Federica Granese

91 Boulevard de l'Hôpital, 75013
Office 405 Paris – FRANCE
32 Av. Henri Varagnat, 93140
Bondy – FRANCE
✉ federica.granese@ird.fr

Visit my webpage:
🔗 www.lix.polytechnique.fr
/Labo/Federica.GRANESE/

Complete list of my publications:
🔗 dblp.org/pid/251/6090.html
🔗 Google Scholar

Scientific interests

○ Safety & Security in AI:
Misclassification Detection, Adversarial Robustness, OOD Detection
○ AI for Health

Education

2023. **PhD in Computer Science**
Institute Polytechnique de Paris
joint with Sapienza University
Paris – FRANCE

2019. **M. S. in Computer Science**
Curriculum: Software Engineering
110/110 cum laude
Sapienza University
Rome – ITALY

2017. **B. S. in Computer Science**
106/110
Sapienza University
Rome – ITALY

Languages

Italian (native)	●●●●
English	●●●○
French	●●○○

Programming skills

Languages (mainly used)	Python, Java, Bash, SQL, HTML, MATLAB, LaTeX
Frameworks	PyTorch, TensorFlow
DBMSs	MySQL, PostgreSQL
OS	MacOS, Linux Windows

Current situation

April 2023 – Sept. 2024. **Postdoctoral Fellow in Artificial Intelligence**

Institut de Recherche pour le Développement (IRD) - Unité de Modélisation Mathématique et Informatique de Systèmes Complexes (UMMISCO)

deepecg4u
Identifying patients at risk of Torsade-de-Pointes, a life-threatening arrhythmia using ECG and deep learning (`https://www.ummisco.fr/?page_id=9361`).
Advised by Edi Prifti & Jean-Daniel Zucker (IRD/Sorbonne University UMMISCO, FRANCE).

Experience

Nov. 2019 – Mar. 2023. **PhD Candidate in Computer Science**

Institute Polytechnique de Paris (FR) joint with Sapienza University (IT).

Securing Machine Learning Algorithms
The goal of the thesis project is to develop rigorous techniques for building safe and trustworthy AI systems. (Abstract: `http://www.theses.fr/s257648#`).
Advised by Catuscia Palamidessi (Lix, Inria, Institute Polytechnique de Paris, FRANCE) & Pablo Piantanida (L2S, CentraleSupélec, CNRS, Université Paris Saclay, FRANCE) and Daniele Gorla (Sapienza, ITALY).

Oct. 2022 – Feb. 2023. **Research Internship**

École de technologie supérieure (ÉTS). Montreal, Quebec H3C 1K3, Canada.

How to detect errors in image segmentation tasks.

Advised by José Dolz (ÉTS, Software and Information Technology Engineering Department, CANADA).

AY 2019 – 2020/2020 – 2021. **TA - Mécanismes de la programmation orientée-objet**

École polytechnique.

Teaching Assistant. The course presents the advanced mechanisms of programming languages, in particular the object-oriented features of Java.

Teaching coordinator: Benjamin Werner (École polytechnique, LIX – FRANCE).

Mar. 2019 – Jul. 2019. **Research Internship** (Erasmus+)

LIX, Inria - Saclay.

Modelling the diffusion of information in social networks to meet privacy and utility guarantees.

Advised by Catuscia Palamidessi (Lix, Inria, Institute Polytechnique de Paris, FRANCE).

Some Publications

Picot, M., Granese, F., Staerman, G., Romanelli, M., Messina, F., Piantanida, P., Colombo, P. **(2023)**. A Halfspace-Mass Depth-Based Method for Adversarial Attack Detection. *Transactions on Machine Learning Research (TMLR).*

Granese, F., Picot M., Romanelli M., Messina F., Piantanida P. **(2022)**. MEAD: A Multi-Armed Approach for Evaluation of Adversarial Examples Detectors. *ECML–PKDD 2022.*

Granese, F., Romanelli M., Gorla D., Palamidessi C., Piantanida P. **(2021)**. DOCTOR: A Simple Method for Detecting Misclassification Errors. *NeurIPS 2021* (**Spotlight**).

