Stefano Zuffi

Amsterdam, Netherlands
stefano.zuffi06@outlook.com
+39 347 399 6906

PROGRAMMING

Python
PyTorch, TransformerLens, einops, NumPy, Pandas, Scikit-learn

JavaScript
Node.js, React

Java
Spring

TOOLS

LaTeX · Git · Jupyter / Colab · Excel

LANGUAGES

Italian
Native

English
Fluent

PROFILE

Logic (MSc) and Philosophy (BA) with research training in formal semantics, model theory, and the conceptual analysis of abstract objects. Currently transitioning into AI safety, with a focus on the foundational and existential-risk dimensions of alignment, and active in building the Dutch AI safety community (AISA, SAIN).

Determined to understand with more depth and breadth artificial intelligence, connecting the dots between the experimental findings, to bear on technical and governance questions in AI x-risk.

RESEARCH INTERESTS

AI x-risk and alignment; conceptual foundations of agency, goals, and value specification; mathematical logic and formal semantics; mechanistic interpretability; philosophy of AI

EDUCATION

MSc in Logic & Mathematics, Universiteit van Amsterdam *Sep 2022 – Apr 2025*

Coursework: Set Theory, Model Theory, Modal Logic, Abstract Algebra, Probability & Statistics. Thesis: *Arbitrary Terms with no Arbitrary Objects* (see Research Experience).

BA in Philosophy, Catholic University of the Sacred Heart, Milan *Sep 2019 – Sep 2022*

Thesis: *GL Modal Calculus and its Applications in Recursion Theory* – modal logic meets computability theory.

RESEARCH EXPERIENCE

MSc Thesis: *Arbitrary Terms with no Arbitrary Objects*, UvA *2024 – Apr 2025*

- Investigated the semantics of arbitrary terms by comparing existing approaches in the literature and proposing a novel "quasi-referential" account
- Combined formal logic with conceptual analysis to clarify the foundations of abstract reference

Research Project: Topology of Concepts, Universiteit van Amsterdam *2023*

Applied topological methods to concept formation following Gärdenfors' conceptual spaces; investigated geometric and topological properties of conceptual representations.

AI SAFETY

Fellow, AFFINE Superintelligence Alignment Seminar, Czech Republic *Upcoming: May 2026*

Selective month-long residential program on the core problems of superintelligence alignment as an existential risk, with mentors from MIRI, DeepMind, Astera Institute, and Mila. Curriculum centered on deep technical understanding through reading, peer teaching, and collaborative debate.

Founding Team Member, Safe AI Netherlands (SAIN) *April 2026 – Present*

Part of the founding team of SAIN, a new Dutch national foundation (*stichting*) formed by merging the AI Safety Initiatives of Amsterdam, Groningen, and Utrecht. SAIN aims to coordinate research, education, and public discourse on AI risks across the Netherlands.

Active Member, AI Safety Amsterdam (AISA), UvA *Dec 2025 – Present*

Participate in meetings, talks, and reading groups within the UvA AI safety community.
Completed BlueDot Impact's Technical AI Safety course (AISA cohort)

Self-directed AI Safety Training *2025 – Present*

Self-taught machine learning and deep learning basics
Currently working through ARENA's mechanistic interpretability curriculum and current alignment/interpretability literature.

TALKS

Speaker, Neuchâtel Undergraduate Philosophy Conference *Oct 2021*

Talk: *Presentism vs Special Relativity* – presentism and relativistic spacetime.

WORK EXPERIENCE

Data Analysis Intern, Nibe s.r.l., Milan *May – Sep 2025*

Logic Tutor, Catholic University of Milan *Oct – Dec 2021*