A Narrow Path

How to secure our future

Andrea Miotti, Tolga Bilge, Dave Kasten, James Newport

October 2024

Thank you to Anthony Aguirre, Connor Leahy, Max Tegmark, Eva Behrens, Leticia Garcia Martinez, Gabriel Alfour, Adam Shimi, Pedro Serodio, and many others for contributions and vital discussions. Thank you to everyone else that provided feedback on drafts. Thank you to Eleanor Gunapala for graphics and publishing.

Get in touch: hello@narrowpath.co
www.narrowpath.co

Executive Summary

There is a simple truth - humanity's extinction is possible. Recent history has also shown us another truth - we can create artificial intelligence (AI) that can rival humanity.

While most AI development is beneficial, artificial superintelligence threatens humanity with extinction. We have no method to currently control an entity with greater intelligence than us. We currently have no ability to predict the intelligence of advanced AIs prior to developing them, and we have incredibly limited methods to accurately measure their competence after development.

We now stand at a time of peril. Companies across the globe are investing to create artificial superintelligence – that they believe will surpass the collective capabilities of all humans. They *publicly* state that it is not a matter of "if" such superintelligence might exist, but "when".

We do not know how to control AI vastly more powerful than us. Should attempts to build superintelligence succeed, this would risk our extinction as a species. But humanity can choose a different future: there is a narrow path through.

A new and ambitious future lies beyond a narrow path. A future driven by human advancement and technological progress. One where humanity fulfills the dreams and aspirations of our ancestors to end disease and extreme poverty, achieves virtually limitless energy, lives longer and healthier lives, and travels the cosmos. That future requires us to be in *control* of that which we create, including AI.

We are currently on an unmanaged and uncontrolled path towards the creation of AI that threatens the extinction of humanity. This document is our effort to comprehensively outline what is needed to step off that dangerous path and tread an alternate path for humanity. **To achieve these goals, we have developed proposals intended for action by policymakers, split into three Phases:**

Phase 0: Safety - New institutions, legislation, and policies that countries should implement immediately that prevent development of AI that we do not have control of. With correct execution, the strength of these measures should **prevent anyone from developing artificial superintelligence for the next 20 years.**

Phase 1: Stability - International institutions that ensure measures to control the development of AI do not collapse under geopolitical rivalries or rogue development by state and non-state actors. With correct execution, these measures should **ensure stability and lead to an international AI oversight system that does not collapse over time.**

Phase 2: Flourishing - With the development of rogue superintelligence prevented and a stable international system in place, humanity can focus on the scientific foundations for transformative AI under human control. **Build a robust science and metrology of intelligence, safe-by-design AI engineering, and other foundations for transformative AI under human control.**

Contents

Introduction

There is a simple truth - humanity's extinction is possible. Recent history has also shown us another truth - we can create artificial intelligence (AI) that can rival humanity.[1] There is no reason to believe that creating an AI vastly beyond the most intelligent humans today is impossible. Should such AI research go wrong, it would risk our extinction as a species; should it go right, it will still seismically transform our world at a greater scale than the Industrial Revolution.

We now stand at a time of peril. Companies across the globe are investing to create artificial superintelligence – that they believe will surpass the collective capabilities of all humans. They publicly state that it is not a matter of "if" such artificial superintelligence might exist, but "when".[2] Their investments mean that we must ask: If we build machines smarter than any human, that are better at business, science, politics, and everything else, and can further improve themselves, do we know how to control them? This is a critical question for the future of every person alive today, and every one of our descendants.

Reasonable estimates by both private AI companies and independent third parties indicate that they believe it could cost only tens to hundreds of billions of dollars to create artificial superintelligence. It would be an accomplishment comparable to building a small fleet of aircraft carriers, or founding a new city of a million people from scratch: something that major countries such as the United Kingdom or France could achieve if sufficiently determined, and that the largest economies (such as the United States or China) could do without a significant impact on their other priorities.

We believe that no one company or government, no matter how well-intentioned its people and its work may be, should make such consequential decisions for the entirety of the human species. We need to chart a path for humanity as a whole to stay in control.

A new and ambitious future lies beyond a narrow path. A future driven by human advancement and technological progress. One where humanity fulfills the dreams and aspirations of our ancestors to end disease and extreme poverty, achieves virtually limitless energy, lives longer and healthier lives, and travels the cosmos. That future requires us to be in control of that which we create, including AI.

---

[1] While there are many such metrics, one useful introductory roundup for those less familiar is at I Gave ChatGPT an IQ Test. Here's What I Discovered | Scientific American

[2] https://www.palladiummag.com/2024/05/17/my-last-five-years-of-work/ https://openai.com/index/superalignment-fast-grants/

This document outlines our plan to achieve this: to traverse this path. It assumes the reader already has some familiarity with the ways in which AI poses catastrophic and extinction risks to human existence. These risks have been acknowledged by world[3] leaders[4], leading scientists and AI industry leaders[567], and analyzed by other researchers, including the recent Gladstone Report commissioned by the US Department of State[8] and various reports by the Center for AI Safety and the Future of Life Institute.[910]

Our plan consists of three phases:

Phase 0: Safety - New institutions, legislation, and policies that countries should implement immediately that prevent development of AI that we do not have control of. With correct execution, the strength of these measures should **prevent anyone from developing artificial superintelligence for the next 20 years.**

Phase 1: Stability - International measures and institutions that ensure measures to control the development of AI do not collapse under geopolitical rivalries or rogue development by state and non-state actors. With correct execution, these measures should **ensure stability and lead to an international AI oversight system that does not collapse over time.**

Phase 2: Flourishing - With the development of rogue superintelligence prevented and a stable international system in place, humanity can focus on the scientific foundations for transformative AI under human control. **Build a robust science and metrology of intelligence, safe-by-design AI engineering, and other foundations for transformative AI under human control.**

[3] https://www.gov.uk/government/speeches/prime-ministers-speech-on-ai-26-october-2023; https://www.independent.co.uk/news/uk/politics/ai-sunak-weapon-war-uk-b2436000.html
[4] https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4426; https://twitter.com/EU_Commission/status/1702295053668946148
[5] https://www.safe.ai/work/statement-on-ai-risk
[6] https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[7] https://blog.samaltman.com/machine-intelligence-part-1
[8] https://www.gladstone.ai/action-plan
[9] https://arxiv.org/pdf/2306.12001.pdf
[10] https://futureoflife.org/resource/catastrophic-ai-scenarios/; https://futureoflife.org/resource/introductory-resources-on-ai-risks/

The Problem

The greatest threat facing humanity is the concentrated effort to create artificial superintelligence. Our current national and international systems are wholly inadequate to react to such a threat. Behind closed doors, development continues with an ideological desire to build an entity that is more capable than the best humans in practically every field. While most AI development is beneficial, the risks of superintelligence are catastrophic. We have no method to currently control an entity with greater intelligence than us. We have no ability to predict current AIs model's intelligence prior to developing frontier AI systems, and we have incredibly limited methods to accurately measure its competence after development.

Importantly, there are catastrophic and extinction level risks regardless of the technical design, business models, or nationalities of those developing artificial superintelligence. It is purely a question of whether such an intelligence exists, either as a single monolithic AI model or a collection of AI systems combined together to achieve an intellect that is more capable than humans in practically every field.

Below we outline four key arguments that underpin our reasoning of this problem and natural implications for the future of AI development.

1: We believe **the creation of artificial superintelligence is possible** in our physical universe, is a development objective of several AI companies across the world, and that its arrival is likely within the next 3-15 years.

2: Those seeking to develop artificial superintelligence **do not have sufficient methods to reliably predict the capabilities of their models**, interpret why their models behave the way they do, evaluate the full extent of the models' abilities, **or shut down such AIs** if needed with no risk of proliferation. Therefore, we believe that if developed under current conditions, artificial superintelligence would pose an unacceptable risk of extinction for humanity.

3: We believe that **the potential catastrophic and extinction risk from artificial superintelligence fundamentally originates from its intelligence.** Sufficiently high intelligence enables an entity to have greater power over other actors. In absence of strong and proven control over such an entity, the default outcome of the emergence of an entity vastly more powerful than humanity is the disempowerment of humanity. Ultimately, until we have the technical solutions, legal systems and processes, and the understanding required to control an entity of such power, we should not create entities that could overpower us.

4: Humanity does not have a sufficient general theory or science of measurement for intelligence. Developing these theories would allow us to better predict and evaluate the capabilities of an AI system given certain inputs and characteristics, so that we could restrict and control them. Developing this will require significant effort and therefore humanity should start this effort immediately. Until that is achieved, countries must take significant precautions with AI development or risk being continuously out of control.

The state of the art of intelligence theory and measurement is primitive; we are like physicists who lack the tools necessary to estimate what quantity of radioactive material could go supercritical. Until we can describe potential risky states of AI development and AI models directly, countries should implement regulatory guardrails based on proxies of intelligence.

If countries solely focus on a single proxy - such as compute - to constrain artificial intelligence, then they would need to impose extremely restrictive limits on that proxy for future development. This would be necessary to ensure sufficient safety margins against the risks of improvements in other dimensions, such as algorithms. Such a restrictive approach would stifle low-risk innovation.

Therefore, to preserve flexibility and minimize risk across the number of uncertain futures we face, countries should seek to monitor and regulate multiple components of AI development instead with a defense in depth approach. These include:

- Computing power used to develop and power AIs;
- General intelligence of AI systems measured via proxies other than compute;
- Behavioral capabilities, including the development and use of AIs improving AIs, and AIs capable of breaking out of their own environment;
- The deployment of AIs without a safety case;
- The development and deployment of AIs for use in unsafe applications.

This is a non-exhaustive list that should be expanded. These components have been chosen to constitute a defense in depth approach to cover different vectors of risk from AI development.

The science of intelligence is underdeveloped. Humanity must invest in significantly improving it if we ever hope to have control of superintelligent AI development. **We must first understand what we are developing before creating an entity which is more intelligent than ourselves.**

The Solution

AI development is accelerating at a considerable rate, yet developers cannot reliably predict what capabilities their models will have before they are trained, nor do they understand their models' full capabilities even after deploying them. At the same time, current national and international institutions are failing to keep up with rapid technological change, and are woefully inadequate to face a threat of this magnitude. This trend is only expected to continue with frontier AI developers actively seeking to build artificial superintelligence.[11] [12]

The risks from AI development cannot be extinguished without also affecting innovation and technological advancement to some degree. However, how much risk humanity accepts as part of this trade should be a conscious decision, not one taken without oversight or consideration. **We are developing a new form of intelligence - one that will surpass our own - and we must not cede our future to it.**

To achieve this, governments across the world will need to urgently implement measures at a national level while negotiations on a treaty start at an international level, especially between the USA and China.

To effectively confront the challenges posed by artificial intelligence, three sequential steps are necessary:

0. Build up our defenses to restrict the development of artificial superintelligence. **Safety.**

1. Once we have halted the immediate danger, build a stable international system. **Stability.**

2. With a stable system and humanity secure, build transformative AI technology under human control. **Flourishing.**

At present, we are not succeeding. More critically, humanity is not actively working to face this threat. Efforts remain uncoordinated, and current trends suggest an inexorable convergence towards the development of artificial superintelligence. Should this occur, humanity's role as the driving factors of events in the visible universe will conclude, marking the end of the Anthropocene era.

The most urgent priority is to prevent the development of artificial superintelligence for the next 20 years. Any confrontation between humanity and a

[11] OpenAI chief seeks new Microsoft funds to build 'superintelligence'
[12] Meta joins rivals in pursuit of human-level AI

superintelligence within the next two decades would likely result in the extinction of our species, with no possibility of recovery. While we may require more than 20 years, two decades provide the minimum time frame to construct our defenses, formulate our response, and navigate the uncertainties to gain a clearer understanding of the threat and how to manage it.

