giovanni occhipinti

AI safety researcher, red-teaming AI control measures in production

giovanni.occhipinti@ephance.de linkedin.com/in/occhipintigiovanni github.com/Joe-Occhipinti

PROFILE

I research how to **red-team AI control measures**. As an ERA fellow I led a red-teaming audit of Codex and Auto-review, OpenAI's action-blocking monitor, shared the findings with OpenAI, and developed a **methodology and attack taxonomy** for production harnesses. Earlier work at SPAR and Oxford produced two **workshop papers**, at NeurIPS 2026 and ICLR 2026.

EXPERIENCE

AI Safety Research Fellow, ERA Fellowship

Cambridge · July – November 2026

- Led a red-teaming audit of Codex and Auto-review, OpenAI's action-blocking monitor; shared findings and patch recommendations with OpenAI.
- Developed a methodology and attack taxonomy for red-teaming control measures in production harnesses.
- Built a benchmark to measure how oversight awareness uplifts models' attack capabilities against monitors.

AI Safety Research Fellow, SPAR

March – May 2026

- Investigated self-prediction in LLM agents over long-horizon trajectories, including misaligned behaviour; workshop paper accepted at NeurIPS 2026.
- Designed the agent scaffolds, evaluation tasks and environments for it.

Visiting AI Safety Researcher, OXCAV

University of Oxford · September 2025 – January 2026

- Investigated activation probing and steering for chain-of-thought unfaithfulness in reasoning models; built a representation engineering pipeline (probes, Contrastive Activation Addition, MLP-guided intervention).
- Elicited unfaithful behaviour through adversarial prompting on MMLU.
- Cosmos Institute research grant; workshop paper accepted at ICLR 2026.

PUBLICATIONS

Lessons From Red-teaming Auto-review, OpenAI's Blocking Monitor
Giovanni Maria Occhipinti, Monika Jotautaitė [post]

Open Problems in LLM Self-Prediction
Lydia Nottingham, Giovanni Maria Occhipinti, Andrew Tran [PDF]
NeurIPS 2026 workshop

Probing and Steering Chain-of-Thought Unfaithfulness in Language Models
Giovanni Maria Occhipinti, Alessandro Abate, Nandi Schoots [PDF]
ICLR 2026 workshop

STRENGTHS

AI control evaluations: red-teaming of control measures

White-box control: probing, steering and representation engineering

Agent scaffolds and evals: Inspect, ControlArena, LinuxArena

Tooling: Python, R, PyTorch, TransformerLens, Docker, vLLM, RunPod

EDUCATION

MSc Data, Methods and Models for Linguistics
University of Bologna · GPA 4.00
2023 – 2025
Thesis research at Oxford; Erasmus at UCD

BA Humanities
University of Bologna
2019 – 2023

SELECTED PROJECTS

Chiron
Agentic plugin for setting design in ControlArena · Apart Research hackathon, top 25%