Samuele Poppi

via Poliziano, 37, 41042 Fiorano Modenese, Modena, Italy

+39 347 113 6182

s.poppi94@gmail.com

samuele-poppi

Born January, 17th 1994

PROFESSIONAL SUMMARY

PhD in Artificial Intelligence at the **Universities of Pisa and Modena**, specialized in **Responsible and Safe AI** for Vision and Language. Completed a research internship at **Meta** with *Cristian Canton Ferrer*, *Oliver Aobo Yang*, *Jianfeng Chi*, and *Diego Garcia-Olano*. Skilled in **research project management**, now supervising two PhD students, and driving research on ethical, impactful AI solutions.

WORK EXPERIENCE

June 2025 – present

Postdoctoral Associate - AI Safety

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi
Supervisor: Prof. Nils Lukas

- AI Safety
- Responsible AI
- ML Security

May 2024 – November 2024

Research Scientist Intern @ Meta GenaAI Trust Team

Summer 2024

GenAI Trust and Safety @ Meta AI, Menlo Park, California, USA
Advisors: Cristian Canton Ferrer, Oliver Aobo Yang, Jianfeng Chi

- Safety of Multilingual LLMs
- Red Teaming of Large Language Models
- Llama Language Models

November 2021 – May 2025

PhD Candidate

Dottorato Nazionale in Intelligenza Artificiale, AI4Society, Università di Pisa and Università di Modena and Reggio Emilia

AImageLab - University of Modena And Reggio Emilia, Modena
Advisors: Prof. Rita Cucchiara and Lorenzo Baraldi

- Conducted research on responsible AI for LLMs, VLMs, and MLLMs, leading to 8 publications.
- Managed research projects and supervised two PhD students on Responsible AI.
- Completed a research internship at Meta, contributing to advancements in ethical AI systems.

June 2021 – September 2021

Research Fellow

HiPeRTLab - University of Modena And Reggio Emilia, Modena

- Computer vision algorithms for underwater and pick and place

February 2021 – June 2021

AI Engineer

HPE srl, Modena

- Computer vision algorithms, data analysis and signal processing

August 2020 – February 2021

Undergraduate Internship

AImageLab - University of Modena And Reggio Emilia, Modena

- Computer vision algorithms for Explainable and Responsible AI
- Supervisors: Prof. Rita Cucchiara, Dr. Lorenzo Baraldi, Dr. Marcella Cornia

PUBLICATIONS

Cappelletti S., Poppi T., Poppi S., Yong Z. X., Garcia-Olano D., Cornia M., Baraldi L., Cucchiara R., "Improving LLM First-Token Predictions in Multiple-Choice Question Answering via Prefilling Attack", *Under Review*, 2025.

Poppi S., Yong Z. X., He Y., Chern B., Zhao H., Yang A., Chi J., "Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks", In Findings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2025.

Poppi S., Poppi T., Cocchi F., Cornia M., Baraldi L., Cucchiara R., "Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models" In Proceedings of the European Conference on Computer Vision, Milano, Italy, 2024

Poppi S., Sarto S., Cornia M., Baraldi L., Cucchiara R., "Unlearning Vision Transformers without Retaining Data via Low-Rank Decompositions" In Proceedings of the 27th International Conference on Pattern Recognition, Kolkata, India, 2024

Poppi S., Sarto S., Cornia M., Baraldi L., Cucchiara R., "Multi-Class Explainable Unlearning for Image Classification via Weight Filtering" IEEE Intelligent Systems, 2024

Poppi S., Rawal N., Bigazzi R., Cornia M., Cascianelli S., Baraldi L., Cucchiara R., "Towards Explainable Navigation and Recounting" In Proceedings of the 22nd International Conference on Image Analysis and Processing, Udine, Italy, 2023, **Honorable Mention ICIAP Best Paper Award**.

Cocchi F.*, Baraldi L.*, Poppi S., Cornia M., Baraldi L., Cucchiara R., "Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis" In Proceedings of the 22nd International Conference on Image Analysis and Processing, Udine, Italy, 2023.

Poppi S., Cornia M., Baraldi L., and Cucchiara R. "Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2021.

RESEARCH MENTORSHIP

Silvia Cappelletti, Ph.D. student at Doctorate on Information and Communication Technology (University of Modena and Reggio Emilia).

- **PhD mentorship**: Supervising Silvia during her PhD, focusing on Responsible AI. Supporting research planning, project development, and publication efforts.

Tobia Poppi, Ph.D. student at Dottorato Nazionale in Intelligenza Artificiale, AI4Society (University of Pisa and University of Modena and Reggio Emilia).

- **MSc thesis mentorship**: mentored Tobia on his MSc thesis, titled "Towards Trustworthy AI: LLM Aligning for Offensive Content Removal", Poppi T., Poppi S., Cornia M., Baraldi L., Cucchiara R., 2023. Provided guidance on research methodology, technical implementation, and academic writing.
- **PhD mentorship**: Supervising Tobia during his PhD, focusing on Responsible AI. Supporting research planning, project development, and publication efforts.