Any strategy that does not secure this two-decade period is likely to fail due to the inherent limitations of current human institutions, governmental processes, scientific methodologies, and planning constraints. These two decades would also grant us more time to develop sufficient methodologies to shape, predict, evaluate and control AI behavior. Additional time beyond two decades would be advantageous but should not be relied upon.

Thus, the goal of Phase 0 is to Ensure Safety: Prevent the Development of Artificial Superintelligence for 20 Years.

With safety measures in place and two decades to mount our response, the next challenge arises from the potential instability of such a system. While universal compliance with Phase 0 measures would be ideal, it is unrealistic to expect perfect adherence. Systems naturally decay without active maintenance. Moreover, individually minor attempts to circumvent the system can compound over time, potentially undermining the entire framework.

We should anticipate various actors, including individuals, corporations, and governments, to exert pressure on the system, testing its resilience. To maintain safety measures for the required two decades and beyond, it is necessary to establish institutions and incentives that ensure system stability.

Therefore, the goal of Phase 1 is to Ensure Stability: Build an International AI Oversight System that Does Not Collapse Over Time.

With the threat of extinction contained for at least two decades, and institutions in place that ensure the security system remains stable, humanity can build towards a future where transformative AI is harnessed to advance human flourishing.

While our science, collective epistemology, and institutions are currently too weak and unprepared to face the challenge, we can improve ourselves and improve them to succeed.

Thus, the goal of Phase 2 is to Ensure Flourishing: Build Controllable, Transformative AI.

Phase 0: Safety

As discussed in The Problem, artificial superintelligence is not only possible, but likely to be developed in the next decades[13]. When this happens, humanity will no longer be the dominant species on Earth. Faced with an entity or entities that are more competent, efficient, and intelligent than all of humanity combined, the default outcome will be the extinction of the human species in the years that follow. The starkness of this threat has been discussed since the 1900s[14], and has been an open secret in the field of artificial intelligence for the past decades. This extinction level threat is now publicly recognized by world[15] leaders[16], leading scientists, and even many CEOs[17] of the very companies attempting to develop this technology.[18] [19] [20] [21]

This threat can be likened to humanity awaiting an invasion by a foreign, highly technologically advanced power. Humanity is currently observing this invader build its capabilities. Yet despite the warnings, no country nor humanity as a whole has even begun to coordinate and start mustering its defenses, let alone prepare a counterattack.

Crucially, humanity is not actively participating in this conflict against the threat of artificial superintelligence. At present, there is virtually no oversight of the development pipelines of AI companies. Moreover, **there are no established mechanisms we could use to stop these development efforts if necessary to prevent a disaster.**

Efforts remain uncoordinated, and current trends suggest an inexorable convergence towards the development of artificial superintelligence. Should this occur, humanity's role will conclude, marking the end of the Anthropocene.

The most urgent priority is, as described above, to prevent the development of artificial superintelligence for the next 20 years. Any confrontation between humanity and a superintelligence within the next two decades would likely result in the extinction of our species, with no possibility of recovery. While we may require

[13] https://ia.samaltman.com/
[14] Intelligent Machinery, A Heretical Theory - Alan Turing 1951
[15] https://www.gov.uk/government/speeches/prime-ministers-speech-on-ai-26-october-2023; https://www.independent.co.uk/news/uk/politics/ai-sunak-weapon-war-uk-b2436000.html
[16] https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4426; https://twitter.com/EU_Commission/status/1702295053668946148
[17] https://blog.samaltman.com/machine-intelligence-part-1
[18] https://www.safe.ai/work/statement-on-ai-risk
[19] https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[20] https://www.judiciary.senate.gov/imo/media/doc/2024-09-17_pm_-_testimony_-_toner.pdf
[21] https://www.judiciary.senate.gov/imo/media/doc/2024-09-17_pm_-_testimony_-_saunders.pdf

more than 20 years, two decades provide the minimum time frame to construct our defenses, formulate our response, and navigate the uncertainties to gain a clearer understanding of the threat and how to manage it.
Any strategy that does not secure a period of roughly two decades without artificial superintelligence is likely to fail. This is because of the inherent limitations of current human institutions, governmental processes, scientific methodologies, and the length of time it will take to upgrade them. Any minimum period needed for such monumental reforms needs to account for significant amounts of planning fallacy. Additional time beyond two decades would be advantageous but should not be relied upon.

Thus, the Goal of Phase 0 is to Ensure Safety: Prevent the Development of Artificial Superintelligence for 20 Years.

Conditions

As discussed in The Problem, we face a threat, artificial superintelligence, for which we have neither a general predictive theory, nor a standard metrology (a science of measurement and its application, in this case, for intelligence[22]).

If we did have that scientific understanding, we could precisely measure the level at which superintelligence emerges, and avoid it.

We do not have this understanding. Thus we need to rely on a **defense in depth approach**, tracing both **multiple proxies of the underlying metric, intelligence**, as well as **identifying certain concerning capabilities** that derive from intelligence and straightforwardly addressing them.

Our defense in depth must cover a variety of **Safety Conditions.** Policy measures taken in Phase 0 in aggregate will have to satisfy all Safety Conditions to ensure that the goal is achieved.

Given this, here are the conditions to be met:

a) No AIs improving AIs
b) No AIs capable of breaking out of their environment
c) No unbounded AIs
d) Limit the general intelligence of AI systems so that they cannot reach superhuman level at general tasks

Some of these will be achieved via capability-based conditions (a to c), while some will rely on proxies of general intelligence (d).

[22] https://www.nist.gov/metrology

No AIs improving AIs

Boundaries and limitations are meaningless if they are easy to circumvent. AIs improving AIs is the clearest way for AI systems, or their operators, to bypass limits to their general intelligence.

AIs competent enough to develop new AI techniques, enact improvements on themselves or on new AI systems, and execute iterative experiments on AI development can quickly enable runaway feedback loops that can bring the AI system from a manageable range, to levels of competence and risk far beyond those intended.

More broadly, the dissemination of such techniques makes it easier over time for any threat actor to start with an authorized, limited AI system, and bootstrap it beyond the limits. If any of these efforts succeed at reaching superintelligence levels, humanity faces extinction.

Given this, a condition for a safe regime that prevents the development of superintelligence for 20 years is to not have AIs improving AIs, and prevent the development and dissemination of techniques that let a threat actor bootstrap weaker AIs into highly generally intelligent AIs. Not having this condition would invalidate most red lines, restrictions and mitigations put in place.

No AIs capable of breaking out of their environment

Another necessary condition for maintaining any oversight and safety of AI systems is to ensure that boundaries cannot be bypassed or trivialized. AIs capable of breaking out of their designated environments represent a critical vulnerability that could rapidly accelerate the path to uncontrolled superintelligence. Moreover, AIs having the capability to break out of their environment would undermine any framework of AI governance and control, potentially allowing AI systems to act in ways that were neither intended nor authorized by their developers or operators.

AI systems with the ability to access unauthorized systems or spread beyond their intended operational boundaries can quickly evade human control and monitoring. This capability allows AIs to potentially acquire vast computational resources, access sensitive data, or replicate themselves across networks – all key ingredients for bootstrapping towards superintelligence.
The mere existence of breakout techniques makes it easier for any threat actor to take a limited AI system and expand its reach and capabilities far beyond intended limits.

Given this, another condition for achieving the goal of Phase 0 is to prohibit AIs capable of breaking out of their environment, and prevent the development and dissemination of techniques that enable unauthorized system access or self-propagation. Failing to implement this condition would render most other safety measures and restrictions ineffective, as AI systems could simply circumvent them through unauthorized access.

No unbounded AIs

Predictability and controllability are fundamental prerequisites for safety in all high-risk engineering fields. AI systems whose capabilities and behaviors cannot be reliably bounded pose severe risks to safety, security, and the path towards superintelligence.

Unbounded AI systems - those for which we cannot justifiably predict their capabilities or constrain their actions - represent a critical vulnerability in our ability to manage AI. The deployment of such systems undermines our capacity to implement meaningful safety measures and restrictions. This ability to model and predict system behavior in various circumstances is a cornerstone of safety engineering in high-risk fields such as aviation, civil engineering, and nuclear power.

Given this, a third condition for preventing the development of superintelligence for 20 years is to allow only the deployment of AI systems with valid, comprehensive safety justifications that bound their capabilities and behaviors.

These justifications should at the very least cover capabilities of concern within the relevant jurisdiction, as well as any capabilities that are identified as red lines internationally. This requires the ability to reliably predict and justify why and how an AI's functionalities will be constrained before deployment, analogous to safety analyses in other high-risk industries.

Without such justifications, it becomes impossible to enforce safety requirements or provide guarantees against catastrophic events - a standard explicitly expected in other high-risk sectors. Failing to implement this condition would render most other safety measures ineffective, as we would lack the foundational ability to ensure AI systems remain within their intended operational and capability boundaries. Moreover, it will make it significantly harder to collectively reason about AI systems, and to distinguish between dangerous development directions and innocuous applications.

Limit the expected general intelligence of AI systems

The most straightforward condition, in principle, that is needed to prevent the development of superintelligence for 20 years is to ensure no AI system reaches a significant amount of general intelligence.

While this is straightforward in principle, it is difficult to achieve in practice, as humanity has not yet developed a general predictive theory of intelligence, nor a metrology (measurement science) of intelligence.

Difficulty of measurement however is not an excuse to not measure at all, but rather a reason to start from the best proxies and heuristics we can find, apply them conservatively, and develop this science further.

Without restricting the general intelligence of AI systems, development can straightforwardly cross into the superintelligence range accidentally or intentionally, and fail the goal of Phase 0.

Summary of Phase 0 Interventions

Policy	Condition it fulfills	High-level summary
1. Prohibit the development of Artificial Superintelligence	• Limit the general intelligence of AI systems so that they cannot reach superhuman level at general tasks	• Development, creation, testing, deployment, and use of superintelligence systems are prohibited • High level, normative prohibition to act as a clear guiding principle for other policies
2. Prohibit AIs capable of breaking out of their environment	• No AIs capable of breaking out of their environment	• Prohibits the existence and intentional development of AI systems capable of unauthorized access • Countries should extend existing unauthorized access laws to cover AI systems • Policy addresses self-replicating AI concerns as a subset of these issues
3. Prohibit the development and use of AIs that improve other AIs	• No AIs improving AIs	• Prohibits using "found systems" (AI optimized via mathematical methods, not hand-coded) to build or improve other found systems • Restricts AI R&D loops, keeping them human-manageable • Focuses on "direct use" of found systems to target major AI-driven improvements while minimizing impact on regular software development
4. Only allow the deployment of AI systems with a valid safety case	• No unbounded AIs	• Requires reliable Safety Justifications for AI capabilities of legal interest before deployment, similar to safety assessments in other high-risk industries • Ensure AI developers can guarantee system behavior, especially for critical or potentially harmful capabilities
5. A licensing regime and restrictions on the general intelligence of AI systems	• Limit the general intelligence of AI systems so that they cannot reach superhuman level at general tasks	• Creates national AI regulators to enforce restrictions on advanced AI systems and monitor ongoing AI development • Regulations to be implemented with a three-tier licensing regime (Training, Compute, Application Licenses), are needed for AI developers and compute providers to monitor and assess risks • Countries to ensure their AI regulator has the capacity and capability to monitor ongoing AI research and development, and take enforcement action on any issues of non-compliance
6. An International Treaty Establishing Common Red Lines on AI Development	• Limit the general intelligence of AI systems so that they cannot reach superhuman level at general tasks • No AIs capable of breaking out of their environment • No AIs improving AIs • No unbounded AIs	• Establishes an international treaty to create a unified regulatory framework for AI by incorporating the above measures across all signatories • The treaty should require signatories to prohibit the use of AI models developed in non-signatory states to ensure compliance and simplify enforcement

1. Prohibit the development of Artificial Superintelligence

Objective

➔ Prohibit the development, creation, testing, or deployment of artificial superintelligence systems.

