

FEDERICA GRANESE

☎ +33 (0) 767170673 ✉ federica.granese@inria.fr 🔗 <https://fgranese.github.io/> 🌐 <https://github.com/fgranese>

I am a permanent research scientist at Inria (ISFP), working on the evaluation of AI systems. My main research line develops principled methods to detect deceptive and manipulative content generated by AI, from sequential detection in text streams to manipulation in multi-agent settings. In parallel, I work on natural language processing, particularly on the design and use of topic models to analyze information ecosystems. I am also interested in tabular foundation models for medical applications. My background spans the reliability and security of machine learning, including misclassification detection and adversarial robustness.

Spoken Languages: Italian (native), English (fluent), French (fluent), Spanish (beginner).

EDUCATION

PhD in Computer Science **April 2023**

Towards Securing Machine Learning Algorithms through Misclassification Detection and Adversarial Attack Detection

<https://inria.hal.science/tel-04407139/>

École Polytechnique, Paris, France

Joint PhD with Sapienza University, Roma, Italy

M.S. in Computer Science **October 2019**

Enhanced models for privacy and utility in continuous-time diffusion networks

110/110 cum laude

Sapienza University, Roma, Italy

B.S. in Computer Science **October 2017**

Polynomially Recognising Graphs Where Saturating Flows are Always Maximum

106/110

Sapienza University, Roma, Italy

EXPERIENCE

Research Scientist (Inria Starting Faculty Position, ISFP) **October 2026 - Present**

Evaluation of AI Systems, Deception and Manipulation Detection, NLP

Inria Center at Université Côte d'Azur, Valbonne, France

- Developing a sequential detection framework for deceptive content in AI-generated text streams.
- Studying deception and manipulation in multi-agent and hybrid human-AI systems.

Fixed-term Researcher (Starting Research Position) **December 2024 - September 2026**

Topic Modeling, Natural Language Processing (NLP)

Inria Center at Université Côte d'Azur, Valbonne, France

- Developed new algorithms for online topic modeling on evolving text streams.
- Designed sentence-level topic models combining lexical and contextual evidence.

Postdoctoral Fellow **April 2023 - November 2024**

DeepECG4U - artificial intelligence in the service of heart health

Sorbonne Université, Paris, France

Institut de Recherche pour le Développement (IRD) UMMISCO, Bondy, France

- Enhance the robustness of Electrocardiogram classifiers specifically for arrhythmia prediction.
- Study the impact of Electrocardiogram denoising on automated classification and its consequential implications on the reliability of the models.
- Investigate generative models for Electrocardiogram reconstruction.

Research Internship **October 2022 - March 2023**

How to detect errors in image segmentation tasks

École de technologie supérieure (ÉTS), Montreal, Canada

- Conducted research on advanced models for segmenting tumors in MRI images.
- Investigated model calibration techniques and their impact on the misclassification detection task.

PhD Candidate

October 2019 - April 2023

Safety and Security in AI

Inria-Saclay, Paris, France

- Contributed to new research topics on misclassification detection, i.e., the problem of identifying whether the prediction of a DNN classifier should (or should not) be trusted.
- Contributed to new research topics on multi-armed adversarial attacks detection, i.e., the problem of identifying simultaneous adversarial attacks perpetrated over the DNN classifier.
- Conducted research on estimating the privacy level of various notions of differential privacy.
- Designed, developed, and maintained code repository to support result reproducibility.

Research Internship

March 2019 - July 2019

Diffusion of information in social networks

Inria-Saclay, Paris, France

- Conducted research on modeling the diffusion of information in social networks to meet privacy and utility guarantees.
- Contributed to a new formulation of utility/privacy policies, aiming to maximize the delivery of information to intended users while minimizing its reach to unintended ones.

Research Internship

March 2017 - September 2017

Algorithm design network flows

Sapienza University, Rome, Italy

- Conducted research on network models where information items flow from a source vertex to a sink vertex, specifically standard flow networks with capacity on edges.
- Contributed to a new polynomial-time algorithm for ensuring that every saturating flow under every capacity-to-edge assignment is maximum.

TEACHING

Technical Foundations of Artificial Intelligence. Lecturer, Master 2 *Droit algorithmique et gouvernance des données*, Faculty of Law, Université Côte d'Azur, Academic Year 2026-2027.

