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

Matthias Dellago

Research Focus

My research aims to develop rigorous theoretical foundations for deep learning by bridging statistical mechanics and algorithmic information theory. While neural networks demonstrate remarkable 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 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 **Mathematics for Economics Exercise Instructor**, *Department of Statistics*, University of Innsbruck

Summer 2018 **Research Assistant**, *Institute for High Energy Physics*, Austrian Academy of Sciences
Quality assurance on radiation-hardened, high-precision particle detector chips for CERN's Large Hadron Collider operations

2018 **Introductory Physics Course TA**, University of Vienna

Publications

2022 **First author: "Characterising 0-day exploit brokers"**, *Matthias Dellago, Daniel Woods, Andrew Simpson*, Workshop on the Economics of Information Security
Results presented at WEIS 2022 at internal Google Chrome Security Team meeting

2022 **First author: "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 **Google Chrome Security Team**, *Invited Research Presentation*
Presented findings on exploit broker behavior

2022 **Workshop on the Economics of Information Security**, *Tulsa*
Invited speaker (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

June 2023 **ML Alignment Theory Scholars**, *Stanford Existential Risks Initiative*, Wentworth Track
6-week workshop on AI-risks, problem solving and scientific writing

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

2021- **MSc Computer Science**, *University of Innsbruck*
Focus on information security and machine learning

2019-2021 **Physics Master's 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 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.