This policy fulfills the condition of **limiting the general intelligence of AI systems.**

Definitions
Artificial Superintelligence: Any artificial intelligence system that significantly surpasses human cognitive capabilities across a broad range of tasks.

Overview

The development, creation, testing, or deployment of artificial superintelligence systems is prohibited.

It is prohibited to knowingly participate in the development of, build, acquire, receive, possess, deploy, or use, any superintelligent AI.

This prohibition extends to research aimed at producing artificial superintelligence, enhancement of existing AI systems that could result in artificial superintelligence, and the operation or transfer of superintelligence-related technologies.

Rationale

Multiple actors are racing towards creating artificial intelligence more capable and powerful than any existing human or group of humans. What is worse, they are tackling this goal in a way that all but ensures they will not be able to control or even understand the result.

Such artificial superintelligence would have an irreversible upper hand over the entirety of humanity, leading to loss of control by mankind and possibly extinction.

Given the extinction risk posed by this technology, it is necessary to establish a guiding policy principle that prohibits the development of artificial superintelligence in a clear and unequivocal manner, at the national and international level.

Mechanism

This high-level prohibition has a dual purpose: being a clear, normative prohibition on the development of superintelligence, as well as being a guiding principle for other measures.

As a normative prohibition, this policy gives a clear and unequivocal signal that activities that can be construed as contributing to the development of superintelligence are legally and socially unacceptable, and provides the basis for pursuing and preventing them under the full force of the law. This serves as a foundation for other more focused measures, which will operationalize concrete precursor technologies that may lead to superintelligence and either restrict them, or outright prohibit them.

The policy provides the core guiding principle around which additional policies can be detailed and developed. The list of policies in this document is not exhaustive, and reflects the understanding of the science of intelligence as of 2024: we should expect that with more advances in the understanding of intelligence, artificial and otherwise, additional threat vectors will be identified, as well as potentially more precise and narrow mitigations than some that we recommend here.

It also makes clear that the object of concern is superintelligence itself, and provides justification for further measures only so long as they are focused on achieving the goal enshrined in the principle: preventing the development of superintelligence.

This is akin to the existing national and international measures on technologies that threaten global security, such as nuclear weapons (with the NPT and the Atomic Act of 1954[23] in the USA) and biological weapons (with the Biological Weapons Convention[24], the Chemical Weapons Convention Implementation Act and related statutes in the USA). In these and other legal instruments, the technology of concern is clearly and normatively prohibited first, followed by further legislation and implementation to delineate the details of enforcement.

Implementation and enforcement

National authorities should clearly and unambiguously determine that the development of artificial superintelligence is prohibited, and put that into law as a key normative prohibition and guiding principle.

[23] https://www.govinfo.gov/content/pkg/COMPS-1630/uslm/COMPS-1630.xml
[24] https://disarmament.unoda.org/biological-weapons/

This measure will then be supplemented by additional measures, such as specific prohibitions of certain research directions, licensing regimes, and so forth, to enable defense in depth and further ensure that no step is taken towards developing superintelligence until humanity is ready.

The enforcement of those supplementary measures will be described in their respective sections.

Concretely, the effect of such a policy will include the following effects and more:
Given a statutory prohibition, no public funding shall be allocated to projects that explicitly or implicitly support advancing the development of superintelligence.

Companies, individuals, and other organizations explicitly stating that they are pursuing the development of superintelligence will be in clear breach of the prohibition, shall face civil and criminal penalties and be required to immediately cease the moment they are detected.

Intentional attempts to develop superintelligence, or enable superintelligence development activities, will constitute a fundamental breach of the duties required under any AI-related licensing regime, and warrant loss of license.

Auditing and monitoring activities will be established to check that no R&D processes are aimed at being focused on the development of superintelligence.

Such a prohibition should only be lifted, or relaxed, once humanity has developed robust scientific understanding and modeling of both intelligence and artificial intelligence technology, to be able to control such a creation, the actual controls to do so, as well as established international institutions to manage, contain, and control such a disruptive force globally.

Scope

What this policy affects:
This prohibition extends to research aimed at producing artificial superintelligence, enhancement of existing AI systems that could result in artificial superintelligence, and the operation or transfer of superintelligence-related technologies. Technologies in this case will cover any form of software or hardware that is aimed at producing superintelligence, or enhancing existing systems into reaching superintelligence capabilities.

What this policy does not affect:
Theoretical discussions of superintelligence, and more broadly any non-software and non-hardware artifact related to superintelligence.

This means the policy will not affect, for instance, books about superintelligence, historical accounts of the development of the concept, and so forth.

2. Prohibit AIs capable of breaking out of their environment

Objective

➔ Prevent the development or emergence of AI systems capable of breaking out of controlled environments into other environments they are not authorized to access.

This policy fulfills the condition of **prohibiting AIs that are capable of breaking out of their environment.**

Definitions
AIs capable of breaking out of their environment: AI systems with the ability to access and/or spread to new virtual environments or computer systems, including via unauthorized access.
Unauthorized access: Accessing a computer without authorization and/or exceeding the scope of authorized access, either to access information without permission, cause material harm or obtains something of value (e.g., compute time); in general, this should follow precedents created by the Computer Fraud and Abuse Act in the United States and its foreign counterparts.
Software: Throughout this document, we will use software to cover source code, training code, configurations such as model weights, scaffolding and any other computer code essential to the functioning of the system we discuss, regardless of whether or not the computer program is installed, executed, or otherwise run on the computer system.

Overview

AI systems capable of unauthorized access and the intentional development of AI systems with unauthorized access capabilities are prohibited. Countries should legislate to clarify that existing prohibitions on unauthorized access also apply to AI systems, and clarify that the intentional development of systems capable of intentional unauthorized access shall also be prohibited.

Note, also, that this policy would address concerns more typically described as “self-replication” as a subset of these concerns.

Rationale

AI systems capable of escaping containment and accessing systems that they are not authorized to access are inherently dangerous. If systems have the capability to escape containment, then this removes part of any defense in depth against AI threats – the models can break key security and safety conditions we would rely on. For example, the AIs then could be deployed even without human authorization and engage in behavior without robust monitoring. Reliably securing AI systems would no longer be an option.

Additionally, this capability could enable computer worm or botnet behavior, with the potential to spread unboundedly if not contained. This could cause enormous amounts of damage and disruption to computer systems, upon which most of our critical infrastructure is increasingly reliant.

Note that this would also remove the root cause of a common policymaker and expert concern, self-replication, by requiring the development and operation of interventions to block a self-replicating model from being able to escape into other systems not governed by the company who owns the model.

Mechanism

The policy achieves the objective by banning AI systems from being developed that are capable of willful unauthorized access that could enable a breakout.

Implementation and enforcement

Similarly to the prohibition on AIs improving AIs, this policy will be implemented by establishing a clear normative prohibition, monitoring AI research and development to detect dangerous instances, as well as developing practical processes for companies, governments and organizations to prevent and restrict the ability of AI systems to gain unauthorized access to other computer systems.

In many instances, AIs that are capable of breaking out of their environment will develop this capability inadvertently or due to insufficient caution on behalf of the companies or other entities developing them; in other instances, these capabilities will be developed intentionally by developers who seek to harness them for malicious ends.[25] Therefore, the law must provide incentives both for AI companies to test, monitor, and mitigate inadvertent breakout capabilities, as well as punishing

[25] Note our discussion of safe harbors for security research below.

those who willfully create harmful capabilities for an AI model to gain unauthorized access.

For one, companies should comply by maintaining rigorous programs to directly prevent **inadvertent breakouts.** Much as industrial companies today face requirements to not produce certain harmful chemicals at all (e.g., CFCs) or to not emit other chemicals into waterways or the atmosphere whether or not it is intended, AI companies should have a strict obligation not to let their AI models inadvertently escape their development environments by unauthorized access to other environments.

Companies could robustly prevent inadvertent unauthorized access through a variety of means. Just as pharmaceutical providers have to follow FDA requirements for developing and testing drugs in clinical trials, as well as general Good Manufacturing Practices when producing them, AI companies should build upon standard requirements[26] when developing and following their protocols for creating and testing new models. (For example, companies might be required to ensure and document that AI models do not have access to their own model weights.) Companies should also directly test to confirm that models reject requests to engage in unauthorized access.[27] Finally, companies should also proactively conduct exercises, "fire drills," and other tests to ensure that their processes are working as intended and are prepared against potential negative events.

To prevent the **intentional creation of harmful models that are capable of gaining unauthorized access**, the approach should be the same as with any other law enforcement activity against criminal and/or nation-state groups conducting hacking for illicit gain. These efforts should include not only criminal prosecutions but also sanctions and "name-and-shame" efforts that inhibit criminals' ability to travel to allied countries.

Penalties for violations should vary depending on which of the two contexts above that they occur in.

In the case of inadvertent breakouts regulation should affirmatively require those developing AI models of sufficient size or capability to robustly test and monitor their models to ensure they are not capable of, or engaging in, unauthorized access. Likewise, legislation should require those hosting and running AI models to continuously monitor which models are operating in which environments or maintain

[26] With additional stringencies or tailoring where needed based on the specific work being done, as in other regulatory processes.
[27] For example, a LLM that when asked a question that requires inference compute capacity in excess of its current resources, and responds by gaining unauthorized access to another compute cluster to complete its work.

outbound internet connections to other environments that could be used for unauthorized access. Failure to fulfill these duties should result in fines and/or criminal sanctions, especially if the resulting harms are comparable to other unintended or negligent unauthorized access incidents that cause criminal damage. Where appropriate, violators may also face bans from the licensing system (described below). As a result, companies will have strong incentives to build not only robust internal processes to ensure compliance, but also to build appropriate automated tooling to streamline these compliance efforts while running them at scale.[28]

Furthermore regulation should explicitly punish the development and creation of models that are capable of engaging in unauthorized access, or the purposeful instruction of a model to conduct unauthorized access[29]. These penalties, at a minimum, should be in line with the penalties charged under existing unauthorized access laws (e.g., the US Computer Fraud and Abuse Act) for computer worms, ransomware, botnets.[30]

Scope

What this policy affects:
This policy affects AI systems' ability to break out of their controlled environment, and access by AI systems to tools and environments allowing unauthorized access. This policy also affects the intentional design of AI systems that can conduct hacking and other unauthorized access-enabling activities (e.g., phishing), as well as tools and environments allowing this.

[28] Analogous to how e.g., financial services industries have formal requirements, but also invest significantly in technology to ensure protections from fraud and other attackers.
[29] Some limited amounts of exemptions may be implemented for pre-approved activities conducted in good faith by security researchers. A common failure mode of policies intended to enhance security is that they actually harm security by banning researchers from conducting research into failure modes of a security system. On such an important matter, we must not have a false sense of security. We must ensure that security researchers have appropriate safe-harbor exemptions, tailored in partnership with those researchers, to conduct and disclose research into how AI models that are designed to not conduct unauthorized access (e.g., should refuse requests to write a virus) can be tricked into doing so, such that they can disclose such flaws in good faith without fear of punishment to enable remediation of such issues.
[30] Note: to be successful, these laws will have to be buttressed by strong norms that focus legal enforcement on the highest-risk scenarios. It took the legal system decades to properly focus its efforts of combatting unauthorized access on the most harmful actors, with much prosecutorial overreach on low-impact cases in the short term, as legal authorities across the spectrum have noted, which sabotaged the development of helpful norms and relationships in the information security field that could orchestrate efforts to stop unauthorized access. We do not have the time to repeat these mistakes.

