JOHANNES TREUTLEIN

johannestreutlein.com ◇ johannestreutlein@berkeley.edu

EDUCATION

Ph.D. Computer Science *8/2022–present*
University of California, Berkeley
Advisor: Prof. Stuart Russell

M.Sc. Computer Science *1/2021–present*
University of Toronto
Research-based master's program; supervision by Prof. Roger Grosse and Prof. Jakob Foerster

Visiting Student in Mathematics and Statistics *10/2019–3/2020*
St Catherine's College, University of Oxford

B.Sc. Mathematics *10/2017–12/2020*
Technical University of Berlin
Overall grade: 1.0 (on a scale of 1.0 (best) to 4.0 (worst)); this puts me in the top 2% of graduates[1]

B.Mus. Double bass *10/2011–9/2015*
State University of Music and Performing Arts Stuttgart
Grade: 1.0; double bass studies at a music conservatory; includes courses in musicology and music theory

EXPERIENCE

Stanford Existential Risks Initiative *6/2022–present*
ML Alignment Theory Scholar, Deceptive AI Stream *Berkeley*

- Working with mentor Evan Hubinger on AI Safety via conditioning predictive models

Center for Human-Compatible AI, UC Berkeley *5/2020–8/2020*
Research Internship *Berkeley (remote)*

- Worked on zero-shot coordination in multi-agent reinforcement learning; developed a mathematical theory and a new deep reinforcement learning method
- Supervision by Prof. Jakob Foerster and Michael Dennis

Machine Learning Group, Technical University of Berlin *6/2019–9/2019*
Student Assistant *Berlin*

- Research on explaining the decisions of anomaly detection models using deep Taylor decomposition

Centre for Effective Altruism *8–9/2018*
Summer Research Fellowship *Oxford*

- Research project on the application of evidential decision theory to Bayesian bargaining games
- Received supervision and input by researchers at the University of Oxford

Effective Altruism Foundation (now Center on Long-Term Risk) *9/2016–7/2019*
Research, Outreach *Berlin*

- Research on decision theory and global priorities; advised more than $ 50,000 in grant-making
- Conducted outreach via social media and in-person

Munich Philharmonic Orchestra *3/2015–8/2016*
Orchestra academist *Munich*

- Performed in concerts and TV and radio recordings with the Munich Philharmonic Orchestra

[1] According to the latest available data at `https://www.pruefungen.tu-berlin.de/fileadmin/ref10/Gradings/SoSe19/Bachelor_Mathematik.pdf`

RESEARCH

Publications

- Cem Anil*, Ashwini Pokle*, Kaiqu Liang*, **Johannes Treutlein**, Yuhuai Wu, Shaojie Bai, J. Zico Kolter, and Roger Grosse. "Path Independent Equilibrium Models Can Better Exploit Test-Time Computation". **NeurIPS 2022**.
- Timon Willi*, Alistair Letcher*, **Johannes Treutlein***, Jakob Foerster. "COLA: Consistent Learning with Opponent-Learning Awareness". **ICML 2022**.
- **Johannes Treutlein**, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. "A New Formalism, Method and Open Issues for Zero-Shot Coordination". **ICML 2021**.
- William MacAskill, Aron Vallinder, Caspar Oesterheld, Carl Shulman, and **Johannes Treutlein**. "The Evidentialist's Wager". **The Journal of Philosophy**, Volume 118, Issue 6, June 2021.

Working papers

- Julian Stastny, Maxime Riché, Alexander Lyzhov, **Johannes Treutlein**, Allan Dafoe and Jesse Clifton. "Normative disagreement as a challenge for Cooperative AI". Accepted at the NeurIPS 2021 StratML and Cooperative AI workshops.
- **Johannes Treutlein** and Caspar Oesterheld. "A representation theorem for decision-theoretic uncertainty". Manuscript.
- **Johannes Treutlein**. "Modeling evidential cooperation in large worlds". Manuscript.

Oral presentations

- "Risks from Learned Optimization in Advanced Machine Learning Systems by Hubinger et al.". AI Safety reading group. Vector Institute for AI. 2021.
- "How the Decision Theory of Newcomblike Problems Differs Between Humans and Machines". Decision Theory and the Future of Artificial Intelligence II. LMU Munich. 2018.
- "Global consequentialism for machines". Utility, Progress, and Technology. 15th Conference of the International Society for Utilitarian Studies. Karlsruhe Institute of Technology. 2018.

SERVICE

- Served as a reviewer on the program committee for the NeurIPS Workshop on Machine Learning Safety *2022*
- Served as a mentor for the Cambridge Existential Risks Initiative (CERI) summer research fellowship, supervising a research project on evidential cooperation in large worlds *6/2022-9/2022*
- Advised on a six-figure grant by a large effective altruist philanthropic organization *6/2022*

SCHOLARSHIPS AND PRIZES

Open Philanthropy AI Fellowship and Vitalik Buterin PhD Fellowship *2022-2027*
Fellowships for a Computer Science PhD at UC Berkeley starting in Autumn 2022, covering tuition and fees and a total stipend of $ 55,000 per year

Center on Long-Term Risk Fund *1/2021–8/2022*
A scholarship of CA$ 160,794 covering living expenses and tuition to pursue a MSc in Computer Science at the University of Toronto

Berlin Mathematical Society Bachelor Prize *2021*
Awarded to Mathematics graduates with outstanding academic records from universities in Berlin and Potsdam

St Catherine's College Book Prize *5/2020*
Awarded for good work in a topology tutorial at St Catherine's College, University of Oxford

Open Philanthropy *10/2019–8/2020*
A scholarship of £ 35,428 covering living expenses and tuition for undergraduate studies in mathematics, including two terms as a visiting student at St Catherine's College, University of Oxford

Deutschlandstipendium *4/2014–3/2015*
Scholarship of € 3,600 awarded based on academic excellence to ca. 1% of German university students

* Equal contribution

EXTRACURRICULAR ACTIVITIES

Oxford University Orchestra *10–12/2019*

- Principal double bass of the Oxford University Orchestra, the university's flagship orchestra, and the Oxford University Sinfonietta, a chamber orchestra made up of the university's best instrumentalists

Effective Altruism Munich local group *2016*

- Co-founded a local Effective Altruism chapter
- Organized and advertized talks and meet-ups (with up to 100 attendants)

Jusos Reutlingen *2009*

- Member of the board of the Reutlingen chapter of the German Social Democrats' youth organization
- Participated in local politics and in electoral campaigns