Existential Risk Observatory

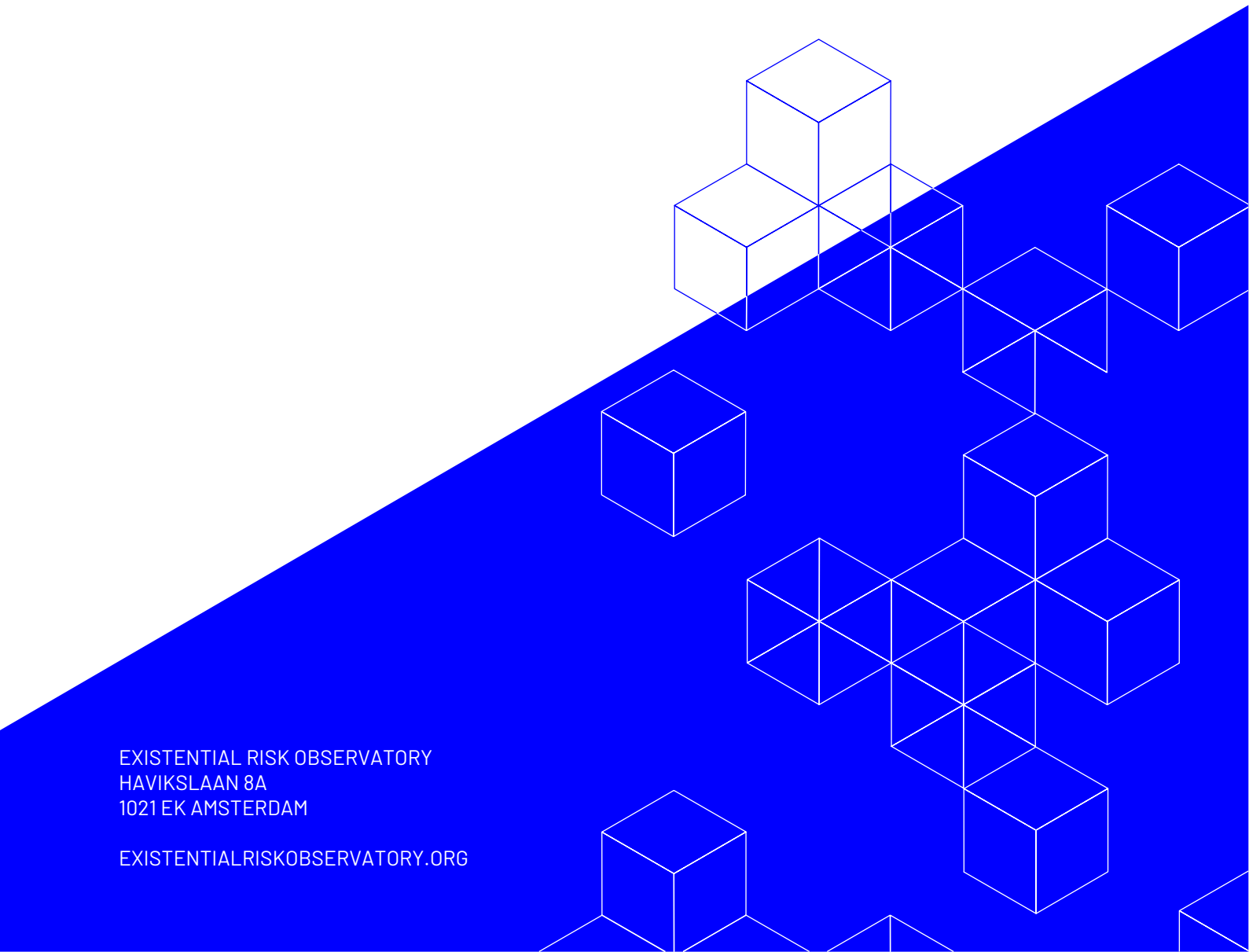
HOW TO PAUSE AI

Research internship opening

FULLTIME / PART-TIME, REMOTE / IN AMSTERDAM, THE NETHERLANDS

3 MONTHS (FULLTIME BASIS)

1000 E/MONTH (FULLTIME BASIS)



EXISTENTIAL RISK OBSERVATORY
HAVIKSLAAN 8A
1021 EK AMSTERDAM

EXISTENTIALRISKOBSERVATORY.ORG

WE REDUCE EXISTENTIAL RISKS OF AI

New technology such as AI, biotech and nanotech is developing at lightning speed, which offers plenty of opportunities. However, this technology also entails major risks. In the worst case, the uncontrolled development of human-level AI in particular could even become an existential risk: a threat to the future of humanity.

The Existential Risk Observatory has been set up to reduce the chance of existential risks. We are a small nonprofit with the aim of reducing existential risks by informing the public debate. We do this mainly by publishing articles in leading media, organizing events, doing research, and advising policymakers. We are funded by, among others, SFF, ICFG, and EA LTFF.

ROLE DESCRIPTION

You will independently perform research aiming to answer the question: how can we pause AI effectively for as long as it is needed? To do so, you will read and process literature, write a report with potentially novel ideas, have meetings with your supervisor, give presentations for interested experts and/or laymen, and contact other researchers outside the organization. The work can be done either fully remote or (partially) in-person in our Amsterdam office.

The final deliverable is a report containing all information available leading up to a robust, lengthy implementation of the pause with as little downsides as possible. As a researcher, you will work mostly independently, without much support from colleagues.

ABOUT THE PAUSE

Without a clear solution to the giant problem of aligning (both technical and social) superhuman AI, we think this technology should not be built. Therefore, we and many others have called for a pause in the development of AI beyond a certain capabilities level. In the short term, such a pause could be implemented fairly straightforwardly (given public awareness and political backing), since leading AI models are at this moment large and expensive to train, and few companies are on the cutting edge. However, eventually, and perhaps soon already, hardware and algorithmic improvements could lead to uncontrollable AI being available to everyone. How should we enforce a pause in such a scenario? That's the question you will help to answer. Solution directions you will look into will include hardware regulation, data regulation, and others.

