dellago.matt@gmail.com
matthiasdellago.github.io
matthiasdellago
etuVuHoAAAAJ
matthiasdellago

Matthias Dellago

Research Focus

I am developing rigorous theoretical foundations for deep learning by bridging statistical mechanics, complexity theory and algorithmic information theory. While neural networks demonstrate remarkable regularities, such as scaling laws, their theoretical understanding remains pre-paradigmatic. I focus on finding physically realizable alternatives to classical Turing machine-based learning theory, drawing on statistical mechanical principles to model learning in finite systems. Just as thermodynamics transformed dangerous steam engines into controllable forces of progress, foundational theory will enable us to harness artificial intelligence safely and reliably.

Experience

Fall 2024	**Visiting Member**, *London Initiative for Safe AI* Grounding deep learning in algorithmic information theory, connecting to singular learning theory.
2024	**Guest Researcher**, *Institute for Machine Learning, Johannes Kepler University Linz* Interpretability of attention weight decay and sparsity. Loss landscape roughness analysis.
Winter 2023	**Guest Researcher**, *Amsterdam Machine Learning Lab, University of Amsterdam* Interpretability of attention, and new interpretable architectures.
2021-2022	**Researcher**, *Information Security and Privacy Lab*, University of Innsbruck Joint project with Oxford: Economic analysis of the 0-day Grey Market.
2020	**Instructor of Undergraduate Mathematics for Economics Exercise Class**, *Department of Statistics*, University of Innsbruck

Publications

In Prep.	**A Simplicity Prior on Semigroups for Learning Theory** Alternatives to the UTM-based Solomonoff prior for finite systems.
2022	**Characterising 0-day exploit brokers**, *Matthias Dellago, Daniel Woods, Andrew Simpson*, Workshop on the Economics of Information Security Results presented at WEIS 2022 and at internal Google Chrome Security Team meeting.
2022	**Exploit brokers and offensive cyber operations**, *Matthias Dellago, Daniel Woods, Andrew Simpson*, The Cyber Defense Review
2022	**Formalising attack trees to support economic analysis**, *Andrew Simpson, Matthias Dellago, Daniel Woods*, The Computer Journal

Invited Talks

2023	**Invited Research Presentation**, *Google Chrome Security Team* Presented findings on exploit broker behavior.
2022	**Invited speaker**, *Workshop on the Economics of Information Security*, Tulsa (Paper presented by co-author).

Academic Events

- 2024: 38th Chaos Communication Congress, Hamburg
- 2024: ICML, Vienna
- 2024: Human Aligned AI Summer School, Prague
- 2024: ICLR, Vienna
- 2023: Singular Learning Theory Retreat, Amsterdam
- 2023: Safe and Trustworthy AI Workshop, Imperial College London

Grants

2023-2024 **Long Term Future Fund Grant**, *Effective Altruism Fund*
Technical AI-alignment research: A new hopfield based approach to mechanistic interpretability of attention.

2024 **Erasmus+ Scholarship**, *European Commission*
For Master's thesis research at Amsterdam Machine Learning Lab.

Education

2022 **Exchange Program**, *Vrije Universiteit Amsterdam*
Geometric deep learning and software reverse engineering.

2021–2025* **Computer Science Master's Studies**, *University of Innsbruck*
*Expected graduation June 2025. Focus on information security and machine learning.

2019-2021 **Physics Graduate Studies**, *University of Innsbruck*
Research in quantum computation, leading to interest in machine learning.

2015-2019 **Physics BSc**, *University of Vienna*
Designed and built a cloud chamber to win a competition for muon (cosmic ray) detection.

2007-2015 **Gymnasium**, *Krottenbachstraße*, Vienna
English-German bilingual high school. Graduation with perfect score.

Languages

German	Native
English	Native

Other Activities

Competitive experience in swimming, triathlon, cross-country skiing, and judo. Alpine climbing and mountaineering.