Joschka Braun

joschkacbraun@gmail.com · (+49)175 8901489 · Website · GitHub · LinkedIn

Work Experience

MATS Research — Berkeley, USA / London, UK / Berlin, Germany
Research Scholar — Jun 2025 – Present
- Researched **exploration hacking** – how LLMs can strategically influence and resist RL training by altering their exploration – supervised by David Lindner, Scott Emmons, and Roland Zimmermann (Google DeepMind).
- Created model organisms that resist RL capability elicitation on biosecurity and AI R&D benchmarks while retaining general performance; evaluated detection methods and audited frontier models for exploration hacking propensity.
- Paper submitted to ICML 2026; earlier version received **Best Paper Runner-up at NeurIPS 2025 Workshop**.

Health NLP Lab — Tübingen, Germany
Part-time, paid Student Researcher — Feb 2025 – May 2025
- Research in activation engineering for text summarization; results accepted to **Findings of EACL 2026**.

KASL - Krueger AI Safety Lab — Cambridge, UK
Full-time, paid Research Internship — Jun 2024 – Dec 2024
- Researching unreliability and limitations of steering vectors; results presented at **ICLR 2025 Workshop**.

Health NLP Lab — Tübingen, Germany
Part-time, Student Researcher — Oct 2023 – Jun 2024
- Research in controlled text generation and topical summarization; results presented at **ICML 2025 Workshop**

RAWLAB GmbH — Reutlingen, Germany
Student Employee, Machine Learning Engineer — May 2021 – Jan 2023
- Trained ML models and applied transfer learning techniques on MobileNetV2 to classify time series spectrograms.

Publications

Exploration Hacking: LLMs Can Learn to Resist RL Training
Joschka Braun*, Eyon Jang*, Damon Falck*, ..., Scott Emmons, Roland S. Zimmermann, David Lindner
Under Review at ICML 2026 — **(Preprint)**

Beyond Multiple Choice: Evaluating Steering Vectors for Summarization
Joschka Braun, Carsten Eickhoff, Seyed Ali Bahrainian
Findings of the Association for Computational Linguistics: EACL 2026 — **(To Appear)**

Resisting RL Elicitation of Biosecurity Capabilities: Reasoning Models Exploration Hacking on WMDP
Joschka Braun*, Yeonwoo Jang*, Damon Falck*, Roland S. Zimmermann, David Lindner, Scott Emmons
NeurIPS 2025 Workshop: Biosecurity Safeguards for Generative AI — **Best Paper Runner-up (Oral)**

Beyond Multiple Choice: Evaluating Steering Vectors for Adaptive Free-Form Summarization
Joschka Braun, Carsten Eickhoff, Seyed Ali Bahrainian
ICML 2025 Workshop on Actionable Interpretability

Understanding (Un)Reliability of Steering Vectors in Language Models
Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, Dmitrii Krasheninnikov
ICLR 2025 Workshop on Foundation Models in the Wild

Education

University of Tübingen — Tübingen, Germany
Master of Science, Machine Learning — Oct 2022 – Jun 2025
- *Final Grade:* 1.2 (German scale: 1.0 best; 4.0 pass)
- *Thesis:* Understanding Unreliability of Steering Vectors in Language Models (1.0 Grade)

University of Tübingen — Tübingen, Germany
Bachelor of Science, Computer Science — Apr 2020 – Sep 2022
- *Final Grade:* 1.3 (Top 5% of class; German scale: 1.0 best; 4.0 pass)
- *Thesis:* Verbal Epistemic Uncertainty Estimation for Numeric Values with GPT-3 (1.0 Grade)

References

David Lindner: Mentor at MATS 8.0; Research Scientist at Google DeepMind
Roland Zimmermann: Mentor at MATS 8.0; Research Scientist at Google DeepMind
Scott Emmons: Supervisor of research project at MATS 8.0; Research Scientist at Google DeepMind
Dmitrii Krasheninnikov: Supervisor of research project at KASL; PhD student at University of Cambridge
David Krueger: Senior Supervisor of research project at KASL; Assistant Professor at University of Cambridge
Seyed Ali Bahrainian: Supervisor of research project at Health NLP Lab; Post-Doc at Health NLP Lab
Marius Hobbhahn: Supervisor of Bachelor's Thesis; CEO of Apollo Research
Contact details available on their websites or upon request.

AI Safety Experience

METR Remote
ML Research Engineering Contractor Jan – Feb, 2025
– Provided task feedback and set human performance baselines for ML Engineering tasks for METR's HCAST benchmark.

ASET - AI Safety Engineering Taskforce Remote
ML Research Engineer Sep 23rd – Nov 1st, 2024
– I've integrated the BIG-Bench Hard benchmark into the Inspect Evals framework of the UK AI Safety Institute.

Cambridge ERA:AI Fellowship Cambridge, UK
Technical AI Safety Fellow Jul 1st – Aug 23rd, 2024
– Conducted research on the limitations of contrastive activation steering in language models.

Center for AI Safety Remote
Facilitator of 8-week Intro to ML Safety course Mar – Apr 2023
– Taught adversarial robustness, data poisoning, anomaly detection, and mitigations like adversarial training.

BlueDot Impact Remote
Facilitator of 12-week AI Safety Fundamentals Alignment Track Feb – Apr 2023
– Facilitated sessions on safety risks, RLHF, scalable oversight and mechanistic interpretability.

Academic Service

- **NeurIPS 2025**: Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET); Mechanistic Interpretability Workshop
- **ICLR 2025**: Workshop on Foundation Models in the Wild; Workshop on Building Trust in LLMs and LLM Applications: From Guardrails to Explainability to Regulation
- **NeurIPS 2024**: Workshop on Socially Responsible Language Modelling Research (SoLaR)

Skills

Programming: Python, PyTorch, TensorFlow, Transformers (Huggingface), NumPy, Pandas, Scikit-Learn, Keras
Relevant University Courses: Deep Learning, Reinforcement Learning, Natural Language Processing, Statistical Machine Learning, Probabilistic Machine Learning, Self-Driving Cars, Linear Algebra
Linguistic Proficiency: German (Native), English (Proficient)

Miscellaneous

Debating: Two-time winner of the German Debating Championship (DDM) 2019 & 2020, won the largest and most prestigious German-speaking debate competition for two consecutive years.
Scholarship: Bertha von Suttner-Studienwerk (2022–2024), awarded to 15 students out of 170 applicants, based on their academic achievements, social engagement, and commitment to humanist values.
AI Safety Group Co-founder: In October 2023, co-founded the AI Safety Group Tübingen to engage more people in AI safety.