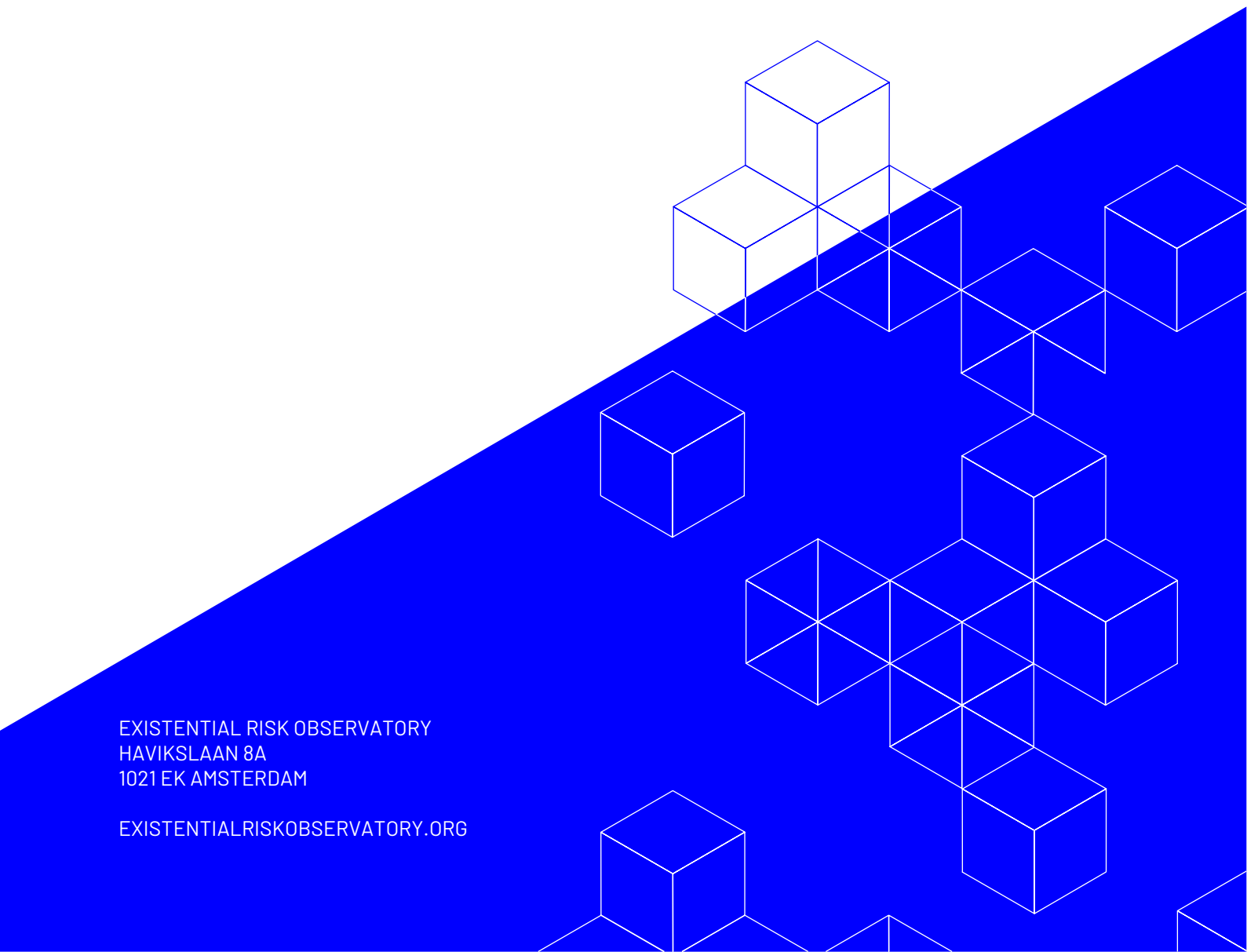
Existential Risk Observatory

AI Policy Proposals

November 2023



EXISTENTIAL RISK OBSERVATORY
HAVIKSLAAN 8A
1021 EK AMSTERDAM

EXISTENTIALRISKOBSERVATORY.ORG

Contents

SAFETY

1. Implement an AI pause

Human extinction risks can materialise already when training a frontier AI model, not only when deploying it. Currently, there is no known method to rule out an extinction event when training any model more capable than existing frontier models (such as GPT-4). As long as this is the case, we should not gamble with humanity by training new models with unknown capabilities that may turn out to be catastrophic. Therefore, we need to pause advancing AI capabilities by capping the allowed size of training runs to GPT-4 level (around 2.2e25 FLOP). This limit will need to decrease in the future to compensate for algorithmic improvements.

This measure should first be implemented by governments of countries with leading labs (specifically the US and UK), and later by all countries. In the medium term (beyond a few years), it is unfortunately not yet clear how to keep a pause in place. This question should therefore be researched by academics, nonprofits, and policymakers with priority.

2. Create a licensing regime

For any frontier model not subject to the Pause, governments should demand a training licence. This licence can only be obtained by meeting safety requirements, such as, but not necessarily limited to, the ones in the areas of model evaluations and auditing described below. Without a licence, no frontier models may be trained.

3. Mandate model evaluations

For any frontier model not subject to the Pause, model evaluations should be mandatory in order to obtain the licence described above. These model evaluations (evals) must be performed both at several points during training and after training, to test for dangerous behaviour.

Examples of such behaviour include the ability for human deception, hacking, and providing information that could lead to harmful application, such as the manufacture of bioweapons. This is not an exhaustive list of evaluations and it should be updated periodically to include capabilities that could lead to catastrophic outcomes.

