JOHANNES TREUTLEIN

johannestreutlein.com ⋄ mail@johannestreutlein.com

EDUCATION

Ph.D. Computer Science *8/2022–present (on leave since 6/2024)*
University of California, Berkeley
Advisor: Prof. Stuart Russell

M.Sc. Computer Science *1/2021–12/2022*
University of Toronto and Vector Institute
Research-based; advised by Prof. Roger Grosse and Prof. Jakob Foerster; GPA: 4.0

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

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

Truthful AI *3/2026–present*
Research Scientist *San Francisco*

- Working with Owain Evans on generalization and reasoning in LLMs and their implications for alignment.

Anthropic *6/2024–2/2026*
Member of Technical Staff, Alignment *San Francisco*

- Built model organisms of misalignment, including synthetic data generation, SFT/RL training, and evaluations.

Constellation *1/2024–6/2024*
Astra Fellowship, Owain Evans group *Berkeley*

- Worked on a research project on out-of-context reasoning in LLMs under Owain Evans' mentorship.

Stanford Existential Risks Initiative *6/2022–9/2022*
ML Alignment Theory Scholar, Deceptive AI Stream *Berkeley*

- Worked 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.

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

- Research project on evidential cooperation in large worlds, applying dependency equilibria to bargaining games.
- Advised by researchers at the University of Oxford.

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

- Research on decision theory and global priorities.
- 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.

RESEARCH

Johannes Treutlein, Samuel R. Bowman, Trenton Bricken, Alex Cloud, Misha Wagner, Rowan Wang, Evan Hubinger, Fabien Roger, Sam Marks. *Pre-deployment auditing can catch an overt saboteur.* Anthropic Alignment Science Blog, 2026.

Rowan Wang, **Johannes Treutlein**, Fabien Roger, Evan Hubinger, Sam Marks. *Evaluating honesty and lie detection techniques on a diverse suite of dishonest models.* Anthropic Alignment Science Blog, 2025.

Mia Taylor, James Chua, Jan Betley, **Johannes Treutlein**, Owain Evans. *School of Reward Hacks: Hacking harmless tasks generalizes to misaligned behavior in LLMs.* arXiv preprint, 2025.

Trenton Bricken, Rowan Wang, Sam Bowman, Euan Ong, **Johannes Treutlein**, Jeff Wu, Evan Hubinger, Samuel Marks. *Building and evaluating alignment auditing agents.* Anthropic Alignment Science Blog, 2025.

Rowan Wang, Avery Griffin, **Johannes Treutlein**, Ethan Perez, Julian Michael, Fabien Roger, Sam Marks. *Modifying LLM Beliefs with Synthetic Document Finetuning.* Anthropic Alignment Science Blog, 2025.

Samuel Marks†, **Johannes Treutlein**†, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R. Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meek, Kei Nishimura-Gasparian, Euan Ong, Christopher Olah, Adam Pearce, Fabien Roger, Jeanne Salle, Andy Shih, Meg Tong, Drake Thomas, Kelley Rivoire, Adam Jermyn, Monte MacDiarmid, Tom Henighan, Evan Hubinger†. *Auditing language models for hidden objectives.* arXiv preprint, 2025.

Nathan Hu, Benjamin Wright, Carson Denison, Samuel Marks, **Johannes Treutlein**, Jonathan Uesato, Evan Hubinger. *Training on Documents About Reward Hacking Induces Reward Hacking.* Anthropic Alignment Science Blog, 2025.

Ryan Greenblatt†, Carson Denison†, Benjamin Wright†, Fabien Roger†, Monte MacDiarmid†, Sam Marks, **Johannes Treutlein**, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Evan Hubinger†. *Alignment faking in large language models.* arXiv preprint, 2024.

Johannes Treutlein*, Dami Choi*, Jan Betley, Samuel Marks, Cem Anil, Roger Grosse, Owain Evans. *Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data.* **NeurIPS 2024**.

Caspar Oesterheld, **Johannes Treutlein**, Roger Grosse, Vincent Conitzer, Jakob Foerster. *Similarity-based cooperative equilibrium.* **NeurIPS 2023**.

Johannes Treutlein. *Modeling evidential cooperation in large worlds.* arXiv preprint, 2023.

Caspar Oesterheld*, **Johannes Treutlein***, Emery Cooper, and Rubi Hudson. *Incentivizing honest performative predictions with proper scoring rules.* **UAI 2023**.

Evan Hubinger, Adam Jermyn, **Johannes Treutlein**, Rubi Hudson, Kate Woolverton. *Conditioning Predictive Models: Risks and Strategies.* arXiv preprint, 2023.

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**.

Julian Stastny, Maxime Riché, Alexander Lyzhov, **Johannes Treutlein**, Allan Dafoe and Jesse Clifton. *Normative disagreement as a challenge for Cooperative AI.* NeurIPS 2021 StratML and Cooperative AI workshops.

William MacAskill, Aron Vallinder, Caspar Oesterheld, Carl Shulman, and **Johannes Treutlein**. *The Evidentialist's Wager.* **The Journal of Philosophy**, Volume 118, Issue 6, **2021**.

† Core research contributor
* Equal contribution

Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. *A New Formalism, Method and Open Issues for Zero-Shot Coordination.* **ICML 2021**.

Johannes Treutlein and Caspar Oesterheld. "A typology of Newcomblike problems". Manuscript, 2017.

SERVICE

- Reviewer for ICLR 2025 *2024*
- Reviewer for NeurIPS 2024 *2024*
- External grant reviewer for the Cooperative AI Foundation *2024*
- Reviewer for the NeurIPS Workshop on Machine Learning Safety *2022*
- Advised on a six-figure grant by a large effective altruist philanthropic organization *2022*

SCHOLARSHIPS AND PRIZES

Open Philanthropy AI Fellowship and Vitalik Buterin PhD Fellowship *8/2022-7/2027*
Fellowships for a Computer Science PhD at UC Berkeley, covering tuition and fees in addition to a stipend of $ 55,000 per year.

Center on Long-Term Risk Fund *1/2021–7/2022*
A scholarship of $ 144,579.10 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.

EXTRACURRICULAR ACTIVITIES

UC Berkeley Symphony Orchestra *9/2023-12/2023*

- Principal double bass of the UC Berkeley Symphony Orchestra, performing in six concerts with three different programs during the Fall 2023 semester.

Oxford University Orchestra *10/2019-03/2020*

- 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.
- Contributed to local politics and electoral campaigns.