kireev, Francesco d'Angelo, Alexander Panfilov) from EPFL, ETH Zürich, and the University of Tübingen. Their work has been accepted at top-tier conferences [NeurIPS'21], [NeurIPS'24c], [ICML'24] and workshops [NeurIPS'24a WS], [NeurIPS'24c WS].

Edoardo's work on `RobustBench` [NeurIPS'21] has received academic recognition through a Best Paper Honorable Mention Prize at an ICLR Workshop [NeurIPS'21]. He continued with a master's thesis at Princeton with one of our RobustBench collaborators and went on a PhD program at ETH Zürich. Hao's work on understanding instruction fine-tuning of LLMs [ICML'24] received a nomination for EPFL Outstanding Master's Thesis. He is now applying for PhD positions and I am glad that I have contributed to this decision.

My general advising approach is trying to understand what topics a student finds interesting, setting a clear vision about a problem that is of common interest, and making sure meaningful progress is regularly made. Rather than having students assist with small parts of larger projects, I prefer them to take leadership roles—typically as first authors on potential publications. While this approach is more challenging, it is also more rewarding. Students learn significantly more when they participate in every stage of research, from conducting initial experiments to writing the final paper.

## References

[ICML'23b] **Maksym Andriushchenko**, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. "A Modern Look at the Relationship Between Sharpness and Generalization". *ICML*. 2023.

[ICML'22] **Maksym Andriushchenko** and Nicolas Flammarion. "Towards Understanding Sharpness-Aware Minimization". *ICML*. 2022.

[NeurIPS'21] Francesco Croce*, **Maksym Andriushchenko***, Vikash Sehwag*, Edoardo Debenedetti*, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. "RobustBench: A Standardized Adversarial Robustness Benchmark". *NeurIPS Datasets and Benchmarks Track*. 2021.

[NeurIPS'24c] Francesco D'Angelo*, **Maksym Andriushchenko***, Aditya Varre, and Nicolas Flammarion. "Why Do We Need Weight Decay in Modern Deep Learning?" *NeurIPS*. 2024.

[NeurIPS'24c WS] Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and **Maksym Andriushchenko**. "Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 lawsuit". *Under submission (a short version appeared at the NeurIPS 2024 Safe Generative AI Workshop)* (2024).

[NeurIPS'24a WS] Hao Zhao, **Maksym Andriushchenko**, Francesco Croce, and Nicolas Flammarion. "Is In-Context Learning Sufficient for Instruction Following in LLMs?" *Under submission (a short version appeared at the NeurIPS 2024 Workshop on Adaptive Foundation Models)* (2024).

[ICML'24] Hao Zhao, **Maksym Andriushchenko**, Francesco Croce, and Nicolas Flammarion. "Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning". *ICML*. 2024.