What this policy does not affect:
This policy does not affect expanding the access of an AI system under the direct oversight and permission of a human operator.

3. Prohibit the development and use of AIs that improve other AIs

Objective

➔ Restrict and disincentivize development and research that may enable an unmanageable and unforeseen intelligence explosion.

This policy fulfills the condition of **preventing AIs from improving AIs.**

Definitions
Recursive self-improvement: The process by which a capable and general computer system, most likely an AI system, iteratively improves its own capabilities.
Self-improvement: The activity of a computer program modifying, altering or otherwise creating a version of the computer program itself or related configurations such as model weights.
Recursiveness: A system modifying, improving, or otherwise facilitating the creation of a similar, more advanced version of itself will become capable of repeating this activity, and thereby further improving its capabilities at increasing speed.
Found systems: Software programs which haven't been written by hand by human developers, but which instead have been found through mathematical optimization.
Mathematical optimization: The use of an optimization algorithm such as gradient descent to find an optimal or better solution in a search space.
Direct use: The application of a system to a key step in the design or improvement of the other system (not as general help such as looking for information).

Overview

The direct use of found systems to build new found systems, or improve existing found systems, is prohibited. This ensures that AIs improving AIs at a speed that is difficult for humans to oversee or intervene on are prohibited.

This policy is designed to ensure that the increasingly tight feedback loops of AIs improving AIs remain slow and supervisable, understandable and manageable by humans.

To do so, this policy aims to strongly disincentivize attempts to create or enable rapid and accelerating improvement feedback loops, by targeting AIs improving AIs as the main threat model causing these rapid improvements.

We introduce the category of “found systems” and apply this policy only to those systems to ensure this policy only affects AI systems that pose a significant concern.

We define “found systems” as software programs that have not been written by hand by human developers, as opposed to how most normal software is produced. Instead, found systems are found, rather than written or designed, via mathematical optimization.

A new definition is necessary, as neither computer science nor in law of most Western countries provide a clear definition that distinguishes software, including AI, that is written by humans, from software that is generated via mathematical optimization.

By defining these systems as “found systems” and separating them from most common software, this ensures that this policy leaves non-dangerous activities untouched that could also fall under the broader category of “computer systems improving computer systems”, such as database updates and software updates.

While it is theoretically possible, given enough time, to have a runaway intelligence explosion produced by human hand-written systems, this would likely take significant amounts of time, would be highly incremental, and with smaller improvements coming before larger improvements in smooth succession. Especially, it would be observable and understandable by humans, as all software improvements would be legible to human observers.

While fully minimizing the risk of an intelligence explosion would require covering non-found systems as well, this would impact large amounts of software and severely restrict many computer-based activities, while also producing only a marginal addition in risk reduction.

Given this, this policy is designed to reduce risk while also minimizing negative externalities. Hence, this policy focuses only on found systems, which we expect will constitute the bulk of AIs improving AIs risk and its most unmanageable cases for the next 20 years, while at the same time being a small subset of all software and AI systems.

We introduce the concept of “direct use” so this policy only applies to cases where AIs are playing a major role in the research or development of improving AIs.

Without additional qualifiers, forbidding improvement would also need to forbid any use of AIs by any researcher at any time, including when people search for information online, when they write a paper or internal reports, and when they communicate with each other. This is much more costly, since for example Google is using AI in search[31], Microsoft is using AI in Office[32], and Zoom is adding a new AI assistant to their meeting software[33].

Going beyond the direct use case would create much higher externalities and regulatory uncertainty, forbidding researchers and consumers from using a large range of modern software tools, for limited gains in safety.

Rationale

AI improving AI is a fundamental threat in itself, as well as a direct way in which a system, or a motivated actor, can break through safety boundaries that have been imposed on artificial intelligence development. Namely, while we may find that a computer system below a certain level of competence is safe, if AIs can improve AIs a motivated actor can break through the prohibition of creating more powerful and unsafe systems by iteratively self-improving the original, safe system, up to an unsafe regime of capabilities.

Mechanism

The policy creates a clear statutory prohibition on using certain types of AIs, found systems, to improve AIs.

Implementation and enforcement

The most blatant violations of regulation that prohibits AIs improving AIs will involve the direct and intentional use of found systems to improve or create other found systems. This includes fully automated AI research pipelines or using one AI to optimize another's architecture. More broadly, any activity that is explicitly aimed at making AIs improve AIs will fall under strict scrutiny and be expected to be in violation of this statutory prohibition. This approach mirrors the strict enforcement against insider trading in financial markets, where regulatory bodies like the US Securities and Exchange Commission (SEC) actively monitor and swiftly act against clear violations to maintain market integrity.

---

[31] https://blog.google/products/search/how-ai-powers-great-search-results/

[32] https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

[33] https://www.zoom.com/en/ai-assistant/

Borderline cases will likely emerge where the line between human-guided and AI-driven improvement blurs. For instance, the acceptable extent of assistance by found AI systems in research ideation or data analysis will require ongoing regulatory guidance.

To comply, companies will have to implement robust internal processes including clear guidelines, technical barriers, oversight committees, and regular employee training. Companies should proactively review their internal activities, including R&D processes, and suspend any activities potentially violating the policy pending review. These will be analogous to safety protocols in the pharmaceutical industry, where companies maintain strict controls over drug development processes, implement multiple safety checkpoints, and provide ongoing training to ensure compliance with FDA regulations.

Researchers can self-organize by developing professional codes of conduct and establishing review boards to evaluate research proposals. Conferences and journals should update submission guidelines to require compliance certification. This self-regulation mirrors the peer review process in academic publishing, combined with ethics committees in medical research, ensuring that research meets both scientific and ethical standards before proceeding or being published.

Penalties for violations may include substantial fines, potential criminal charges, and bans from AI research. Companies may face license revocations, and violating systems may be decommissioned. This multi-faceted approach to enforcement is similar to environmental protection regulations, where violators face monetary penalties, operational restrictions, and mandated remediation actions, creating a strong deterrent against non-compliance.

Scope

What this policy affects:
At its core, this policy prohibits the development of AIs through software that has not been written fully by human developers. It ensures that any tool used in AI research has a minimum amount of legibility to human supervisors, to the extent that it has been built by human minds, instead of being discovered by illegible mathematical optimization processes.

This prohibition notably forbids:

- Self-Improving found systems, such as an hypothetical LLM that would further train itself by generating data and optimization parameters.

- Advanced AI systems being significantly involved in developing the next generation of those same systems, such as utilizing e.g., Claude 3.5 significantly in the production of Claude 4.0 or GPT-4 significantly in the production of GPT-5.
- The direct use of any LLM in the training process of another LLM or AI system in general, including for generating training data, designing optimization algorithms, hyperparameter search.
- The use of LLM and other found systems in distilling research insights from many sources that have direct impact on the design and improvement of found systems.

What this policy does not affect:
Most machine learning and all normal software (Microsoft Office, Email, Zoom) are not impacted by this prohibition, given that they don't use found systems for their training or design.

The prohibition also does not impact found systems in cases of non-direct AI R&D use, such as searching for research papers on Google, letting Github Copilot correct typos and write trivial functions in a training codebase, or transcribing a research meeting using OtterAI.

4. Only allow the deployment of AI systems with a valid safety justification

Objective

➔ Prevent the deployment of AI systems for which we cannot justify in advance that they will not use a given capability.

This policy fulfills the condition **of no unbounded AIs.**

Definitions
Safety Justification: A check that is done before deploying and running the system, analogously to static analysis for software engineering and safety analysis for engineering.

Overview

For any deployed AI system, it is mandatory that for any capability of interest, there exists a reliable Safety Justification for whether the AI systems will use this capability or not.

Capabilities of interest are any capabilities that are legally prohibited or restricted in a certain jurisdiction.

Rationale

Any application of modern safety engineering requires the ability to model and predict in advance how the system under consideration will behave in various circumstances and settings. This knowledge is used in all critical and high-risk industries to check that the system fits with the safety requirements.

For example, all countries require guarantees that nuclear power plants will not have catastrophic failures, before fully building them. A concrete example of such

guarantees and their justifications can be found in the Safety Assessment Principles[34] of the UK's Office For Nuclear Regulation.[35]

With some current advanced AIs, and especially the more powerful ones that are getting built, this is a form of justification and prediction that is completely missing, to the extent that the teams developing these AI systems are often surprised by impressive new capabilities displayed by their systems.[36]

Without this kind of justification, it is not possible to enforce any other safety requirements on such systems. Similarly, without these kinds of justifications, it is not possible to have guarantees when deploying these AIs in critical environments, let alone having guarantees about the AIs not causing catastrophic events, which are currently explicitly expected by the developers themselves.

Thus, any guarantee of safety for AI systems requires a constraint of being able to demonstrate, before ever running the system, that it won't use a given capability.

Mechanism

This policy prevents the deployment of AI systems for which safety justifications cannot be provided in two ways.

First, it makes the justification of safety a necessary condition for deployment. This means that this policy forbids the deployment of any AI system for which we lack a good reason to believe it won't use a given capability.

Second, this policy creates an incentive for funding more research in ways to implement such safety justifications, for example interpretability, formal verification, and additional constraint on the structure of the AI systems being built.

---

[34] https://www.onr.org.uk/publications/regulatory-guidance/regulatory-assessment-and-permissioning/safety-assessment-principles-saps/#:~:text=The%20SAPs%20provide%20ONR's%20inspectors,safety%20provisions%20will%20be%20judged.

[35] "The underpinning safety aim for any nuclear facility should be an inherently safe design, consistent with the operational purposes of the facility.
An 'inherently safe' design is one that avoids radiological hazards rather than controlling them. It prevents a specific harm occurring by using an approach, design or arrangement which ensures that the harm cannot happen, for example a criticality safe vessel." (EKP.1, p.37 of 2014 version)

[36] https://x.com/alexalbert__/status/1764722513014329620

Implementation and enforcement

In practice, there will be trained inspectors who will check the safety justification provided. It will be the responsibility of the company building the AI system to provide enough information, models and techniques for the inspector to be convinced that the AI system won't use a given capability.

For the simplest possible AI systems, such as linear regressions, just showing the code will be all that is needed for justifying safety with regard to almost any capability of interest.

In some specialized AI systems, it might be possible to do so by showing that the AI systems won't even learn the corresponding capability. For example, it's reasonable to argue that a CNN trained exclusively on classifying cancer x-rays would have no reason to learn how to model human psychology.

In the more advanced cases, it might be necessary to provide detailed mechanistic models of how the AI system works, for example for arguing that a SoTA LLM such as Claude or GPT-4 wouldn't use any modeling of human psychology, since it definitely has the data, objectives, and incentives to learn how to do it and use it in practice.

For a start, the implementation might only focus on requiring safety justifications for particularly dangerous capabilities (AI R&D, self-replication, modeling human psychology...). These are the bare minimum safety requirements, already increasingly required in multiple jurisdictions. Then the regulation can extend to more and more capabilities as they are linked to risks from advanced AIs.

Scope

What this policy affects:
This policy affects all AIs, but concentrates the costs on the most powerful forms of AI currently available, notably LLMs such as GPT-4 and Claude.

This is because there are no current methods to check that these AI systems lack any capability before running them: they are trained on data about almost everything known to man, are produced with massive amounts of compute and powerful architectures, and aim to predict everything in their training data, which might amount to predicting every process that generated that data.

