AI SAFETY RESEARCH PORTFOLIO

Safety research for systems under pressure.

Adversarial evaluation, multimodal alignment, model robustness, and scalable safety methodologies for generative AI systems.

Samuele Poppi

Postdoctoral Associate in AI Safety at MBZUAI

01 RISK	Deployment-aware threat modelling
02 EVALUATION	Benchmarks, datasets, and protocols
03 MODALITIES	Text, image, and multimodal systems
04 DELIVERY	Research leadership and clear decisions

Website Scholar GitHub Email

EXECUTIVE PROFILE

Research depth, delivery discipline.

A senior research profile built around rigorous evaluation and clear safety decisions.

Samuele is a Postdoctoral Associate at MBZUAI with a PhD in Artificial Intelligence focused on Responsible AI for vision and language. His research spans AI safety, security, robustness, constitutional AI, multilingual safety, multimodal systems, and model evaluation.

His work turns safety questions into testable methodologies: benchmarks, evaluation frameworks, adversarial tests, datasets, and experimental pipelines. He supervises MSc and PhD researchers, coordinates collaborative projects, reviews experimental quality, and translates technical findings into decision-relevant recommendations.

CORE EXPERTISE

A cross-modal safety toolkit.

Technical evaluation, system security, and research leadership.

GenAI red teaming	Adversarial evaluation	Multimodal AI safety	Constitutional AI
Model robustness	Multilingual safety	Safety benchmarking	Machine unlearning
AI security	Watermarking	Evaluation design	Dataset design
Research leadership	Technical communication		

SELECTED RESEARCH

Safety questions turned into testable systems.

Case studies spanning post-alignment drift, multimodal safeguards, multilingual robustness, and AI security.

01

Gravitational Fine-Tuning Reversion

ARXIV 2026

Training-history geometry of post-alignment drift

RESEARCH QUESTION Can fine-tuning reversion be explained through training-history geometry and causally reduced without sacrificing task utility?

PROBLEM	SAFETY RISK
Harmless fine-tuning can partially undo alignment, revive unlearned capabilities, or restore latent behaviours from earlier training.	Benign post-alignment updates may reintroduce harmful behaviour while preserving downstream task performance.
MY CONTRIBUTION	**MAIN FINDING**
First author; connected fine-tuning reversion to training-history geometry and evaluated v_rev as a causal mediator.	Drift rapidly aligns with v_rev; blocking that direction reduces harmfulness with little task cost in the evaluated setup.

Project **Paper**

02

SPQR

ECCV 2026

Safety alignment under benign model adaptation

RESEARCH QUESTION How should safety alignment be evaluated when text-to-image systems undergo benign deployment-time adaptation?

PROBLEM	SAFETY RISK
Safety controls can weaken after ordinary fine-tuning even when downstream data and tasks are benign.	A customized model may remain useful on standard quality measures while becoming less reliable on safety.
MY CONTRIBUTION	**MAIN FINDING**
Co-authored the work and advised benchmark design, experimentation, and publication development.	Safety alignment must be re-evaluated after benign adaptation rather than treated as a fixed pre-deployment property.

Project **Paper**

SELECTED RESEARCH

Safety questions turned into testable systems.

Case studies spanning post-alignment drift, multimodal safeguards, multilingual robustness, and AI security.

03

Safe-CLIP

ECCV 2024

Removing NSFW concepts from vision-language models

RESEARCH QUESTION Can unsafe concepts be removed from CLIP representations while preserving useful vision-language capabilities?

PROBLEM	SAFETY RISK
Vision-language representations can preserve unsafe concept associations that propagate to downstream systems.	Representation-level associations can make harmful content difficult to filter without damaging useful behaviour.
MY CONTRIBUTION	**MAIN FINDING**
First author; led the research, experimental evaluation, and technical communication.	Safe-CLIP reduces unsafe concept associations while preserving useful vision-language behaviour across evaluated tasks.

Project **Paper** **Code**

04

Multilingual AI Safety at Meta GenAI

MAY-NOV 2024

Industry research experience

RESEARCH QUESTION How robustly does safety behaviour transfer across languages, and how does adaptation change that robustness?

PROBLEM	SAFETY RISK
Safety behaviour that is reliable in one language may not transfer consistently or survive later fine-tuning.	Uneven safeguards create cross-lingual attack surfaces in globally deployed systems.
MY CONTRIBUTION	**MAIN FINDING**
Conducted research on multilingual LLM safety, red teaming, and Llama models within Meta GenAI Trust and Safety.	Public research shows that fine-tuning attacks in one language can degrade multilingual safety alignment.

Public output **Paper**

SELECTED RESEARCH

Safety questions turned into testable systems.

Case studies spanning post-alignment drift, multimodal safeguards, multilingual robustness, and AI security.

05

Fragility of Multilingual LLMs

NAACL 2025

Post-deployment robustness

RESEARCH QUESTION How does fine-tuning affect safety across languages, and what does that imply for post-deployment governance?

PROBLEM	SAFETY RISK
Fine-tuning can alter established safety behaviour after a model leaves its original training environment.	Adaptation may create multilingual degradation that standard pre-deployment tests do not capture.
MY CONTRIBUTION	**MAIN FINDING**
First author; led the multilingual attack study, experimental analysis, and communication of safety implications.	Attacks in one language can break safety alignment across languages, suggesting partly language-agnostic safety information.