Granese, D., Gorla G., Palamidessi, C. **(2021)**. Enhanced Models for Privacy and Utility in Continuous-Time Diffusion Networks. (Journal extended version). *International Journal of Information Security, Springer Verlag.*

Gorla, D., Granese, F., Palamidessi, C. **(2019)**. Enhanced Models for Privacy and Utility in Continuous-Time Diffusion Networks. *ICTAC: 313-331.*

Federica Granese
91 Boulevard de l'Hôpital, 75013
Office 405 Paris – FRANCE
32 Av. Henri Varagnat, 93140
Bondy – FRANCE
✉ federica.granese@inria.fr

Visit my webpage:
www.lix.polytechnique.fr
/Labo/Federica.GRANESE/

Complete list of my publications:
dblp.org/pid/251/6090.html
Google Scholar

Oral communications

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD2022). 19-23 September 2022, Grenoble, France. Paper.

Génération IA 2030Colloque européen sur l'Intelligence Artificielle. 14-18 March 2022, DigitalCity.Brussels, Bruxelles. Pitch.

Conference on Neural Information Processing Systems (Neurips2021). 6-14 December 2021, Virtual. Paper.

DATAIA Workshop « Safety & AI » 2021. 13 December 2021, Centre Inria Saclay. Pitch.

Data Science & Artificial Intelligence DAY – 5eme Edition. 21 October 2021, Paris - Saclay. Poster and Pitch.

6th International Colloquium on Theoretical Aspects of Computing (ICTAC2019). Oct 30 - Nov 4 2019, Hammamet, Tunisia. Paper.

Awards and Scholarships

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 €).

ACM Celebration of Women in Computing (womENcourage2019). 16-18 September (**2019**), Rome, Italy. Challenge 3: AI and Welfare.

Erasmus+ (**AY 2018 – 2019**), 6 months mobility. (1.000 €).

Thesis summary

PhD thesis

Securing Machine Learning Algorithms

Deep Neural Networks (DNNs) have seen significant advances in recent years and are widely used in various applications. When it comes to safety-critical systems, developing methods and tools to make these algorithms reliable, particularly for non-specialists who may treat them as "black boxes" with no further checks, constitutes a core challenge. This thesis aims to investigate various methods that can enable the safe use of these technologies. In the first part, we tackle the problem of identifying whether the prediction of a DNN classifier should (or should not) be trusted so that, consequently, it would be possible to accept or reject it. In this re- regard, we propose a new detector that approximates the most powerful (Oracle) discriminator based on the probability of classification error with respect to the true class posterior probability. Two scenarios are investigated: Totally Black Box (TBB), where only the soft predictions are available, and Partially Black Box (PBB), where gradient-propagation to perform input pre-processing is allowed. The proposed detector can be applied to any pre-trained model. It does not require prior information about the underlying dataset and is as simple as the simplest available methods in the literature. We address in the second part the problem of simultaneous adversarial example detection. The detection methods are generally validated by assuming a single implicitly known attack strategy, which does not necessarily account for real-life threats. Indeed, this can lead to an overoptimistic assessment of the detectors' performance and may induce some bias in comparing competing detection schemes. To overcome this limitation, we propose a novel multi-armed framework for evaluating detectors based on several attack strategies. Among them, we use three new objectives to generate attacks. The proposed performance metric is based on the worst-case scenario: detection is successful if and only if all different attacks are correctly recognized. Moreover, following this setting, we formally derive a simple yet effective method to aggregate the decisions of multiple trained detectors, possibly provided by a third party. While every single detector tends to underperform or fail at detecting types of attack that it has never seen at training time, our framework successfully aggregates the knowledge of the available detectors to guarantee a robust detection algorithm. The proposed method has many advantages: it is simple as it does not require further training of the given detectors; it is modular, allowing existing (and future) methods to be merged into a single one; it is general since it can simultaneously recognize adversarial examples created according to different algorithms and training (loss) objectives.