Tom Lieberum

Education

08/2020 – 08/2022 **Master Sc. (Artificial Intelligence)**, *University of Amsterdam, The Netherlands*.
current GPA 9.3 (roughly equivalent to A+ in the UK system)

09/2019 – 08/2020 **Self-studying Computer Science and Machine Learning**.
Studying theoretical computer science, deep learning and reinforcement learning to prepare a switch from physics to machine learning.

10/2016 – 09/2019 **Bachelor Sc. (Physics)**, *RWTH Aachen University, Germany*.
GPA 1.3 (roughly equivalent to A- in the UK system)

Master Thesis (ongoing)

Working Title *Effective Synthesis of Non-shared Information in Multi-View Reinforcement Learning*

Supervision MSc. David Kuric (University of Amsterdam)

Grade Ongoing.

Description In this project we are investigating information-theoretic and variational methods to effectively combine different observational inputs in the multi-view RL setting to learn a representation and corresponding transition model without relying on the simplifying assumptions used in the existing literature. The goal is to develop a more general framework to effectively approach the multi-view RL setting.

Bachelor Thesis

Title *Machine Learning for Top Tagging at the LHC*

Supervision Prof. Michael Krämer and Dr. Alexander Mück (RWTH Aachen University)

Grade Thesis: 1.3 ($\sim$ A-); Talk: 1.0 ($\sim$ A+)

Awards

Summer 2021 **3rd place in MineRL BASALT - NeurIPS 2021 Competition**.
3rd place + most creative research in the MineRL BASALT 2021 competition, organized by the Center for Human Compatible AI and AICrowd, accredited under the NeurIPS 2021 competition track. Focus of the challenge was on learning from demonstrations and feedback without an explicit reward signal. In addition to the above, I was awarded a community support award.

2019 **Dean's List**.
Awarded for being in the top 5 % of students in my year, during the academic year 2018-2019.

2018 **Dean's List**.

Highlighted Machine Learning Experience

09/2022 – present **Research Engineer**, *DeepMind*, Alignment Team.
TBD.

01/2022 – 07/2022 **Mechanistic Interpretability Tooling**.
I received a grant from the Long-Term Future Fund to work on the interpretability library Unseal[a] and to pursue other research topics in mechanistic interpretability.

[a] `https://unseal.readthedocs.io/en/latest/index.html`

Summer 2021 **AGI Safety Fundamentals Course**, *Effective Altruism Cambridge*.
Participating in the AGI Safety Fundamentals course offered by Effective Altruism Cambridge. Contents of the course included a broad overview over the AI alignment problem, particular sub-problems and notable research agendas.

Summer 2021 **Project AI**, "Representation learning for model-based RL in Minecraft", Grade: 8.5 (roughly equivalent to A+ in UK system).
University project with the goal of implementing and comparing different model based reinforcement learning methods in the MineRL environment.

01/2021 **Replication Project**, “Fairness without demographics through Adversarially Reweighted Learning”, Co-authors: Erik Jenner, Frederik Paul Nolte, and Nadja Rutsch.
University project with the goal of replicating a paper from the Fairness in AI literature. Submitted to the Machine Learning Reproducibility Challenge.

Technical skills

Python	Intermediate	Pytorch	Intermediate
LaTeX	Intermediate		

Languages

German	Native	English	C2 (IELTS UKVI 8.5)