- *AI: Introduction to Deep Learning*. Lecturer and tutorials (Python). Université Côte d'Azur, Academic Year 2025-2026.

- *MARIANNE Journal Club (Exploring LLMs from the ground up, covering seminal papers, key advancements, and their limitations)*. Organizer and coordinator. Inria Center at Université Côte d'Azur, Academic Year 2024-2025.

- *Mécanismes de la programmation orientée-objet, Object-oriented programming mechanics*. TA (Java). École Polytechnique, Academic Year 2019-2020.

- *Mécanismes de la programmation orientée-objet, Object-oriented programming mechanics*. TA (Java). École Polytechnique, Academic Year 2020-2021.

PROGRAMMING SKILLS

Programming Languages: Python, Java, Bash, SQL, HTML.

Frameworks and Libraries: PyTorch, TensorFlow, SciKit.

Deployment: Git, Jupyter Notebooks.

Document Preparation Systems: LATEX, Markdown.

DBMSs MySQL, PostgreSQL.

Operative Systems: MacOS, Linux (Ubuntu Desktop), Windows.

RESEARCH FUNDING

- PEPR LMI (France 2030), Project 6 – Inria-VIGINUM joint laboratory (2026–2029). *Comparative analysis of topic coverage across sources: topic modeling and argument mining robust to source artifacts*.

- INESIA (2026–2028). *Detection of manipulative behaviors in AI systems*. Funds a 24-month postdoctoral position.

- ANR/AID ASTRID *Résistance* (2025). SHARD: *Hybrid AI-human systems for resistance to disinformation*. Senior participant, in charge of the machine learning components. Consortium: Inria MARIANNE (coordinator), Inria SIRA, Arlequin AI. 36 months (**387.000 Eur**).

AWARDS AND SCHOLARSHIPS

- IVADO-INRIA (2025): Semestre franco-qubécois d'échanges autour de l'intelligence artificielle. 5-weeks mobility (4.500 Eur).
- Prix de thèse en Apprentissage Automatique de la Société Savante Francophone en Apprentissage Machine (2024). *Honorable mention* (250 Eur).
- Bourse de recherche Mitacs Globalink (2022), 6 months mobility (6.000 CAD).
- Vinci, Chapter II (2020). Competition for financing thesis in cotutelle between Italy and France (4.700 Eur).
- Erasmus+ (AY 2018 – 2019), 6 months mobility (1.000 Eur).
- ACM Celebration of Women in Computing (womENCourage2019). Hackathon: AI and Welfare, 16-18 September 2019.

(SELECTED) PUBLICATIONS

Complete list available at: <https://scholar.google.com/citations?user=P3p6x5kAAAAJ>.

* Equal contribution.

Proceedings of International Conferences and Journal Papers

Stick-Breaking ETM with Continuous Optimal Transport for Online Analysis of Document Streams.

Federica Granese, Serena Villata, Charles Bouveyron.

AISTATS 2026.

ECGtizer: An Open-Source, Fully Automated Pipeline for Digitization and Signal Recovery from Paper ECGs.

Alex Lence, Ahmad Fall, Samuel David Cohen, Federica Granese, Jean-Daniel Zucker, Joe-Elie Salem, Edi Prifti.

Biomedical Signal Processing and Control, 2026.

Merging Embedded Topics with Optimal Transport for Online Topic Modeling on Data Streams.

Federica Granese, Benjamin Navet, Serena Villata, Charles Bouveyron.

ECML/PKDD 2025.

ECGrecover: A Deep Learning Approach for Electrocardiogram Signal Completion.

Alex Lence, Federica Granese*, Ahmad Fall*, Blaise Hanczar, Joe-Elie Salem, Jean-Daniel Zucker, Edi Prifti.*

KDD 2025.

On Estimating the Strength of Differentially Private Mechanisms in a Black-Box Setting.

Daniele Gorla, Louis Jalouzet, Federica Granese, Catuscia Palamidessi, Pablo Piantanida.

IEEE Transactions on Dependable and Secure Computing, 2025.

Optimal Zero-Shot Detector for Multi-Armed Attacks.

Federica Granese, Marco Romanelli*, Pablo Piantanida.*

AISTATS 2024.