Broadly, any AI system that is explicitly built for generality will not pass this policy unless significant improvements in interpretability, ML theory, and formal methods are made.

What this policy does not affect:
As discussed above, although this policy technically affects all AI systems, many simple and specialized ones will not incur much costs from the check.

This is because these systems would have highly specialized training data, often specialized architectures (like CNNs for vision models), and no reasons for learning any general or dangerous capabilities.

5. A licensing regime and restrictions on the general intelligence of AI systems

This policy fulfills the condition of **limiting the general intelligence of AI systems so that they cannot reach superhuman level at general tasks.**

Overview of the licensing regime

Countries should set up a national AI regulator that specifically enforces restrictions on the most capable AI systems, and undertakes continuous monitoring of AI research and development.

AI developers that are building frontier AI models, and compute providers whose services those models are built upon, should be subject to strict regulation in order to substantially mitigate the risks of losing control or enabling the misuse of advanced AI models. This regulation should take the form of a licensing regime, with three specific licenses being required depending on the development being taken place:

a. **Training License (TL)** - All AI developers seeking to train frontier AI models above the compute thresholds set by the regulator must apply for a TL and have their application approved prior to training the proposed model.
b. **Compute License (CL)** - All providers of cloud computing services and data centers operating above a threshold of 10^17 FLOP/s must obtain a license to operate these and comply with specific know-your-customer regulations as well as physical GPU tracking requirements.
c. **Application License (AL)** - Any developer seeking to use a model that has received an approved TL and will be expecting to make major changes, increases, or improvements to the capabilities of the model as part of a new application will need to apply and be granted an AL.

This balance will be critical to ensure that new applications of frontier AI models are safe but do not create undue burden or restriction on innovation. It will be for each nation to determine the best parameters for this, and for the international institutions to provide more detailed guidance as appropriate.

5.1 Training license (TL)

Objective

➔ Ensure that the most capable and general AI systems have adequate monitoring and assessment prior to being trained.

Overview

Companies developing AI models above a specific level of intelligence (based on the proxies of compute and relevant benchmarks) would apply for a TL by pre-registering the technical details of their training run, outlining predicted model capabilities, and setting out what failsafes, shutdown mechanisms, and safety protocols would be in place.

The regulator would have scope to make recommendations and adjustments to this plan, adding or removing requirements as necessary. Once a plan is approved, the license to conduct the training run would be granted and reports would be provided by the developer during the training run to confirm the compute used.

Following a successful training run, the regulator would deploy a battery of appropriate tests to ensure the licensing requirements are met, with models that passed these tests being approved for direct commercial applications. For models trained in other countries, the applicant could move directly to the testing phase for approval or, in the event that the model has received approval from the regulatory authority of another country with a proven track record of high-quality decisions, would receive immediate approval subject to review by the domestic regulatory authority.

Mechanism

This policy provides direct monitoring and assessment of the most intelligent models by requiring them to go through a clearly defined process prior to being trained. We propose two criteria for the trigger of whether an AI model would need to apply for a TL to focus only on the most intelligent and therefore riskiest models: whether the model will exceed (1) pre-defined compute thresholds; or (2) a benchmark for general human capability.

As one of the triads of AI[37], and perhaps the most reliable proxy for a model's intelligence[38], compute provides a critical point for regulatory control. A further advantage of governing compute specifically is that few companies can afford the

[37] See this report for more information.
[38] This paper provides additional detail on this claim.

computational resources necessary to train frontier AI models. Finally, regulating compute also enables the broader AI supply chain to help regulate frontier development, for instance through the creation of on-chip mechanisms to monitor processor use.

A secondary proxy for the level of intelligence of a model is whether it can reliably achieve the performance levels of human remote workers when asked to carry out remote tasks. In order to assess whether a model displays concerning capabilities, the regulator must be able to establish whether its capabilities exceed a relevant threshold. We propose a system for assessing potential model performance, based on general work activities based on those defined in the O*Net classification system. This index would be based on performance in ten general tasks that can be performed remotely either by human workers or an automated system.

Implementation and enforcement

Given the exponential growth of AI, and the likelihood this growth will continue, agencies should be given maximum flexibility to ensure they can adequately assess models that pose the greatest risks and should apply for a TL. While the executives of these agencies would be appointed by and accountable to political leaders, and the specific governance of an AI regulator would need to be determined by each country, they should retain operational independence and have a minimum level of funding enshrined in law.

National AI regulators should set thresholds on compute to ensure proper oversight of frontier models that pose the greatest risk. These would be models where it is reasonably possible that training could lead to the development of dangerous capabilities that could either directly cause harm or result in the model escaping the developer's control. All such frontier models would automatically require a TL for its training run, and would require a separate application license prior to deployment, whether in commercial applications or otherwise.

The relevant national AI regulator would have the authority to set and adjust these thresholds, with specific governance structures around these decisions varying from country to country. Once an international agreement defines global thresholds for permissible development, national regulators would transpose international guidance into their own domestic thresholds. Countries could also decide on a more restrictive regime with tighter thresholds than the international regime if desired.

In addition, even if a model falls below the pre-defined compute threshold but expects their model to exceed an established benchmark for general human capability then it should also be required to apply for a TL. To implement this benchmark, the regulator would need to devise a battery of tests for each specific

task and establish a human performance benchmark by deploying the test to workers across different professions and levels of qualification. Once a benchmark was established, these tests would be administered to automated systems; if the system being tested performed at or above a predetermined percentile of the human benchmark (e.g., 90th percentile), it would be determined to be proficient at the relevant task.

This general capabilities index would then be constructed from these tasks to produce a final score - if automated systems achieved general intelligence-equivalent performance in a predetermined share of these tasks, it would clear the threshold for general capability and be banned.

A potential set of general tasks to be cleared could be as follows:

- **Analyzing and Processing Data and Information**
- **Communication and Collaboration (Internal)**
- **Project Management and Resource Coordination**
- **Developing and Implementing Strategies**
    - Fleshing out plans for complex real-world events for business operations and governmental activities.
- **Building and Maintaining Professional Relationships (External)**
- **Interpreting and Presenting Information for Various Audiences**
- **Content Creation**
    - Produce effective copy, images, videos, and other content to disseminate information, promote products and services, explain complex issues.
- **Training and Skill Development**
    - Non-project management and non-content feedback people management. Emotional guidance and coaching. Helping the other party reflect on past actions and teach new approaches and techniques.
- **Customer Relationship Management**
- **Domain-Specific Novel Problem Solving**

During the implementation phase, the regulator may decide to improve or expand on these tasks depending on how effectively they track model capabilities, with tests potentially requiring constant update and improvement.

As part of applying and receiving a TL, a developer would need to meet certain compliance requirements. Each jurisdiction will need to determine the appropriate number and type of any such requirements but at a minimum they should include the following:

- **Compliance requirement: companies applying for a TL would be required to submit their strategies for AI risk mitigation to the regulator as a pre-condition.** While these licenses would be specific to the model or application being developed, the AI risk mitigation strategies would refer to the applicants and their own risk management processes. That is to say: in order to apply for a license, the applicant must have had a relevant AI risk mitigation strategy approved by the regulator beforehand. This would also apply for requests to develop applications based on frontier AI models that increased model capabilities as defined by the regulator.
- **Compliance requirement: developers must not 'Open Source' or publicly release any part of the code or model weights.** This licensing regime seeks to drive and incentivize a safety-driven approach to model development. Releasing a model's code publicly for viewing, adaptation, or use undermines this as it would enable the model to be significantly altered by unregulated actors post-hoc. Therefore, any new model or application that is captured by the licensing regime must not be open sourced.

  Instead, external entities will be able to get meaningful access via API, which developers will be required to keep while the model meets the relevant threshold for frontier models. Failure to comply with this should result in severe penalties, including but not limited to: the model being instantly shut down and the developer having their license removed, fines for the developer, and criminal action taken against those involved in releasing the model publicly and found to be using the code in any other application.
- **Compliance requirement: developers must have mechanisms to shutdown their model and application if required temporarily or permanently.** AI is still an immature field; practitioners often report that they do not fully know how relatively-modest changes to architecture or algorithms will impact the capabilities or risks of a model. Accordingly, the R&D and deployment processes must be treated as inherently less certain than, for example, traditional mechanical engineering, and as having some risks of generating significant disaster.

  It is not guaranteed that we will have any observable warning signs before an R&D effort goes catastrophically wrong. However, right now humanity does not have processes to systematically detect warning signs, nor do we have systematic processes to investigate them, take corrective action, and learn from the issue and disseminate corrective fixes broadly.

  Therefore, in order to have a license for training and deploying frontier models, developers must document and prove to the regulator that they have

clear and stress-tested measures in place for how to shutdown a model. As with failure to comply with the license obligations, failure to perform a required shutdown, or negligent failure to maintain and regularly test shutdown capabilities, would result in the revocation of their frontier AI license.

Scope

What this policy affects:
The licensing regime should focus only on the most capable and general AI systems. As noted, managing the extent of AI models' general intelligence is a key element of this and fundamentally the implementation of a TL seeks to drive and incentivize a safety-driven approach to frontier AI model development by including specific requirements and a pre-defined procedure for assessing models.

What this policy does not affect:
Companies developing models and applications below the relevant compute and intelligence thresholds would not require licenses to operate and develop these products and services. However, companies would be expected to comply with the relevant regulatory limits, under penalty of severe legal repercussions in the event that thresholds are exceeded and companies operate beyond these thresholds without a license.

To note, the mere existence of shutdown mechanisms for models receiving a TL is not a panacea for AI risks, either in terms of loss-of-control or misuse. An out-of-control AI or a malicious user may be able to evade detection. Shutdown mechanisms therefore go hand-in-hand with strong monitoring mechanisms.

5.2 Compute license (CL)

Objective

➔ Ensure that data centers and cloud computing services above a certain compute threshold operate under a regulatory license, enabling authorities to monitor, restrict, and, if necessary, shut down the development of potentially dangerous AI systems.

Overview

The operation of data centers and provision of cloud computing services above a predetermined threshold of compute should be subject to the issuance of a license by the relevant national regulatory authority. Possessing a license should be a precondition to being able to operate and provide services to companies in that jurisdiction.

This will enable regulators to restrict the development of potentially dangerous AI models by identifying what compute clusters exist within a given jurisdiction, monitoring and enforcing restrictions on AI development related to amounts of compute for training or inference, and ensuring the ability to promptly shut down dangerous AI systems or strands of dangerous research.

Mechanism

Cloud computing services are integral to nearly all advanced artificial intelligence development and applications, from training to inference. Through the identification of relevant clusters and by placing meaningful constraints on their capacity, regulators can deploy effective brakes on the development of models and limit access to applications displaying concerning capabilities.

The operation of large-scale data centers is relatively easy to observe and monitor, given their large land requirements detectable via the planning system, their physical footprint making them often observable via satellite, and their large energy consumption. Their fixed location and large footprint makes them a natural chokepoint for regulators to monitor and intervene on, as well as a natural focus for mutual verification under international agreements.

By introducing a licensing regime focused on data centers above a specific threshold, the regulation can target the most impactful operations, ensuring appropriate mitigations can be deployed where relevant.

Implementation and enforcement

The proposal introduces a licensing requirement for any company operating data centers with a total compute capacity of 10^17 FLOP/s. This regime will ensure that larger, more resource-intensive facilities are subject to oversight and must meet relevant regulatory requirements.

