Frédéric BERDOZ

PhD Student at ETH Zürich

fberdoz@ethz.ch
+41 79 949 07 35
Brandstrasse 55, 8952 Schlieren, Switzerland
Scholar
Website
Github

INTRO

Second-year PhD student under the supervision of Prof. Roger Wattenhofer, primarily focused on AI alignment and developing formal safety guarantees for autonomous agents. Also broadly interested in AGI and distributed systems, including distributed alignment, digital-era democracies, and distributed machine learning.

EDUCATION

		(GPA)		
ETH	*PhD*, Distributed Computing Group	-	*Zürich*	2023 - now
EPFL	*MSc*, Computational Science & Eng.	(5.73/6)	*Lausanne*	2019 - 2022
↳ **MIT**	*Master's thesis*, MIT Media Lab	(6.00/6)	*Boston*	spring 2022
EPFL	*BSc*, Mechanical Engineering	(5.66/6)	*Lausanne*	2016 - 2019
↳ **McGill**	*Full-year Academic Exchange*	(3.94/4)	*Montreal*	2018 - 2019

SKILLS

Programming: *Proficient*: Python | *Good*: C, C++, Bash | *Familiar*: Java, Fortran
Tools and Libraries: Docker, Git, Pytorch, Pandas, Numpy, TensorFlow, JAX
App & Web Development: *Good*: Swift | *Familiar*: Javascript, HTML, CSS
Languages: *Proficient*: French (Native), English (C1-C2) | *Familiar*: German (B2)

EXPERIENCE

Techinal Project Manager *Lausanne*, 11/2022 - 07/2023
Neural Concept
Led efforts to drive the adoption of 3D deep learning in real-world engineering (CAD and CAE) applications. Managed end-to-end project implementation, from securing and nurturing high-value leads to developing proof-of-concepts, deploying, and scaling solutions, resulting in successful deals exceeding $100k.

Research Assistant *Lausanne*, 08/2022 - 10/2022
Machine Learning and Optimization Lab, EPFL
Research on the topic of decentralized and privacy-preserving machine learning.

Data Science Intern *Zürich*, 08/2021 - 01/2022
Beyond Gravity (formerly *Ruag Space*)
Internship in data science within the aerospace industry, developing pipelines to analyze data early in the testing process and predict failures that could occur later in costly space mechanism tests, thereby reducing overall testing time and costs.

SELECTED PROJECTS

Can an AI safely run a government?: Developed a framework for the safe deployment of autonomous agents in social decision processes. Introduced a novel alignment definition based on utility and social choice theory, proposing "probably approximately aligned" policies with formal guarantees. Also presented a method to ensure agents' actions are verifiably safe. This work was accepted at NeurIPS 2024 [1].

Task-agnostic data valuation: Developed a task-agnostic method for valuing a data seller's data for a buyer, based on statistical differences in diversity and relevance without using validation metrics. The approach uses secure queries without exposing raw data, and experiments show it effectively captures the value of the seller's data. This work was accepted at AAAI 2023 [2].

Scalable collaborative machine learning: Developed a privacy-preserving machine learning framework that enables efficient collaboration by sharing averaged last-layer activations without exposing private data. This method reduces communication costs

and server dependency, while improving model performance in cross-device applications compared to traditional federated learning. This work received the runner-up Best Paper Award at the TSRML workshop, co-located with NeurIPS 2022 [3].

Analyzing and improving the robustness of voting advice applications: Investigated the risks VAAs pose to democracy under adversarial attacks, identifying 11 manipulation strategies and their impact. Showed how altering matching methods and question selection significantly influences party recommendations. Proposed robustness properties and mitigation strategies for securing future AI-based VAAs. This work is currently under review [4].

PAPERS

[1] **Frédéric Berdoz** and Roger Wattenhofer, *Can an AI Agent Safely Run a Government? Existence of Probably Approximately Aligned Policies*, in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, December 2024.

[2] Mohammad Mohammadi Amiri, **Frédéric Berdoz**, and Ramesh Raskar, *Fundamentals of Task-Agnostic Data Valuation*, in *Proceedings of the 37th Conference on Artificial Intelligence (AAAI)*, 2023.

[3] **Frédéric Berdoz**, Abhishek Singh, Martin Jaggi, and Ramesh Raskar, *Scalable Collaborative Learning via Representation Sharing*, *Workshop on Decentralization and Trustworthy Machine Learning in Web3 (TSRML@NeurIPS)*, 2022. *(Runner-up Best Paper Award).*

[4] **Frédéric Berdoz**, Dustin Brunner, Yann Vonlanthen, and Roger Wattenhofer, *Recommender Systems for Democracy: Toward Adversarial Robustness in Voting*, 2024. *(Under review).*

SELECTED COURSES

Advanced Algorithms by Michael Kapralov
Foundation of Data Science by Rüdiger Urbanke (Claude-Shanon Award 2023)
Machine Learning by Martin Jaggi and Rüdiger Urbanke
Deep Learning by François Fleuret
Image Processing I & II by Michael Unser

MISC.

Teaching: Student TA for multiple courses at EPFL during BSc and MSc, TA for multiple courses at ETH during PhD.

Volunteering: Organizing community events in my hometown, finding and managing up to 300 volunteers over events that last up to 2 weeks.

Associative: Former member of the Junior Entreprise EPFL (managing projects realized by students for private customers) and the EPFL Rocket Team (simulation team).

Compulsory: Military service in the swiss air force. Honorary rank for distinguished service as aviation soldier and accountant for 150+ people. Served 245/245 days.

Scholarships: Attended the competitive CIFAR DLRL Summer School 2024 in Toronto ($4k covered by the SNF). Recipient of the Hasler scholarship during my Master's thesis at MIT (50% of the total cost, amounting to $7k).

REFERENCES (upon request)

Roger Wattenhofer: Full professor at ETH Zürich, current supervisor.
Martin Jaggi: Associate Professor at EPFL, former supervisor.
Ramesh Raskar: Associate Professor at MIT, former supervisor.
Mohammad M. Amiri: Assistant Professor at Rensselaer Poly. Inst., co-author.