Project **Paper**

06

AI Security, Watermarking and Forensics

2024-2026

Adversarial research portfolio

RESEARCH QUESTION Which security and authenticity claims remain valid under fine-tuning, forgery attempts, and adaptive attacks?

PROBLEM	SAFETY RISK
Security mechanisms can fail under model modification or threat models that differ from their original evaluation.	Watermarks, monitors, and authenticity signals may be bypassed or forged under adaptive pressure.
MY CONTRIBUTION	**MAIN FINDING**
Co-authored work on activation watermarking and randomized-key defenses; mentored watermarking and deepfake research.	Robustness under adaptation and adaptive attack should be part of the security claim, not a secondary check.

Activation watermarking **Watermark forgery**

RED-TEAMING METHODOLOGY

From system context to retested mitigation.

A general research approach for evaluating generative AI across languages, modalities, and model versions.

01

Define the system

Map capabilities, interfaces, users, constraints, and deployment context.

02

Construct the threat model

Name assets, adversaries, access, incentives, and security boundaries.

03

Map failure modes

Prioritise likely misuse, safety failures, and high-consequence edge cases.

04

Design adversarial cases

Build prompts and scenarios for direct, indirect, and adaptive attacks.

05

Create evaluation data

Curate reproducible datasets with coverage across risks and user contexts.

06

Define success criteria

Combine quantitative metrics with structured qualitative review.

07

Stress boundaries

Compare languages, modalities, model versions, and system configurations.

08

Analyse root causes

Move from failure counts to vulnerability patterns and causal hypotheses.

09

Propose mitigations

Match controls to failure mode, deployment layer, and residual risk.

10

Retest

Verify mitigations and check for regressions or displaced risk.

11

Communicate decisions

Translate evidence into findings for technical and senior stakeholders.

General research methodology for evaluating generative AI systems.

LEADERSHIP AND DELIVERY

Raise the quality of the work - and the team doing it.

Evidence from supervision, methodology design, experimental review, teaching, and collaborative delivery.

MENTOR RESEARCHERS

Supervise MSc, PhD, and independent researchers across constitutional alignment, world-model jailbreaking, content safety, watermarking, and deepfake analysis.

DESIGN METHODOLOGY

Define threat models, datasets, baselines, metrics, ablations, and quality gates before experiments scale.

REVIEW FOR QUALITY

Review experimental design, implementation, failure cases, technical writing, and publication strategy.

COORDINATE DELIVERY

Align parallel research tracks through responsibilities, review points, and shared standards.

TEACH AND TRANSLATE

Deliver graduate lectures and turn technical evidence into risk, limitation, and mitigation statements.

SET DIRECTION

Translate open-ended safety questions into scoped hypotheses and sequenced workstreams.

EXPERIENCE

Academic rigor, industry context.

A path connecting Responsible AI foundations with current generative and multimodal safety work.

MBZUAI

Postdoctoral Associate in AI Safety

June 2025 - Present | Abu Dhabi, UAE

AI safety and security for language, vision-language, and generative models; MSc and PhD supervision; graduate teaching.

META GENAI TRUST AND SAFETY

Research Scientist Intern

May 2024 - November 2024 | Menlo Park, California, USA

Multilingual LLM safety, cross-lingual red teaming, fine-tuning attacks, and Llama models.

UNIVERSITY OF PISA AND UNIVERSITY OF MODENA-REGGIO EMILIA

PhD in Artificial Intelligence - AI4Society

November 2021 - May 2025 | Pisa and Modena, Italy

Responsible AI in Vision and Language: Ensuring Safety, Ethics, and Transparency in Modern Models. Completed with honors.

SELECTED PUBLICATIONS

Research outputs by safety domain.

Peer-reviewed papers and current preprints across AI safety, security, and multimodal systems.

01 ECCV 2026

SPQR: A Multi-Dimensional Benchmark for Safety Alignment under Benign Model Adaptation

Mohammed Talha Alam, Nida Saadi, Fahad Shamshad, Nils Lukas, Karthik Nandakumar, Fakhri Karray, Samuele Poppi

02 ARXIV 2026

A Gravitational Interpretation of Fine-Tuning Reversion

Samuele Poppi, Nils Lukas

03 ICML FAGEN WORKSHOP 2026

When the Chain of Thought Knows Better: Failure Modes in Multi-Turn Reasoning Models

Sai Kartheek Reddy Kasu, Nils Lukas, Samuele Poppi

04 ARXIV 2026

Robust Safety Monitoring of Language Models via Activation Watermarking

Toluwani Aremu, Daniil Ognev, Samuele Poppi, Nils Lukas

05 NAACL FINDINGS 2025

Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, Jianfeng Chi

06 ECCV 2024

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

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

TALKS AND SERVICE

Research in the room.

Invited lectures, conference activity, peer review, and community-building for trustworthy AI.

February 2026	**Not All Attackers Are Malicious: When Safety Degrades Without Harmful Intent**	Symposium on Security in the Age of AI, MBZUAI
May 2025	**From Text to Vision: Ensuring Responsibility and Safety in Modern AI**	National PhD School in AI, Scuola Normale Superiore
2023-2025	**Peer review**	ACM MM, CVPR, ECCV, ICCV, and BMVC
2024-2026	**Trustworthy AI workshop organization**	ECCV 2024 and MBZUAI 2026

CONTACT

Let's make safety claims testable.

Available for conversations about AI safety research, red teaming, multimodal evaluation, and research leadership.

Email **LinkedIn** **GitHub** **Scholar** **Website**