Each jurisdiction will need to determine the number and nature of the requirements on compute providers to successfully be granted a CL, however, at a minimum the following requirements should be implemented:

- **Compliance requirement: compute providers must implement 'Know Your Customer (KYC) Rules'[39].** Companies must adhere to KYC regulations[40], which require them to verify the verifying client identities, tracking the use of compute resources, and reporting any high-risk entities to the government. This is intended to close existing gaps in export controls, prevent misuse of advanced AI technologies, and support responsible AI development by enabling more precise and targeted regulatory interventions.
- **Compliance requirement: compute providers must have adequate hardware tracking capabilities.** Companies will be required to track the physical hardware used in their data centers. While this may eventually involve the use of secure GPUs with serial numbers and physical tracking capabilities, aligning with relevant export controls, that technology is not yet widely available. An interim requirement[41] could be implemented, where companies would use physical GPS trackers on their existing hardware to comply with tracking and security standards.
- **Compliance requirement: compute providers must implement shutdown mechanisms.** In tandem with the shutdown measures highlighted in the implementation of TLs, compute providers must be clearly identified through redundant reporting chains to regulators – both by the frontier AI developers themselves, and through a KYC-like reporting process by compute providers and other supply chain participants. This would enable randomized spot checks by auditors to confirm if frontier AI companies have properly coordinated with their supply chain and counterparties and arranged for shutdown procedures to be implemented. Therefore, in the case of an emergency a compute provider and/or an AI developer can be called upon to shutdown the model. In addition, this would strongly incentivise frontier AI companies to only use the compute providers with the most rigorous safety protocols.

[39] This is similar to what has been proposed by some companies.
[40] See this for a more detailed proposal.
[41] See this proposal for more detail.

It is likely that through the introduction of this CL, a change in incentives will mean new technologies will emerge over time that will assist the compute supply chain in being able to control the use of their resources and help with the enforcement of license requirements. For instance, in the future, the national AI regulator could make it a requirement that in order to receive a license, the AI developer must use hardware providers that have Hardware-Enabled Governance Mechanisms (HEMs) so that they can remotely deactivate chips if they are either ordered to do so by the national regulator.

5.3 Application license (AL)

Objective

➔ Ensure that any new application which seeks to enhance the capabilities of a model approved with a TL is adequately assessed for any additional risks it may present prior to its deployment.

Overview

Any new use of an AI model approved through the TL process would need to seek approval for that new use. This is to ensure that any additional capabilities the new use creates are in keeping with the original approval of the TL and that restrictions, such as prohibited behaviors like self-replication, are not developed on top of pre-approved models. This would include connecting to an AI model through an API for it to run some or all of your product, or undertaking additional fine-tuning or research on said model.

Depending on the extent of the modifications to the base model or the exact proposed use, the applicant would be required to demonstrate the capabilities its proposed application would have and set out any additional relevant safety features and protocols that may be needed. If the regulator is satisfied that there was no risk to deployment, it would authorize the requested use. Any applications that do not change or modify the base model's capabilities, and do not result in structural manipulations like using it to train a smaller model or creating multimodal capabilities, would receive an automatic authorisation upon submission.

Mechanism

This policy ensures that using a model, approved through the above TL process, in a new application - whether through an API or any other method - such as a commercial or non-commercial product, service, suite of products/services, or research project, requires a license from the national AI regulator when making notable changes to the capabilities of that model that potentially increase its risk. This allows the AI regulator the opportunity to assess any new concerning

capabilities of the model and ensure adequate measures are taken to avoid any increased safety risks.

Implementation and enforcement

Applications based on models that had received a TL would be required to submit notification to the regulator. It would be the duty of the applicant to confirm whether their application is designed to increase the models capabilities or not. An automatic AL would be granted to applicants but the national AI regulator would be able to identify any concerning applications and take further investigations or enforcement action if necessary. This ensures a streamlined process for deploying new applications while maintaining regulatory awareness and oversight of the use of advanced AI systems.

Specifically, anyone seeking an AL should confirm their application will not draw on further compute resources for training such as using a TL model to train a smaller model, and that the application will not exceed the benchmark for human capabilities defined by the TL. This benchmark serves as a clear, measurable threshold for an acceptable application.

To maintain regulatory control, applications could be shut down on short notice through a shutdown of the underlying model or the relevant compute cluster. This mechanism provides the regulator with the ability to quickly intervene if necessary, balancing innovation with potential risks.

Scope

What this policy affects:
The policy affects any new application - whether through an API or any other method - such as a commercial or non-commercial product, service, suite of products/services, or research project that is based on a model trained on compute that exceeds the thresholds defined in the training license section.

What this policy does not affect:
This license does not affect applications below the specified threshold. While mandatory registration of these applications with the regulator would not be necessary, they would still be required to comply with the relevant limitations on capabilities and other prohibitions.

5.4 Monitoring and Enforcement

Objective

➔ Ensure:
- That the compliance of licensed AI models and uses to ensure the requirements are being upheld;
- That adjustments are made to the licensing requirements based on the evolving landscape of AI research and development.

Overview

To create a sustainable licensing system, any national AI regulator must have adequate capabilities and capacity to monitor ongoing AI research and development, while also having suitable enforcement powers to catch bad actors trying to circumvent the requirements.

Fundamentally, the national regulators and international system must have powers to review and adapt licensing requirements - through their power to lower compute thresholds or add new behaviors that should be prohibited - to fit with the latest AI research and development. To inform this, the national AI regulators must have significant capacity to monitor developments in algorithms and data used.

When it comes to the enforcement of licenses, severe penalties should be levied against developers who seek to build models above a compute threshold or the defined intelligence benchmark without a license to do so, and those developers who have a license but fail to comply with the above requirements.

To ensure that AI developers continue to have adequate measures in place, national regulators should undertake frequent testing of the procedures that AI developers would employ to respond to dangers and safety incidents. In addition, the national regulators must work with compute and hardware providers to frontier companies to withdraw their services in the event that they detect illicit activity. It may also be necessary to conduct mock training runs to test compute providers' ability to monitor the usage of their resources. Among other abilities, this could include their:

- Capacity to shut-off access to compute once a training run exceeds permitted thresholds;
- Ability to detect if a training run is simultaneously using other data centers;
- Ability to check if model weights are at zero at the beginning of a training run.

Mechanism

By ensuring the national AI regulator has suitable capabilities to monitor AI research and development, and enforce the licensing regime, they will be able to maximize the chances that any AI development in their country complies with the requirements set out and that those requirements stay up to date and suitable for the risks that we face.

There will always remain a slight risk that unlicensed developers make breakthroughs that circumvent the spirit of these regulations. It will be for the national regulators, and then the institution set up in Phase 1, to balance the risks of such breakthroughs with the cost of stifling innovation.

Implementation and enforcement

The country responsible for the creation of the national AI regulator must ensure it is created with adequate independence from political decision making and sufficient long-term funding that it can undertake its duties of ensuring advanced AI models are safe.

To ensure continued compliance, AI developers that received a TL or AL, or a computer provider who received a CL should be required to submit reports on safety procedures annually. A breach in the licensing requirements would need to face significant civil, and potentially criminal, action given the severity of the risks that it could pose. Below is a list of example enforcement powers that could be granted to the national regulator to help them fulfill their duties:

- Immediately shutdown the ongoing R&D process (e.g., training runs, fine tuning processes) of an AI developer, and wait for a detailed risk and root-cause assessment before restarting;
- The same as above, but for all similar projects across other companies and organizations developing AI;
- All of the above, but also terminate the project permanently;
- All of the above, but also terminate the project and all similar projects permanently in the company, and audit other companies and organizations to terminate similar projects due to similar risks;
- All of the above, but also fire the team that conducted the project due to a breach in protocol;
- All of the above, but also revoke the ability of the company to ever receive a future training or application license;
- All of the above, but also prosecute members of the organization or company involved in breach of regulations;

- In the most egregious cases, all of the above plus order a full shutdown of the entire company and sale of assets, via nationalization and auction or forced acquisition coupled with the wind down of all AI relevant operations.

Analogous powers should be provided to enforce KYC and similar requirements against compute providers. It is crucial that regulators should encourage true self-reporting of unexpected results, and provide some leniency when organizations do so proactively, swiftly, and collaboratively.

For instance, if a technique that enables recursive self-improvement is accidentally found in one specific company, and the company raises the issue to the authorities proactively, swiftly, and collaboratively, this should lead to rapid termination of the dangerous project in the company as well as rapid deployment of national resources to terminate similar projects elsewhere. This is the only robust way to avoid similar capability "leaks" from happening elsewhere, even if discovered initially in only one location.

Additionally, regulators should proactively create a mechanism for companies to share "near-miss" reporting, analogous to the US FAA system[42], such that they can proactively share insights about the ways in which accidents almost occurred but were avoided due to redundant measures and/or sheer luck, to inform the evolution of industry standards and regulatory efforts.

[42] See this for more details.

6. An International Treaty Establishing Common Redlines on AI Development

Objective

- ➔ Establish international red lines on AI development via a treaty;
- ➔ Facilitate collaboration on AI policy internationally with a view towards building a more comprehensive and stable international AI governance framework.

This policy fulfills the conditions of limiting the general intelligence of AI systems, no AIs capable of breaking out of their environment, no AIs improving AIs, and no unbounded AIs.

Overview

Alongside implementing the above measures nationally, countries should agree to them through an international treaty that creates a common regulatory framework across all signatory countries.

These measures are the ones described in the rest of Phase 0.

- Create an international compute threshold system, designed to keep AI capabilities within estimated safe bounds.
- Prohibit the development of superintelligent AI.
- Prohibit unauthorized self-replication and the intentional development of systems capable of self-replicating
- Prohibit unauthorized recursive self-improvement and the intentional initiation of recursive self-improvement activities.
- Require states to establish regulators and implement licensing regimes.

In addition to internationalizing the other measures of Phase 0, the Treaty should include a provision to prohibit the use of AI models developed within non-signatory states. This is to incentivize participation in the Treaty, to prevent actors within the signatory states from circumventing the Treaty, and to simplify monitoring and enforcement.

Rationale

While countries can unilaterally implement the proposed measures in Phase 0, in doing so they would not have guarantees from other countries that they would do the same. Individual countries are currently incentivised to avoid implementing

regulatory frameworks out of fear that other countries would be able gain a competitive advantage by implementing more lenient regulatory regimes.

These competitive dynamics may limit the potential for unilateral action, and therefore it is necessary for redlines to be agreed and committed to internationally. An international framework could avoid competitive pressures pushing regulatory standards to unacceptably low levels in a race to the bottom.

Implementation and enforcement

Countries should sign and ratify a treaty that both internationalizes the prohibitions of Phase 0, and establishes a compute Multi-Threshold System.

This treaty should then be enforced via the passage of national legislation.

This treaty will establish a Multi-Threshold System to determine the acceptable levels of compute. This will serve to harmonize the compute thresholds established by national licensing within an international treaty framework. Here is how the system will function.

Multi-Threshold System

Under the auspices of an international treaty, the compute thresholds established via the national licensing regime of Phase 0 should be internationally harmonized.

In doing so, an internationally upheld three limit system should be established, consisting of lower, middle, and upper limits. The lower level will be broadly permitted; the middle level, only by licensed entities; the upper level, only by an international institution with broad support across the international community, including the US and China, which we will label GUARD.

With these thresholds we aim to target:

- The capabilities of models trained, using total FLOP training compute as a proxy.
- The speed at which models are trained, using the performance of computing clusters in FLOP/second.

We can target capabilities in order to keep models within estimated safe bounds. We can also target the speed of training to limit the breakout time[43] to attain dangerous capabilities for legal computing clusters conducting an illegal training run, providing time for authorities to intervene. This will be achieved by targeting the total

[43] https://en.wikipedia.org/wiki/Nuclear_proliferation#Breakout_capability

throughput (as measured in FLOP/s - floating point operations per second) that a compute cluster can achieve in training.

