Mette Friis Andersen

m.f.andersen@uva.nl | github.com/mettefriis

University of Amsterdam
Institute of Logic, Language and Computation (ILLC)
Amsterdam, Netherlands

Research Areas

Mechanistic Interpretability, Faithfulness

Education

PhD Neurosymbolic AI, University of Amsterdam — September 2026 - current

Currently studying whether VLMs implement binding using tensor product representations.
Supervisor: Martha Lewis

MSc Logic and Computation, University of Amsterdam — September 2026 - September 2026

Thesis: *Causally sufficient, semantically misaligned*
Supervisors: Sandro Pezzelle, Giovanni Cina

BSc Philosophy, Logic and Scientific Method, LSE — September 2020 - August 2023

Classification: First-class Honours
Thesis: *On the Logical Character of Emotive Language: Evaluating Williamson's Anti-luminosity Argument* (awarded 80, high distinction)
Supervisor: Anna Mahtani

Papers

Andersen, M. F., Cina, G. & Pezzelle, S. (2026). "Causal Sufficiency Without Semantic Alignment: How Causal Subspaces Can Masquerade as Semantic Concepts". ICML 2026 Workshop on Mechanistic Interpretability.

Andersen, M. F., Li, S. & Lewis, M. (2026). "The Geometry of Metaphor". Proceedings of the 9th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP

Andersen, M. F., Heuss, M. & Lucic, A. (2025). "Three Desiderata for Faithfulness in Machine Learning Explanations: The Case for Causal Abstraction." 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability

Andersen, M. F. (2024). "On the Logical Character of Emotive Language: Evaluating Williamson's Anti-luminosity Argument." Apophansis (King's College London's peer-reviewed journal in philosophy)

Summer schools

Summer School in Logic and Formal Epistemology, Carnegie Mellon University (CMU) — June 2024

TimFest 2023, Magdalen College, University of Oxford — June 2023

Talks

Towards a probabilistic model of ordinary ability knowledge: An Edgingtonian reply to Vetter
Metaphysics workshop (2023), University College London

Williamson's anti-luminosity argument
Epistemology workshop (2022), London School of Economics and Political Science

Grants and Awards

Viet-Jacobsen Fonden (15.000 DKK towards studies at University of Amsterdam),
Aage og Johanne Louis-Hansens Fond (20.000 DKK towards studies at University of Amsterdam)

Reviewer service

Reviewer, ICML MechInterp Workshop (expertise: Methods (probing, steering, causal interventions), Concept Discovery (e.g., SAEs, dictionary learning)