Francesco Bellei, Master's student at the University of Modena and Reggio Emilia.

- Thesis title: "Class-wise Machine Unlearning for Vision and Swin Transformers"
- Year: 2023

Fabio Quattrini, Master's student at the University of Modena and Reggio Emilia, currently a Ph.D. student at Doctorate on Information and Communication Technologies (Università di Modena and Reggio Emilia).

- Thesis title: "Introducing Vision Transformers in Denoising Diffusion Probabilistic Models"
- Year: 2022

EDUCATION

November 2021 – May 2025

PhD Candidate in Artificial Intelligence

Dottorato Nazionale in Intelligenza Artificiale, AI4Society, Università di Pisa and Università di Modena and Reggio Emilia

AImageLab - University of Modena And Reggio Emilia, Modena
Advisors: Prof. Rita Cucchiara and Lorenzo Baraldi

- Responsible and Trustworthy AI
- Explainable Artificial Intelligence for Image Classification and Vision-and-Language Navigation
- Machine Unlearning techniques for Explainability
- Safety and Robustness benchmarking for LLMs

2018 – 2021

Master's degree in Computer Engineering

Department Of Computer Engineering "ENZO FERRARI" - University of Modena And Reggio Emilia

- Mark: 110/110 CUM LAUDE
- Thesis title: Towards Visual Explanations of CNNs: An Experimental Analysis
- Supervisors: Prof. Rita Cucchiara, Dr. Lorenzo Baraldi, Dr. Marcella Cornia
- Subjects: Computer Vision, Machine Learning and Deep Learning, Multimedia Processing Systems, Big Data Analysis, Robotics, Automotive Cyber Security

2013 – 2017

Bachelors's degree in Computer Engineering

Department Of Computer Engineering "ENZO FERRARI" - University of Modena And Reggio Emilia

- Mark: 89/110
- Thesis title: Modellistica e Controllo di un Generatore di Potenza Idraulica
- Supervisor: Prof. Roberto Zanasi

SKILLS

Languages

Italian – Mother tongue
English – European level B2
- Diploma of English (Level B2.3), Trinity College, 2013, European level B2

French – Level B1

Programming

Programming languages
- Python, C++, Matlab

Libraries
- PyTorch, Numpy, Hugging Face, Pandas, Scikit-Learn, OpenCV

RESEARCH INTERESTS

Explainable and Responsible Artificial Intelligence
Machine Unlearning
LLMs
MLLMs

PARTICIPATION TO RESEARCH GROUPS

October 2021 – present AImageLab - University of Modena And Reggio Emilia, Modena

June 2021 – September 2021 HiPeRTLab - University of Modena And Reggio Emilia, Modena

August 2020 – February 2021 AImageLab - University of Modena And Reggio Emilia, Modena

PARTICIPATION TO RESEARCH PROJECTS

December 2022 – present European Lighthouse of AI for Sustainability (ELIAS)

- Research activities: Research and design of next-generation content moderation and filtering algorithms. Study on the creation of robust datasets encompassing diverse content types, including text, images, and videos. Exploration of advanced AI techniques such as large language models (LLMs) and vision-language models (VLMs) for content classification. Focus on ethical guidelines, bias mitigation, and explainable AI to ensure transparency and user trust in moderation systems.
- co-funded by European Union (EU)

December 2022 – present European Lighthouse on Safe and Secure AI (ELSA)

- Research activities: Research and design of a challenge on Deep Fake Detection. Study on the generation of a suitable Dataset of real and fake images, prompt engineering and deep fake detection state-of-the-art models.
- co-funded by European Union (EU)

June 2021 – September 2021 Connected Electric Modular Powertrain (CEMP) Project

- Research activities: Study and design of a driving support system of a motorcycle for obstacle detection and consequent definition of warning signals to the driver.
- co-funded by Lombardia Region

INVITED TALKS AND SEMINARS GIVEN

2025 Seminar: "From Text to Vision: Ensuring Responsibility and Safety in Modern AI"

National PhD School in Artificial Intelligence – Responsible Generative AI course at **Scuola Normale Superiore di Pisa**

- Delivered a 2-hour lecture as part of the national PhD curriculum in AI4Society.
- Covered methodologies for red-teaming and blue-teaming of LLMs and Multimodal Models (VLMs, MLLMs).
- Included practical insights from research internship at Meta on multilingual model robustness and content moderation.
- Discussed mitigation techniques for Multimodal GenAI architectures.

TALKS AT INTERNATIONAL CONFERENCES AND WORKSHOPS

2024 Poster presentation at the 27th International Conference on Pattern (ICPR), Kolkata, India, 2024.

2024 Poster presentation at the 18th European Conference on Computer Vision (ECCV), Milan, Italy, 2024.

2023 Oral presentation at 22nd International Conference on Image Analysis and Processing, Udine, Italy, 2023.

2023 Poster presentation at 22nd International Conference on Image Analysis and Processing, Udine, Italy, 2023.

2021 Oral presentation at the "Responsible Computer Vision Workshop", IEEE/CVF Conference on Computer Vision and Pattern Recognition.