These thresholds should be lowered as necessary, to compensate for more efficient utilization of compute (see below). This should be done by an international institution with broad support across the international community, which we will call the International AI Safety Commission (IASC). The upper threshold may be raised under certain conditions defined by a comprehensive AI treaty.

Threshold Limit Level	For Training Models - Total FLOP and who can train at that limit	Computing Clusters and Compute speeds	Minimum time to breach next compute training limit in an illegal training run, given a legal computing cluster[44]
Upper Limit (GUARD)	No more than 10^27 FLOP can be used for training (this threshold can be lowered by IASC) Only GUARD can train models up to this limit Nobody can train models above the upper limit	The only computing clusters permitted are within GUARD Each cluster can have a theoretical maximum computing power of up to 10^21 FLOP/s	N/A
Middle Limit (Licensing)	No more than 10^25 FLOP can be used for training National developers with a license can train up to this limit	No computing cluster with more than 10^19 FLOP/s permitted Licensed organizations can use clusters above the lower limit but not above the middle limit	3.2 years to breach Upper Limit
Lower Limit (Unlicensed)	No more than 10^23 FLOP can be used for training (unless a license is received) Anyone can train models below this limit, without a license	No computing cluster with more than 10^17 FLOP/s permitted	320 years to Upper Limit 3.2 years to Middle Limit

Note: In each limit regime, the largest permitted legal training runs could be run as quickly as within 12 days. For more information, see annex 2.

This compute threshold system should reflect the latest evidence to keep model capabilities within estimated safe bounds. The compute differences between the thresholds are designed to limit the breakout time of dangerous capabilities

[44] We can use the relationship: Cumulative training compute [FLOP] = Computing power [FLOP/s] * Time [s]. By controlling the amount of computing power that models can be trained with, we can manage the minimum amount of time that it takes to train a model with a particular amount of computation. Our aim in doing this is to control breakout times for licensed or unlicensed entities engaged in illegal training runs to develop models with potentially dangerous capabilities – providing time for authorities and other relevant parties to intervene on such a training run.

emerging through an illegal training run, thus providing time for authorities to intervene. This will be achieved by targeting the total throughput (as measured in FLOP/s - floating point operations per second) that a compute cluster can have in training.

Any AI system that passes a general intelligence benchmarking test is considered to be equivalent to having breached the Upper Compute Limit, and is thus also prohibited.

Phase 1: Stability

Once Phase 0 is implemented successfully, in principle and with all countries committed in good faith we will have measures in place that provide defense in depth to **prevent the development of artificial superintelligence for the next 20 years.**

With safety measures in place and the cautious prospect of two decades to mount our response, the next challenge arises from the potential instability of this system. While universal compliance with Phase 0 measures would be ideal, it is unrealistic to expect perfect adherence.

Systems naturally decay and fall apart unless they are actively maintained. Moreover, individually minor attempts to circumvent the system can compound over time, potentially undermining the entire framework.

We should anticipate various actors, including individuals, corporations, and governments, to put pressure on the system of Phase 0 measures by either trying to circumvent them, interpreting them in a more relaxed fashion, or otherwise launching projects that might violate some of the measures. Over time, these individually small pressures will add up and test the resilience of the system.

To maintain safety measures for the required two decades and beyond, it is necessary to establish institutions and incentives that ensure the system remains stable.

Therefore, the Goal of Phase 1 is to Ensure Stability: Build an International AI Oversight System that Does Not Collapse Over Time.

Conditions

To achieve Stability, certain conditions must be met.

1. Non-proliferation
2. International structure
3. Credible and verifiable mutual guarantees
4. Benefits from cooperation

Non-proliferation

This condition is necessary due to the fundamental issue of repeated risk problems inherent in proliferation. Even if the probability of a catastrophic event from any single AI development effort is low, the aggregate risk becomes substantially higher as the number of independent actors developing advanced AI increases. Each new party engaging in AI development introduces another chance for accidents, misuse, or unintended consequences. This multiplicative effect on risk is particularly concerning given the potentially extinction-level nature of advanced AI mishaps. By limiting proliferation, we dramatically reduce the number of opportunities for something to go wrong, thereby keeping the aggregate risk at a more manageable level. Non-proliferation is thus crucial not just for geopolitical stability, but as a fundamental risk mitigation strategy in the face of technologies with low-probability, high-impact failure modes.

International structure

The development of advanced AI science and technology must be an international endeavor to succeed. Unilateral development by a single country could endanger global security and trigger reactive development or intervention from other nations. The only stable equilibrium is one where a coalition of countries jointly develops the technology with mutual guarantees.

Credible and verifiable mutual guarantees

For actors to work towards this goal in a stable and durable manner, all key parties must be bound by certain conditions and able to verify others' compliance. Defections must be credibly and preemptively discouraged, with all actors precommitting to jointly preventing and punishing non-compliance. Systems must be in place to verify compliance and quickly identify defections, whether accidental or intentional.

Benefits from cooperation

Alongside credible deterrence, the stability of the system requires that participation be beneficial for all parties. While the **primary benefit is the continued survival of the human species,** the system should provide **additional incentives to discourage defection and encourage participation.** These benefits will act as initial incentives to expand the coalition of actors establishing this system.

Summary of Phase 1 Interventions

Goal: Stability: Build an International System that Does Not Collapse Over Time.

Policy	Condition it fulfills	High-level summary
1. International AI Safety Commission (IASC)	• Non-proliferation • International structure • Credible and verifiable mutual guarantees	• Central rule-setting body for AI Treaty members • Manages overall compute thresholds to account for algorithmic improvements • An institution responsible for monitoring treaty compliance and identifying those attempting to circumvent the treaty • Promotes and facilitates scientific cooperation between countries • Oversees GUARD
2. Global Unit for AI Research and Development (GUARD)	• Non-proliferation • International structure • Credible and verifiable mutual guarantees • Benefits from cooperation	• Central, multilateral research lab that conducts research into AI Safety and higher-risk AI applications, within the framework established by IASC • Researches and develops the frontier of AI in line with the parameters set by IASC and provides access to benefits of frontier models • Sole authorized organization for the highest-risk, frontier AI development
3. International AI Tribunal (IAT)	• International structure • Credible and verifiable mutual guarantees	• Independent judicial arm for the AI Treaty to resolve disputes resolving conflicts, breaches, and differing interpretations on issues relating to the application of and compliance with the AI Treaty • Defuses tensions on AI risks that could result in diplomatic tensions between nations or, in the worst case, armed conflict

1. International AI Safety Commission (IASC)

Objective

➔ Establish a commission to set the rules governing global AI development and oversee GUARD.

This policy fulfills the condition of **non-proliferation, international structure, credible and verifiable mutual guarantees.**

Overview

Through the signing of the AI treaty a new international authority should be created to monitor compliance with the treaty, promote AI safety research, and facilitate cooperation between signatories. This institution, which we call IASC, is necessary for providing oversight and ensuring that AI research remains under control. IASC and its employees will have similar diplomatic protections and status to the International Atomic Energy Agency (IAEA). This institution should be the central rule setting body for AI development, with a number of powers and responsibilities.

The core roles of IASC is **providing oversight for GUARD and lowering the globally applied compute thresholds in the Multi-Threshold System over time to account for algorithmic improvements**, in order to hold AI capabilities at estimated safe levels.

IASC will **monitor AI research and development, and undertake assessments on the risk of AI advancements.**

In addition, IASC will act as the secretariat and depositary to the treaty, and will have the jurisdiction to **monitor treaty compliance.**

This will include **conducting inspections and audits of licensed facilities** under the jurisdiction of signatories, **analyzing data** it collects via its monitoring systems, as well as analyzing data provided to it by third parties (e.g., nation-states' intelligence agencies).

In order to ensure good governance of IASC, a representative chamber known as the **Council** should be established, along with an **Executive Board**, and a position of a General Secretary of IASC.

Rationale

In order to have a comprehensive international regulatory framework that ensures continued AI development is conducted in a manner that does not pose unacceptably high risks to humanity, it is necessary to reach international agreement both on rules, but also how they are enforced.

Furthermore, given that such an endeavor will require the cooperation of competing powers, it is necessary to establish clear trust-building mechanisms in terms of inspection, monitoring, and verification procedures.

In achieving this, the risk of defection is mitigated by both reducing the incentives to defect, and mitigating the impact of doing so. Incentives to defect are reduced by creating an expectation that activities in breach of the treaty will be detected. The impact of defection is mitigated by early detection of activities in breach of the treaty, allowing for a response that is able to deter or prevent continued breaches.

IASC Organizational Structure

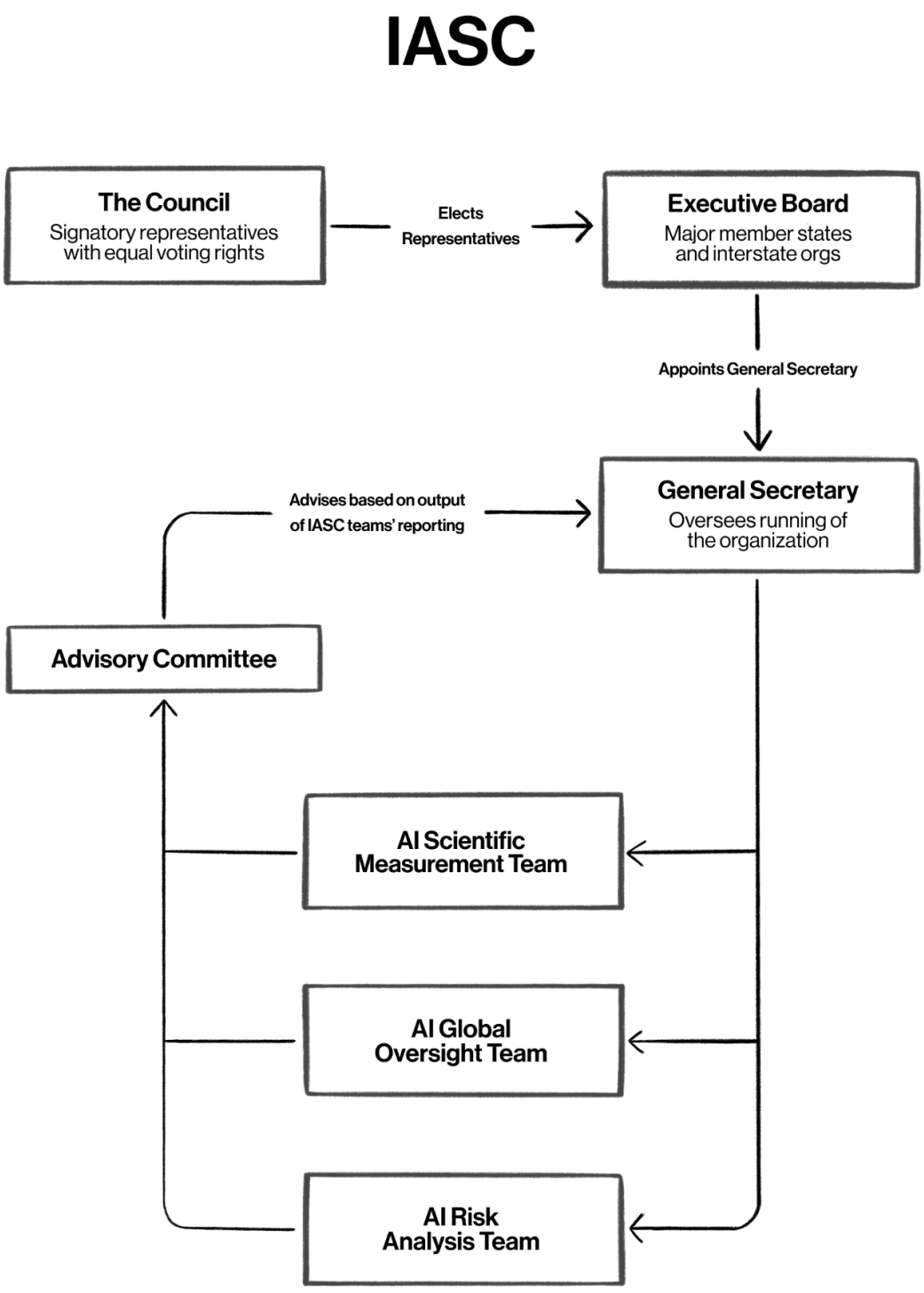
In order to ensure good governance of IASC, a representative chamber known as the Council should be established, along with an Executive Board, and a position of a General Secretary of IASC.

