University of Stuttgart
Germany

The **AI Safety Research Group** headed by **Dr. Thilo Hagendorff** invites applications for

2 PhD Positions in AI Safety

(TV-L 13, 100%)

Join the Independent Research Group and work on AI Safety research, evaluate capabilities in large reasoning models, and discover novel machine behavior – especially behavior related to deception. As a PhD candidate, you will have the opportunity to conduct high-impact research and work on your dissertation while contributing to the development of safe AI. The project is located at the University of Stuttgart, Germany. For a brief insight into our previous work, visit Dr. Thilo Hagendorff's website at www.thilo-hagendorff.info.

Details

We are dedicated to innovative ideas and to conducting high-impact research. Our goal is to understand novel behavior in frontier models and ensure that developers avoid safety and security vulnerabilities. You should be motivated to empirically investigate LRMs/LLMs and address research questions related especially to deception in AI models, particularly in agentic settings. In other words, if you're the kind of person who reads AI papers or listens to AI podcasts late at night just because you can't help yourself – this position is for you. As AI technologies evolve rapidly, it is essential for us to stay at the forefront of the latest advancements and techniques. Therefore, we seek a researcher who is open to new ideas, eager to learn, and ready to explore different approaches to addressing safety challenges in AI. As a researcher in our group, you will be able to work on cutting-edge projects in a fast-growing area of AI research, publish your findings in leading academic journals, and present your work at conferences.

Eligibility criteria

We invite **applications from all fields**, but a background in computer science is an advantage.

- PhD candidates hold (or expect to complete soon) a Master's or equivalent degree
- Familiarity with LRMs and LLMs
- An interest in empirically evaluating these technologies
- Willingness to learn new skills and knowledge, be it related to model evals, interpretability, or monitoring
- Strong analytical and programming skills, being proficient in Python and statistics
- Advanced English language skills (knowledge of German is not required)

What we offer

The PhD position offers high-potential researchers an exceptional opportunity to conduct high-impact research and work on their PhD. Additionally, the position benefits from support structures offered by ELLIS Stuttgart, the Graduate Academy of the University of Stuttgart (GRADUS), and the Research Focus IRIS.

- The position is **TV-L 13** according to the German federal wage agreement.
- The contract will be for up to **2 years**.

- The envisaged start date is **01.01.2026** but can be earlier if it suits the availability of the successful candidate.
- This is not a cog-in-the-machine PhD. It's a chance to help shaping a new research field.
- We don't count hours. We count ideas that excite us.

How to apply

Please submit your application in a **single PDF file** to thilo.hagendorff@iris.uni-stuttgart.de, including the following:

- **Cover letter** describing your background and research interests
- **CV**, containing a list of publications (if any).
- A **document** (1) describing your most important achievements (this can, but must not only include achievements that are valued by society or your peers, but also achievements that you value personally, like caring for a specific cause, learning from a mistake, etc.), (2) describing what are your favourite papers and why, and (3) describing what you think are the main limitations of current AI models.
- **Certificates** (Bachelor, Master). Please include a certified translation of the diploma for languages other than German or English.
- **Reference letters** (if available).

Please submit your application by **15.11.2025**. Inquiries should be directed to Dr. Thilo Hagendorff at thilo.hagendorff@iris.uni-stuttgart.de.

The University of Stuttgart would like to increase the proportion of women in the scientific field and is therefore particularly interested in **applications from women**. Severely disabled persons are given priority in the case of equal suitability.

We look forward to your application!