2021 Poster presentation at the "Responsible Computer Vision Workshop", IEEE/CVF Conference on Computer Vision and Pattern Recognition.

INTERNATIONAL CONFERENCES AND WORKSHOPS ATTENDED

2024 International Conference on Pattern Recognition (ICPR)

2024 European Conference on Computer Vision (ECCV)

2023 IEEE/CVF International Conference on Computer Vision (ICCV)

2023 22nd International Conference on Image Analysis and Processing (ICIAP)

2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

SEMINARS AND COURSES ATTENDED

2023 "Responsible AI in the age of GenAI: the case of Llama2", Cristian Canton Ferrer (Meta), ELLIS Summer School 2023

2023 "An Egocentric Approach to Social AI", Jim Rehg (UIUC), ICVSS 2023

2023 "Machine learning meets physical models of image formation for photography and astronomy applications", Jean Ponce (Ecole normale supérieure-PSL, NYU), ICVSS 2023

2023 "Motion and sounds: Using video to reveal invisible motions and to suppress unwanted sounds", William T. Freeman (MIT), ICVSS 2023

2023 "Computer Vision after the Victory of Data", Alyosha Efros (UC Berkeley), ICVSS 2023

2023 "Long-Range Video Understanding", Angela Yao (National University of Singapore), ICVSS 2023

2023 "Learning to Understand Video Through Language", Lorenzo Torresani (FAIR, Meta), ICVSS 2023

2023 "Robot Learning In The Wild: Continual Improvement by Watching and Practicing", Deepak Pathak (CMU), ICVSS 2023

2023 "A Quiet Revolution in Robotics", Vladlen Koltun (Apple), ICVSS 2023

2023 "The sensorimotor road to artificial intelligence", Jitendra Malik (UC Berkeley), ICVSS 2023

2023 "Visual Localization or Where Am I and Who Knows", Torsten Sattler (CTU), ICVSS 2023

2023 "From Videos to 4D Worlds and Beyond", Angjoo Kanazawa (UC Berkeley), ICVSS 2023

2023 "Structure from Motion and Neural Radiance Fields", Frank Dellaert (Georgia Tech), ICVSS 2023

2023 "Foundational Issues in AI: Views from the real and the ideal worlds", Stefano Soatto (UCLA, AWS), ICVSS 2023

2023 "GPT-4 from Scratch", Andrej Karpathy (OpenAI), ICVSS 2023

2023 "Generative Models as Data++", Phillip Isola (MIT), ICVSS 2023

2023 "Advanced AI approaches to Digital Humanities applications and beyond", Silvia Cascianelli, Angelo Porrello (UNIMORE)

2022 "GENDER UNBALANCED AI", Silvia Zuffi (IMATI-CNR)

2022 "From Handcrafted to End-to-End Learning, and Back: a Journey far Multi-Object Tracking", Laura Leal-Taixé (NVIDIA and University of Munich)

2022 "Graph Signal Processing for Machine Learning: Challenges and Use-cases", Laura Toni (University College of London)

2022 "Machine Learning applications in trading and Portfolio management", Petter Kolm (New York University)

2022 "Hyperbolic deep Learning", Pascal Mettes (University of Amsterdam)

2022 "Explainability in Artificial Intelligence 1-2", Natalia Díaz Rodriguez (University of Granada)

2022 "Bringing Perception to Social Robots: An introductory course", Paolo Rota, Yiming Wang (University of Trento),
Hours: 20, Mark: 30/30

2022 "Multimodal Machine Learning", Wei Wang, Cigdem Beyan (University of Trento),
Hours: 20 Mark: 10/10

2022 "Explainable Artificial Intelligence", Fosca Giannotti (University of Pisa),
Hours: 30, Mark: 30/30 CUM LAUDE

2021 "Research in Videogames: Use of Deep Learning for Saliency Estimation and Cheating Prevention", Dr. Iuri Frosio (NVIDIA)

2020 "Deep Scene Perception without Labeled Data", Prof. Luigi Di Stefano (University of Bologna)

2020 "Sportscar Vehicle Architecture: the starting point of conceiving a new car", Ferrari S.p.A.

2020 "The use of artificial intelligence in the automotive industry: products and processes", Ing. Marco Fainello (Executive Director - Addfor S.p.a.)

Service

Reviewer: ACM MM (2023), CVPR (2024), ECCV (2024), ACM MM (2024)
Organizing Committee: Workshop at ECCV 2024 ("Trust What You learN" Workshop about Trustworthy AI, Machine Unlearning and Deepfake Detection)

Awards

Honorable Mention for Best Paper Award at the 22nd International Conference on Image Analysis and Processing, Udine, Italy, 2023.