Matthias Dellago

✉ dellago.matt@gmail.com
🌐 matthiasdellago.github.io
matthiasdellago
etuVuHoAAAAJ
matthias-dellago-55060327b

Publications

Characterising 0-day exploit brokers, *Matthias Dellago, Daniel Woods, Andrew Simpson*, Workshop on the Economics of Information Security, 2022.
Invited to present findings to Google Chrome Security Team.

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, 2022.

Awards and Recognitions

2023-2024 **Long Term Future Fund Grantee**, *Effective Altruism Fund*,
Technical AI-alignment research: A new approach to mechanistic interpretability of attention, based on modern Hopfield networks and statistical physics.

Experience

Currently **Guest Researcher**, *Institute for Machine Learning*, Johannes Kepler University Linz.
Interpretability of attention – weight decay and sparsity. Loss landscape 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 University: Combining Exploit Market Data with Attack Trees.

2020 **Instructor of Undergraduate Mathematics for Economics Exercise Class**, *Department of Statistics*, University of Innsbruck.

Summer 2018 **Lab Assistant**, *Institute for High Energy Physics*, Austrian Academy of Sciences.
Measured and approved detectors for use at CERN.

2018 **Teaching Assistant for Introductory Physics Course**, University of Vienna.

Education

Currently **Master's Thesis with Erasmus Scholarship**, *Amsterdam Machine Learning Lab*.

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

2022 **Exchange Program**, *Vrije Universiteit Amsterdam*.
Software reverse engineering and machine learning on graphs.

2020-Present **Computer Science Master's Studies**, *University of Innsbruck*.
Focus on machine learning and information security.

2019-2020 **Physics Master's Studies**, *University of Innsbruck*.
Focus on adiabatic quantum computation. Switched to computer science in 2020.

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	Near-Native
French	Maladroit

Other Interests

Competed in swimming, triathlon, cross country skiing and judo. Currently also mountaineering.