A Halfspace-Mass Depth-Based Method for Adversarial Attack Detection.

Marine Picot, Federica Granese*, Guillaume Staerman, Marco Romanelli, Francisco Messina, Pablo Piantanida, Pierre Colombo.*

Transactions on Machine Learning Research (TMLR), 2023.

MEAD: A Multi-Armed Approach for Evaluation of Adversarial Examples Detectors.

Federica Granese, Marine Picot, Marco Romanelli, Francisco Messina, Pablo Piantanida.

ECML/PKDD 2022.

Enhanced Models for Privacy and Utility in Continuous-Time Diffusion Networks.

Federica Granese, Daniele Gorla, Catuscia Palamidessi.

International Journal of Information Security, 2021.

DOCTOR: A Simple Method for Detecting Misclassification Errors.

Federica Granese, Marco Romanelli*, Daniele Gorla, Catuscia Palamidessi, Pablo Piantanida.*

NeurIPS 2021 (Spotlight).

Workshops

TSPFN: A Temporal Tabular Foundation Model for Physiological Time Series Classification.

Jérémie Stym-Popper, Clément Rambour, **Federica Granese**, Nicolas Thome, Olivier Bernard.

STACOM Workshop, MICCAI 2026.

Adapting TabPFN for Zero-Inflated Metagenomics Data.

Giulia Perciballi, **Federica Granese**, Ahmad Fall, Farida Zehraoui, Edi Prifti, Jean-Daniel Zucker.

Table Representation Learning Workshop, NeurIPS 2024.

(SELECTED) ORAL COMMUNICATIONS

Unless otherwise specified, the formats are defined as follows: flash talks refer to short oral presentations of approximately 5 minutes; oral presentations refer to standard conference talks of approximately 15 minutes (excluding questions); seminar presentations refer to invited talks of approximately 45 minutes, followed by a discussion. Round tables denote participatory discussion formats.

Posters, Oral Presentations, and Flash Talks

- *When Pragmatic AI Agents Misalign: Safety Risks in Multi-Agent AI Systems – A Focus on Deceptive Behavior.* Inria-IVADO Closing Event, 2–3 December 2025, Montréal, Québec, Canada. **Oral Presentation**; English.

- *Towards Securing Machine Learning Algorithms from Single-Agent to Multi-Agent Systems.* Visit of the Minister of Higher Education of Quebec at the International Laboratory on Learning Systems (ILLS), Montréal, Québec, Canada. **Flash Talk**; French.

- *Towards Securing Machine Learning Algorithms through Misclassification Detection and Adversarial Attack Detection.* *Conférence sur l'Apprentissage Automatique* (CAp 2024), PhD Thesis Prize session (honorable mention, SSFAM prize), 1–3 July 2024, IMT Nord Europe and Université de Lille, Cité Scientifique Campus, France. **Poster and Flash Talk**; English.

- *Génération IA 2030 – Colloque européen sur l'Intelligence Artificielle.* 14–18 March 2022, Brussels, Belgium. **Flash Talk**; English.

- *DATAIA Workshop “Safety & AI”.* 13 December 2021, Inria Saclay, France. **Oral Presentation**; English.

- *Data Science & Artificial Intelligence Day (5th Edition).* 21 October 2021, Paris-Saclay, France. **Poster**; English.

Seminars

- *Topic Models for Online Analysis of Document Streams.* 3IA Côte d’Azur Seminar, 22 May 2026, Sophia Antipolis, France; English.

- *When Pragmatic AI Agents Misalign: Safety Risks in Multi-Agent AI Systems – A Focus on Deceptive Behavior.* Invited seminar, International Laboratory on Learning Systems (ILLS), 27 November 2025, Montréal, Québec, Canada. Invited seminar; English.

- *Multi-Armed Adversarial Attacks Framework.* Sosysec Seminar (Software and Systems Security), 31 May 2024, Inria Centre at the University of Rennes, France. Invited seminar; English.

Round tables

- Franco-Canadian Dialogue on AI. Round table on AI and Safety, 28 November 2025, Mila – Quebec AI Institute, Montréal, Québec, Canada; English.

- COMCYBER/IA Seminar: “*Fiabilité, confiance, éthique : quelle sécurité des IA ?*” 31 January 2024, Maison des Polytechniciens, Paris, France; French.