JAN WEHNER

AI Safety PhD Student

jan.wehner@cispa.de
Saarbrücken, Germany
jan-wehner-811682216
janweh

TECH STACK

Languages: Python, Matlab, LaTex, Java

Frameworks: Pytorch, Transformers, Matplotlib

SKILLS

Academic Writing
ML Engineering
Communicating Ideas
Teaching
Critical Thinking
Data Visualization

Team Management
Leadership
Public Speaking

LANGUAGES

German: Native

English: C2

SCHOLARSHIPS

European Laboratory for Intelligent Systems (ELLIS) PhD program

A selective PhD meta-program supporting co-supervision and research visits throughout Europe

Open Philanthropy Career Development and Transition Funding scholarship program

Funding to conduct independent research and upskilling in AI Safety

ABOUT ME

I am a PhD student working on AI Safety focusing on interpretability and alignment of LLMs. I am driven by my curiosity and commitment to solving pressing societal problems using my skills in Machine Learning. I love understanding how things work, developing new ideas and working on hard technical problems.

EDUCATION

PhD Candidate | CISPA Helmholtz Center for Information Security

Jun. 2024 – Present — Saarbrücken, Germany

- Researching LLM Alignment and Interpretability advised by Prof. Mario Fritz, co-advised by Prof. David Krueger

MSc. Computer Science - AI Technology Track | University of Technology Delft

Sep. 2021 – Nov. 2023 — GPA: 8.5/10

- Relevant courses: Deep Reinforcement Learning, Computer Vision by Deep Learning, Machine Learning 1&2, Artificial Intelligence Techniques

BSc. Business Informatics | Otto-Friedrich-University Bamberg

Apr. 2017 - May 2021 — GPA: 1.2 (cum laude)

- Thesis: Efficient inference of qualitative temporal information for robust planning

RESEARCH EXPERIENCE

Participant | AI Safety Camp

Jan. 2024 - Mai 2024 — Remote

- Conducted and published research to characterize and prevent harmful fine-tuning attacks

Master Thesis | University of Technology Delft

Feb. 2023 - Nov. 2023 — Delft, Netherlands

- Developed and executed novel research ideas on the intersection of AI Alignment and XAI
- Topic: Counterfactual Explanations of Learned Reward Functions

Research Fellow | Swiss Existential Risk Initative

Jul. 2022 - Aug. 2022 — Bern, Switzerland

- Employed interdisciplinary thinking to identify data requirements for Inverse Reinforcement Learning to learn human values
- Experimentally validated theoretical shortcomings and technical challenges for Reward Learning

REFERENCES

Prof. Mario Fritz
CISPA Helmholtz Center for Information Security
fritz@cispa.de

Prof. Luciano Siebert
University of Technology Delft
L.CavalcanteSiebert@tudelft.nl

Research Assistant | Otto-Friedrich-University Bamberg

Jun. 2021 - Aug. 2021 — Bamberg, Germany

- Conducted empirical research on knowledge representation in temporal domains
- Published and presented research in journals and conferences

TEACHING AND ORGANISING EXPERIENCE

Founder | Delft AI Safety Initiative

Nov. 2022 - Dec. 2023 — Delft, Netherlands

- Devised strategy and structure for a new student organisation focused on addressing risks from AI
- Delivered courses and regular events on AI Alignment to >50 students
- Presented introductory talks about research in AI Safety

Founder | Effective Altruism Delft

Jan. 2022 - Jul. 2023 — Delft, Netherlands

- Initiated, developed and led a student organisation supporting students in achieving positive societal impact
- Developed and delivered courses on Effective Altruism to >100 students
- Assisted students in career and donation choices through advising, discussions and organising events

Teaching Assistant - Algorithms and Data Structures | Otto-Friedrich-University Bamberg

Apr. 2019&2021 - Aug. 2019&2021 — Bamberg, Germany

- Held weekly classes and discussions explaining key concepts in Computer Science
- Graded assignments and providing feedback

RESEARCH PROJECTS

Wehner, J., Abdelnabi, S., Tan, D., Krueger, D., Fritz, M. *Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models.* arXiv preprint.

Seth, I., **Wehner, J.**, Abdelnabi, S., Binkyte, R., Fritz, M. *Safety is Essential for Responsible Open-Ended Systems* arXiv preprint

Rosati, D., **Wehner, J.**, Williams, K., Bartoszcze, Ł., Atanasov, D., Gonzales, R., ... & Rudzicz, F. (2024). Representation noising effectively prevents harmful fine-tuning on LLMs.

Wehner, J., Oliehoek, F., & Siebert, L. C. (2024). *Explaining Learned Reward Functions with Counterfactual Trajectories*. AIEB Workshop ECAI 2024.

Wehner, J., Sioutis, M. & Wolter, D. *On robust vs fast solving of qualitative constraints*. J Heuristics 29, 461–485 (2023).