Existential Risk Observatory

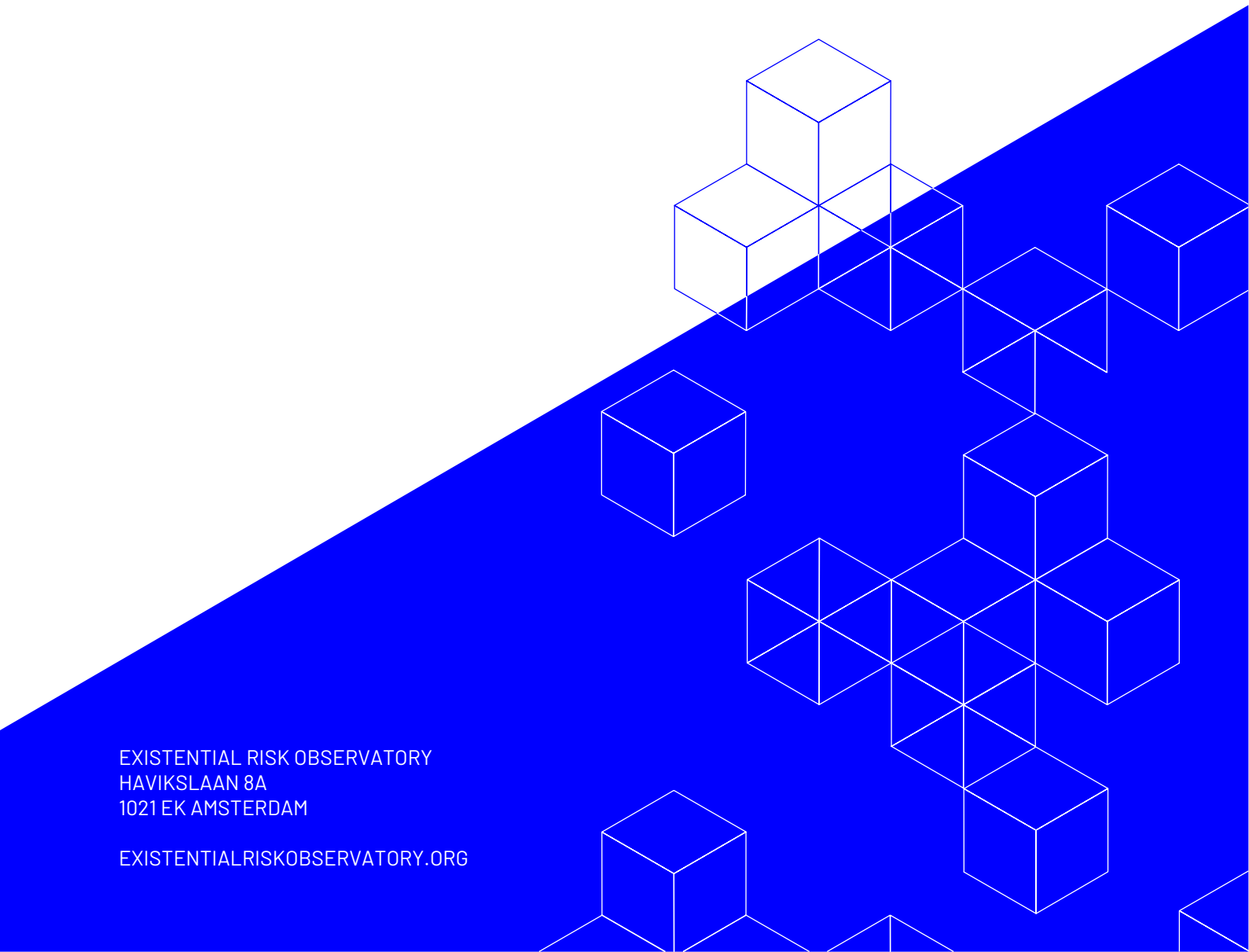
CONDITIONAL AI SAFETY TREATY

Research internship opening - international policy

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

AI is developing rapidly, which offers plenty of opportunities, but also entails major risks. In the worst case, the uncontrolled development of human-level AI could become an existential risk: a threat to the future of humanity.