These evaluations should, when the infrastructure is complete, be run by governments or public institutes, rather than for-profit companies.

4. Mandate third-party auditing

For any frontier model not subject to the Pause, third-party auditing should be mandatory in order to obtain the licence described above. Auditing should be performed in three layers: governance audits (of technology providers that design and disseminate LLMs), model audits (of LLMs after pre-training but prior to their release), and application audits (of applications based on LLMs).

5. Track frontier AI hardware

Track all hardware that can be used to train frontier AI models. At the moment, this concerns mostly high-bandwidth GPUs, as opposed to consumer-grade hardware. Governments should keep track of the location of such hardware and verify that no unlicensed AI models are being trained at locations with high concentrations of frontier AI hardware.

6. Prohibit frontier capabilities research

Research breakthroughs could arguably lead to several orders of magnitude less compute being required for the same capabilities level. This could drastically increase extinction risk. Therefore, all AI research that seems likely to increase frontier AI capabilities, as judged by committees of experts, should be prohibited. This should include the publication of capabilities research, such as novel training algorithms, agent runtimes and AI architectures.

7. Publicly fund AI Safety research (but do not purchase hardware)

There are many open questions in the field of AI Safety that may benefit enormously from more research. For example, it is deeply problematic that there is no academic consensus yet on whether AI presents a risk of human extinction, and if so, through which mechanism exactly. Also, there is no known method yet for implementing an AI Pause beyond a few years duration. Additional research could, among many other outputs, yield consensus threat model descriptions and regulation measures to tackle these risks that would work long-term.

Additionally, funding AI Safety research publicly could avoid potential conflicts of interest between research funders and the public, increasing alignment between research outcomes and the public interest.

Public funding for AI Safety should however not be spent on hardware (compute). In the past, companies such as OpenAI, Google Deepmind, Anthropic, and others, started out as labs that were at least partially focused on safety. Currently, however, many see these companies as mostly accelerating AI timelines. We think this mistake should be avoided with governments. This does however mean that technical alignment, insofar this approach is helpful in reducing extinction risk, should be left to AI labs, which are best positioned for this task.

DEMOCRACY AND OPENNESS

8. Recognize AI extinction risk and communicate this to the public

It is crucial to establish a ground truth of AI risks and to do so explicitly and publicly. 262 AI scientists agree that human-level AI poses an extinction risk to humanity. UN Secretary-General Antonio Guterres, staff from the White House, the EU Commission, and the UK's Rishi Sunak have all acknowledged this. Other government leaders should follow suit and officially acknowledge the human extinction risk that comes with developing frontier AI.

Once governments have acknowledged AI extinction risk, they should communicate this risk to their citizens. Also, parliamentary hearings should be held with AI-developing company CEOs and leading AI academics present. Finally, private education and informative sessions on AI x-risk should be held for politicians and other policymakers.

9. Make the AI Safety Summit the start of a democratic and inclusive process

Create recurring global summits every six months and appoint a permanent commission that implements decisions taken at previous summits and prepares for future summits. Expand the number of countries and (civil society) stakeholders involved in regulating frontier AI, where care should be taken to ensure gender balance, include the Global South, and include underrepresented communities.

10. Organise AGI and superintelligence referendums

It is not a given that superintelligence, or an AI that is significantly more capable than humans at a wide range of cognitive tasks, should be built or that the public wants these to be built. Given the human extinction risk due to the large technical problems with keeping this technology under control, but also issues such as mass unemployment, runaway inequality, and concentration of power, among many others, we should be open to the possibility to stop AGI and/or superintelligence from being developed. Therefore, referendums should be held to decide democratically on these issues.

11. Make AI labs' control and alignment plans public

Currently, ordinary citizens lack insight into what would concretely happen if AI labs would deliver on their ambition to build AGI, and after that superintelligence. This is despite the decisions being made at those labs having a possibly enormous impact on these citizens' lives. Technical decisions on how to build, align, or control a superintelligence could radically alter social outcomes. Therefore, the public needs to have insight into leading AI labs' plans and decisions regarding control and alignment.

12. Demand a reversibility guarantee for increases in frontier AI capabilities

Humanity does not have much experience with irreversible technology. For all technology that has been built so far, we could choose where, when, and how to apply it. We could also choose where, when, and how not to apply it any longer, and we are regularly doing so. Superintelligence, however, may not be controllable. This means that, even in an aligned scenario, negative side effects that can only be observed after occurrence, cannot be corrected anymore. These side effects may be very large. One could think, for example, of climate change as a side effect of the steam and combustion engines. To allow us to correct unforeseen consequences of further AI development, we propose to mandate that increases in frontier AI capabilities must be guaranteed to be reversible.

GOVERNANCE

13. Establish an International AI Agency

AI extinction risk will not respect borders. Therefore, meaningful regulation and enforcement will need to be global. An International AI Agency should be set up to create and enforce AI regulation effectively.

14. Establish liability for AI-caused harm

With anyone creating AI models fully liable for damage done using them, companies will become significantly more safety-conscious. It is important that companies training the original models should not escape liability for events occurring downstream. Those companies are the ones who might be able to take actions to increase models' safety, and the ones deciding to train the models in the first place. Therefore, it is appropriate that liability is located where the risk is created: at the companies training the AI.

15. Do not allow training LLMs on copyrighted content

A significant part of current frontier models' capability is created by training on copyrighted materials without consent, credit or compensation to authors. We think this should not be tolerated, both to protect authors' rights and to increase safety by slowing down capability advancement.