WHAT WE OFFER

- Working on the AI Pause means working on one of the most impactful developments of the century. Becoming an expert in this field and building a network could be a great step in your career.
- You work mostly independently, with a lot of freedom to do your own work in the way you want and in the hours you prefer.
- Internship compensation of 1000 euros/month based on full-time employment (40 hours/week).
- Part-time work (at least 16, preferably 24 hrs/week) is possible and compensation will be proportional.
- Working fully remote is possible, although occasional in-person collaboration in our Amsterdam office would be ideal.
- The internship duration is around 3 months for fulltime work. Starting date is flexible.

REQUIREMENTS

- We're looking for someone who is deeply motivated to pause AI and already has a good understanding of why this needs to be done and perhaps ideas on how to do it.
- STEM background.
- Knowledge of hardware required to train LLMs is a significant plus. If you don't have this, demonstrate in your cover letter how you might acquire it. This knowledge is important since hardware regulation is seen as one of the more promising paths to an enforceable pause.
- Existing network in the existential risk world would be a significant plus (but adherence to any particular group such as Effective Altruism of Rationalism is neither an advantage nor a disadvantage).
- Academic educational level, preferably you've finished your Bachelor.
- Strong analytical skills, demonstrated by academic credentials.
- Good understanding of AI existential risk, especially loss of control. Ideally you know about the different existential risk threat models by Yudkowsky/Bostrom, Paul Christiano, and Bengio/Hinton, proposed solutions, and where they could break down.
- You can write a decent report in English.
- First principles thinker: showing where you have used your amazingly good world model to come up with an original result would be a plus.
- Self-starter: you don't mind to work without much supervision and without close collaboration with colleagues. Ideally, you have shown that you can work well independently and get things done.

APPLICATION PROCEDURE

Please send your motivation letter and CV to info@existentialriskobservatory.org. We review applications on a rolling basis.