The Existential Risk Observatory is 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, The Dreamery Foundation, and EA LTFF.

ROLE DESCRIPTION

You will help draft an international policy paper which you will co-author. The paper will propose the Conditional AI Safety Treaty, an international treaty proposing to pause international AI development if and when global AI capabilities reach a sufficiently unsafe level. The paper will describe treaty rationale, preferred outline, and advantages and disadvantages of different options, but will not contain a legal draft text.

Your work is crucial for the organization, since the Conditional AI Safety Treaty will be a focal point of our work the coming year. It is also highly likely that AI treaties, especially existential risk focused ones, will play a very important role in international politics the coming years. Therefore, this internship is a great launch for your international policy career at the forefront of AI regulation.

You will mostly work independently, but will have regular meetings with your direct supervisor, the director. Additionally, you are expected to reach out, supported where needed, to other experts in the field and potential co-authors. You are also expected to work together with others in the field (outside the organization) independently where relevant. Work may be carried out at the Amsterdam office, or remote from anywhere.

ABOUT THE CONDITIONAL AI SAFETY TREATY

AI has gotten much better in recent years, and we think progress will continue. That means it will be a matter of time until we reach superhuman AI, with its associated existential risks. Without a clear solution to the giant problem of aligning (both technical and social) superhuman AI, which we don't expect to be found any time soon, we think this technology should not be built (precautionary principle). Therefore, we and many others have called for a pause in the development of AI. We are now combining this call with a novel proposal by many leading researchers, namely to make policy capability-dependent.

A pause faces two main issues:

1. If one country pauses, another countries may not (coordination issue).
2. Some say it is too soon to pause already, since AI, according to them, is not yet close to superhuman level (timing issue).

The Conditional AI Safety Treaty aims to resolve both these issues. Coordination issues are alleviated by the very concept of an international treaty, while timing issues are resolved by making the proposal dependent on AI capability progress, and lack of convincing alignment. At the same time, preparing the treaty now, makes sure we have already passed lengthy treaty negotiations by the time AI capabilities arrive at existential levels.

These are the main outlines of our Conditional AI Safety Treaty proposal. It will be up to you to develop this to a full policy proposal, together with co-authors. At the end of the internship, our policy paper should be ready, we should have thought about every aspect in detail and we should feel secure that there does not exist a better AI treaty proposal than ours.

WHAT WE OFFER

- Working on geopolitical AI coordination means working on one of the most impactful aspects of the century, and the ones after.
- Becoming an expert in this field, publishing a paper, and building a network could be an amazing launch of your career.
- You work mostly independently, with a lot of freedom to do your own work 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 regular in-person presense in our Amsterdam office would be ideal.
- The internship duration is around 3 months for fulltime work, preferably longer for part-time. The starting date is flexible.

REQUIREMENTS

For crucial work, we are looking for an excellent candidate:

- Excellent academic results, preferably at a top university. Most suitable for Master or PhD-level.
- An international policy, politics, or international law background would be suitable (to have technical coursework and understanding in addition to this background would be greatly appreciated).

- An AI background would also be suitable (in this case, it would be important to demonstrate that you also possess generalist skills and understanding).
- Other academic backgrounds will be considered.
- Work experience and/or an existing network in one of the following would be highly valued:
    - The existential risk space.
    - International organizations (e.g. UN, OECD, RAND, etc.).
    - AI academia or leading industry.
    - Governments.
    - Other academia.
- Excellent communication skills and professional presentation, both orally and in writing.

APPLICATION PROCEDURE

To apply, please send your motivation letter and CV to info@existentialriskobservatory.org before 30 September 2024. We look forward to your application and exploring how you can contribute to our mission.