Tobia **Poppi**

PhD Candidate · Artificial Intelligence Engineer

Modena, Italy

tobia.poppi@gmail.com | tobiapoppi.github.io | tobiapoppi | tobia-p-890793140

"Empowering AI to be a Safe & Trustworthy tool for Society."

Summary

I'm a PhD candidate in the National PhD in Artificial Intelligence at the University of Pisa, specializing in **AI Safety**, **Responsible AI**, and **Trustworthy AI**. My research focuses on advancing **Generative AI** and **Multimodal architectures**, aiming to bridge the gap between cutting-edge deep learning technologies and ethical alignment with human values. Passionate about leveraging AI to solve complex societal challenges, I am dedicated to ensuring AI systems are **safe**, **transparent**, and **aligned** with responsible principles, driving innovation with a human-centric approach.

Knowledge and Technical Skills

Topics	Responsible AI, Safe AI, Generative AI, Multimodal Learning, LLM & LMM Finetuning, NLP, Deep Learning
Frameworks and Other	PyTorch, Numpy, Pandas, Scikit-learn, Slurm, Git, Unix, OpenCV, Tensorflow
Programming	Python, Bash, C++, C, JAVA, LaTeX

Experience

AImageLab - University of Modena and Reggio Emilia
Modena, Italy

PhD Candidate — *Nov. 2023 - Present*

- Research during the Doctorate at AImageLab covers these topics in Artificial Intelligence: Multimodal Learning, Generative AI, Responsible AI, AI safety, LLMs Finetuning, Natural Language Processing.
- Fine-tuned an open source LLM to convert between safe and unsafe sentences, starting from a small self-curated dataset.
- Designed and implemented a new CLIP-like Multimodal architecture to prevent unsafe and inappropriate content retrieval and generation.
- Author of a Large-scale Multimodal dataset in AI Safety and Responsible AI domain.

AImageLab - University of Modena and Reggio Emilia
Modena, Italy

Research Fellow Intern — *May 2023 - Nov. 2023*

- Curricular intern in Trustworthy AI for Deep generative models.

Hipert Lab - University of Modena and Reggio Emilia
Modena, Italy

Research Fellow — *Feb. 2022 - Aug. 2022*

- Implementation of geometric methodologies based on the model and implementation of machine learning techniques, driven by data, in the context of 6D pose estimation.

Hipert Lab - University of Modena and Reggio Emilia
Mantova, Italy

Research Fellow Intern — *May 2021 - Sep. 2021*

- Internship for Bachelor's Degree; Thesis Title: "Labeling: the best tools available and their application for detection of people e road signs and for poses estimation of 3D objects".

GPI s.r.l.
Modena, Italy

Computer Assistant — *Jun. 2017 - Aug. 2017*

- Computer maintenance, repair and installation at the Modena Polyclinic Hospital.

Publications

2025 Conference on Computer Vision and Pattern Recognition (CVPR) — *Nashville, Tennessee, USA*

Hyperbolic Safety-Aware Vision-Language Models

Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, Rita Cucchiara

Addressing the retrieval of unsafe content from vision-language models such as CLIP is an important step towards real-world integration. Current efforts have relied on unlearning techniques that try to erase the model's knowledge of unsafe concepts. While effective in reducing unwanted outputs, unlearning limits the model's capacity to discern between safe and unsafe content. In this work, we introduce a novel approach that shifts from unlearning to an awareness paradigm by leveraging the inherent hierarchical properties of the hyperbolic space. We propose to encode safe and unsafe content as an entailment hierarchy, where both are placed in different regions of hyperbolic space. Our HySAC, Hyperbolic Safety-Aware CLIP, employs entailment loss functions to model the hierarchical and asymmetrical relations between safe and unsafe image-text pairs. This modelling – ineffective in standard vision-language models due to their reliance on Euclidean embeddings – endows the model with awareness of unsafe content, enabling it to serve as both a multimodal unsafe classifier and a flexible content retriever, with the option to dynamically redirect unsafe queries toward safer alternatives or retain the original output. Extensive experiments show that our approach not only enhances safety recognition, but also establishes a more adaptable and interpretable framework for content moderation in vision-language models.

2024 European Conference on Computer Vision (ECCV) *Milano, Italy*

Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara

Large-scale vision-and-language models, such as CLIP, are typically trained on web-scale data, which can introduce inappropriate content and lead to the development of unsafe and biased behavior. This, in turn, hampers their applicability in sensitive and trustworthy contexts and could raise significant concerns in their adoption. Our research introduces a novel approach to enhancing the safety of vision-and-language models by diminishing their sensitivity to NSFW (not safe for work) inputs. In particular, our methodology seeks to sever "toxic" linguistic and visual concepts, unlearning the linkage between unsafe linguistic or visual items and unsafe regions of the embedding space. We show how this can be done by fine-tuning a CLIP model on synthetic data obtained from a large language model trained to convert between safe and unsafe sentences, and a text-to-image generator. We conduct extensive experiments on the resulting embedding space for cross-modal retrieval, text-to-image, and image-to-text generation, where we show that our model can be remarkably employed with pre-trained generative models.

Education

University of Pisa *Modena, Italy*

PhD in Artificial Intelligence; *National PhD in Artificial Intelligence for Society* *Nov. 2023 - Present*

University of Modena and Reggio Emilia *Modena, Italy*

Master's Degree in Computer Engineering; *Grade: 110/110 with honours* *Sep. 2021 - Oct. 2023*

Norwegian University of Science and Technology *Trondheim, Norway*

Erasmus+ Exchange Semester *Aug. 2022 - Dec. 2022*

University of Modena and Reggio Emilia *Mantova, Italy*

Bachelor's Degree in Computer Engineering; *Grade: 105/110* *Sep. 2018 - Oct. 2021*

Enrico Fermi Technical Industrial Institute *Modena, Italy*

High school Diploma in IT and telecommunications *Sep. 2013 - Oct. 2018*

Participation to National and European Projects

- Horizon Europe project **"ELIAS - European Lighthouse of AI for Sustainability"**, co-funded by the European Union
- PNRR project **"FAIR - Future Artificial Intelligence Research"**, co-funded by the European Union
- Horizon Europe project **"ELSA - European Lighthouse on Safe and Secure AI"**, co-funded by the European Union

Program Committees

2024 **Peer Reviewer,** Computer Vision and Image Understanding (CVIU) - Journal

2024 **Peer Reviewer,** Pattern Recognition (PR) - Journal

2024 **Peer Reviewer,** European Conference on Computer Vision (ECCV) *Milan, Italy*

2024 **Peer Reviewer,** IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) *Seattle, USA*

Languages

Italian Mother tongue

English Writing: *Fluent*, Reading: *Fluent*, Listening: *Fluent*

Interests

Professional Artificial Intelligence, Responsible AI, Generative AI, AI Safety, Trustworthy AI, Alignment Problem, Computer Vision

Personal Pool-Billiard, Sport, Music, Art, Cinema