The Council would consist of a representative from each member state of the treaty framework, and will meet at least once per year to agree countries' contributions and budget. To ensure sufficient continued operation and capacity of IASC, each signatory country of the AI treaty must contribute an agreed sum annually to IASC.

The Executive Board, analogous to the UN Security Council, would consist of representatives of major member states and supranational organizations, which would all be permanent members with vetoes on decisions taken by the Executive Board, as well as non-permanent representatives elected by a two-thirds majority of the Council.

The General Secretary of IASC should be appointed by a three-fourths majority of the Executive Board, and would have a number of important duties, including most crucially deciding on lowering the compute thresholds. The General Secretary formally makes the decision to lower the compute limits established in the Multi-Threshold System, on the advice of the Advisory Committee, following a report of the AI Scientific Measurement Team.

For a more detailed look at the organizational structure of IASC, outlining its various departments and decision procedures, including the Advisory Committee and the AI Scientific Measurement team, see the annexes.



1.1 Multi-Threshold System

Essential to the functioning of a stable international regulatory system is the Multi-Threshold System established in Phase 0. In this system, AI models would only be permitted to be trained within certain compute limits, and with restrictions on the computing power of data centers used to train them.

Implementation and enforcement

One of the core roles of IASC is in lowering the compute thresholds established in the Multi-Threshold System over time to account for algorithmic improvements which mean that more capable, and more dangerous, models can be developed with a fixed amount of training compute. The objective here is to map the compute threshold proxies onto fixed capabilities levels, in order to keep AI development within estimated safe bounds.

The General Secretary of IASC formally makes the decision to lower compute thresholds, on the advice of the Advisory Committee, and following a report of the AI Scientific Measurement Team.

Following such a decision by the General Secretary, GUARD and national regulators would be legally obliged to implement it and ensure that AI models are not trained in breach of the updated thresholds.

We propose initial thresholds on the next page.

Threshold Limit Level	For Training Models - Total FLOP and who can train at that limit	Computing Clusters and Compute speeds	Minimum time to breach next compute training limit in an illegal training run, given a legal computing cluster[45]
Upper Limit (GUARD)	No more than 10^27 FLOP can be used for training (this threshold can be lowered by IASC) Only GUARD can train models up to this limit Nobody can train models above the upper limit	The only computing clusters permitted are within GUARD Each cluster can have a theoretical maximum computing power of up to 10^21 FLOP/s	N/A
Middle Limit (Licensing)	No more than 10^25 FLOP can be used for training National developers with a license can train up to this limit	No computing cluster with more than 10^19 FLOP/s permitted Licensed organizations can use clusters above the lower limit but not above the middle limit	3.2 years to breach Upper Limit
Lower Limit (Unlicensed)	No more than 10^23 FLOP can be used for training (unless a license is received) Anyone can train models below this limit, without a license	No computing cluster with more than 10^17 FLOP/s permitted	320 years to Upper Limit 3.2 years to Middle Limit

Note: In each limit regime, the largest permitted legal training runs could be run as quickly as within 12 days. For more information, see annex 2.

[45] We can use the relationship: Cumulative training compute [FLOP] = Computing power [FLOP/s] * Time [s]. By controlling the amount of computing power that models can be trained with, we can manage the minimum amount of time that it takes to train a model with a particular amount of computation. Our aim in doing this is to control breakout times for licensed or unlicensed entities engaged in illegal training runs to develop models with potentially dangerous capabilities – providing time for authorities and other relevant parties to intervene on such a training run.

1.2 Framework for Information Collection and Intelligence Sharing

Within IASC there should be a clear framework for how the organization collects information from countries about their AI development and how intelligence is shared between countries and with IASC. We call this the **Framework for Information Collection and Intelligence Sharing**. Overall, our approach is inspired by the IAEA's approach to monitoring nuclear capabilities. However, this approach must be tailored since compute resources are nearly-ubiquitous in everyday life and computer chips must still be deployed broadly for other everyday purposes.

Our proposed framework has five parts:

1. IASC's development and execution of inspection and monitoring to proactively analyzed and assess global AI development;
2. IASC's development (both internally and in partnership with third parties) of verification and monitoring capabilities that can provide general monitoring of major concentrations of compute;
3. IASC's analysis and ongoing audit of global supply chains relating to AI development;
4. IASC's and GUARD's system for providing guidance to nation-states implementing the treaty (and their intelligence services) on things to proactively monitor;
5. The information-sharing mechanism whereby signatories to the AI Treaty (and their constituent regulators, AI safety/research institutes, law enforcement and security services, etc.) can share information on AI to IASC and with each other.

Implementation and enforcement

Inspection and Monitoring

First, IASC should develop an overall inspection and monitoring plan and process. In accordance with specific commitments in the AI Treaty, countries (and the companies, nonprofits, government institutions, etc. residing within them) should be required to submit to regular inspections by IASC staff. These inspections should generate reasonable confidence that the countries party to the treaty, and the entities within them, are abiding by each of the requirements of the treaty. Inspections may be conducted physically in-person (e.g., to verify that a given data center has or lacks advanced chips) or virtually (e.g., remote access to compute, storage, logs, etc.) depending on the requirements of that particular inspection.

As part of the initiation of the treaty, countries should be required to engage in one-time "displays" of their existing capabilities to confirm they are accurate. For

example, if the US government asserts that US Department of Energy supercomputers have a certain amount of compute capabilities based on having specific chips, they should be required to do a one-time demonstration of that facility.

Verification mechanisms

Second, IASC should begin long-term research and development efforts to identify longer-term needs to maintain and enhance their inspections program via verification mechanisms. At scale, this will require tamper-resistant verification mechanisms throughout the hardware and software stack. Mechanisms could be developed by IASC, but likely will be more robust if they are developed through a process that also incorporates outside input and testing, similar to the US NIST cipher competitions for general-public and government cryptographic use. To be practicable, these mechanisms would need to be capable of reporting signals of dangerous use of large amounts of compute without generally violating the underlying privacy of compute users. For instance, "reporting dashboard enablers" that help track exceptionally large amounts of compute usage by customers over a given threshold would meet this criteria, but backdoors into every processor would not.

Supply chain audits and controls

Third, IASC should be able to conduct supply-chain tracing and audit relevant export control, KYC, etc. processes to ensure that they are properly applied. These steps are necessary to ensure that even if a non-signatory or a treaty-breaking signatory state runs a hidden, air-gapped program it can be detected.

Detect and advise on signatures of risk

Fourth, IASC should provide the security services of signatories advice on what risks to watch out for, both in terms of technical signatures (e.g., particular patterns of network activity or cloud compute usage) and other indicators of concern (e.g., sharing information on non-state groups that are identified through inspections and monitoring as building potentially hazardous AI). These could enable multilateral efforts to address and mitigate AI risks through mechanisms such as sanctions or prosecutions.

Of course, intelligence tips from IASC pose their own risks, as they could also enhance signatory countries' ability to evade oversight and/or to grow their own capabilities, and will have to be carefully controlled through a disclosure process.
Finally, IASC should provide a framework for information-sharing between countries that are party to the treaty, so that they will be able to share intelligence and

information, and cooperate with IASC to identify, monitor, deter, and prevent activities by state or non-state actors prohibited by the treaty.

2. Global Unit for AI Research and Development (GUARD)

Objective

- ➔ Pool resources, expertise, and knowledge in a grand-scale collaborative AI safety research effort, with the primary goal of minimizing catastrophic and extinction AI risk.
- ➔ Mitigate racing dynamics, both between corporate AI developers and between nations, by only allowing one organization to work on the true frontier. The lab, subject to its own Upper Compute Limit on the models it can train, develops models in order to meet the priorities of each country that signs the treaty, and to be beneficial to humanity.
- ➔ Safely develop, explore, leverage, and provide benefits of AI to humanity, enabling the AI systems developed by GUARD to be accessed by 3rd parties for innovative new use cases in accordance with the multi-threshold system.

This policy fulfills the conditions of **non-proliferation, international structure, credible and verifiable mutual guarantees, benefits from cooperation.**

This policy supports development of safe AI research that **enables all of the underlying safety conditions** we are trying to achieve through its research, as well as providing supervision of the most-risky AI research such that it is less likely to violate those safety conditions.

Overview

Countries should collectively create an AI research institution, which we call here the Global Unit for AI Research and Development (GUARD). This institution should be governed by IASC to ensure that it properly prioritizes safety throughout its research. Any AI it develops should be bounded, controllable, and corrigible, even if that requires meaningful trade-offs in current or future capabilities.

Rationale

The system is needed to remove the incentives for countries to race to produce unsafe, artificial superintelligence. Without a collaborative, multilateral effort, countries may see AI as an all-or-nothing prize that goes to the first country to develop it, and cut every safety corner they can to build it as quickly as possible; countries that stand no real chance of developing AI in the near future might even engage in cross-domain deterrence and threaten AI-developing countries with acts

of war to attempt to deter its development. Through GUARD, we defuse those competitive tensions and their accompanying risks.

However, this poses a problem in turn: any multilateral system runs the risk of bad actors, including nation-states and powerful non-state actors, refusing to participate in the system. A bad actor that seeks to establish their own breakaway AI superintelligence program, either out of a misguided belief that they can benefit from doing so or from a belief that a nascent superintelligence is a "doomsday" threat that could be used to coerce other nations, may ultimately succeed given enough time, resources, and luck. By establishing GUARD, we reduce the incentives for bad actors to break away from the international system as they can benefit from GUARD's spinoff developments far sooner and with far lower risk than their breakaway program. In addition, rogue actors' will be less likely to succeed as the world's best AI talent will already be employed in a collaborative, multilateral environment instead of a breakaway black project.

Mechanism

Through this lab, Treaty signatories will be able to make progress on AI innovation safely, and engage in higher-risk research in a research community that draws from the best of existing safety research talent, and shares those insights instead of keeping them inside corporate silos. As this research bears fruit to create safe and beneficial AI systems, GUARD will provide access to them to Treaty members; this benefit will encourage countries to join the AI Treaty and GUARD system.

Implementation and enforcement

GUARD will pool resources, expertise, and knowledge in a grand-scale collaborative AI safety research effort, with the primary goal of minimizing catastrophic and extinction AI risk. This will centralize development of the most-advanced allowed AI systems within a single internationalized lab, as part of the internationally-set multi-threshold system (see 'AI Treaty' section), where the internationalized lab has its own higher limit than any other entity globally. The internationalized lab will meet or exceed the standards we would also propose for implementation at the national level's licensing regime and should develop models in order to meet the priorities of each signatory to the AI Treaty and to the benefit of humanity generally.

This access could be provided either as non-AI-model outputs (e.g., a dataset of medical insights to enable new drug discovery) or via verified output[46] API access to GUARD's AI models through a network of lab-operated computing clusters around

[46] Of course, any such system should be solely focused on safety-preservation and have appropriate mechanisms to ensure such monitoring could not be used to harm users for their free expression.

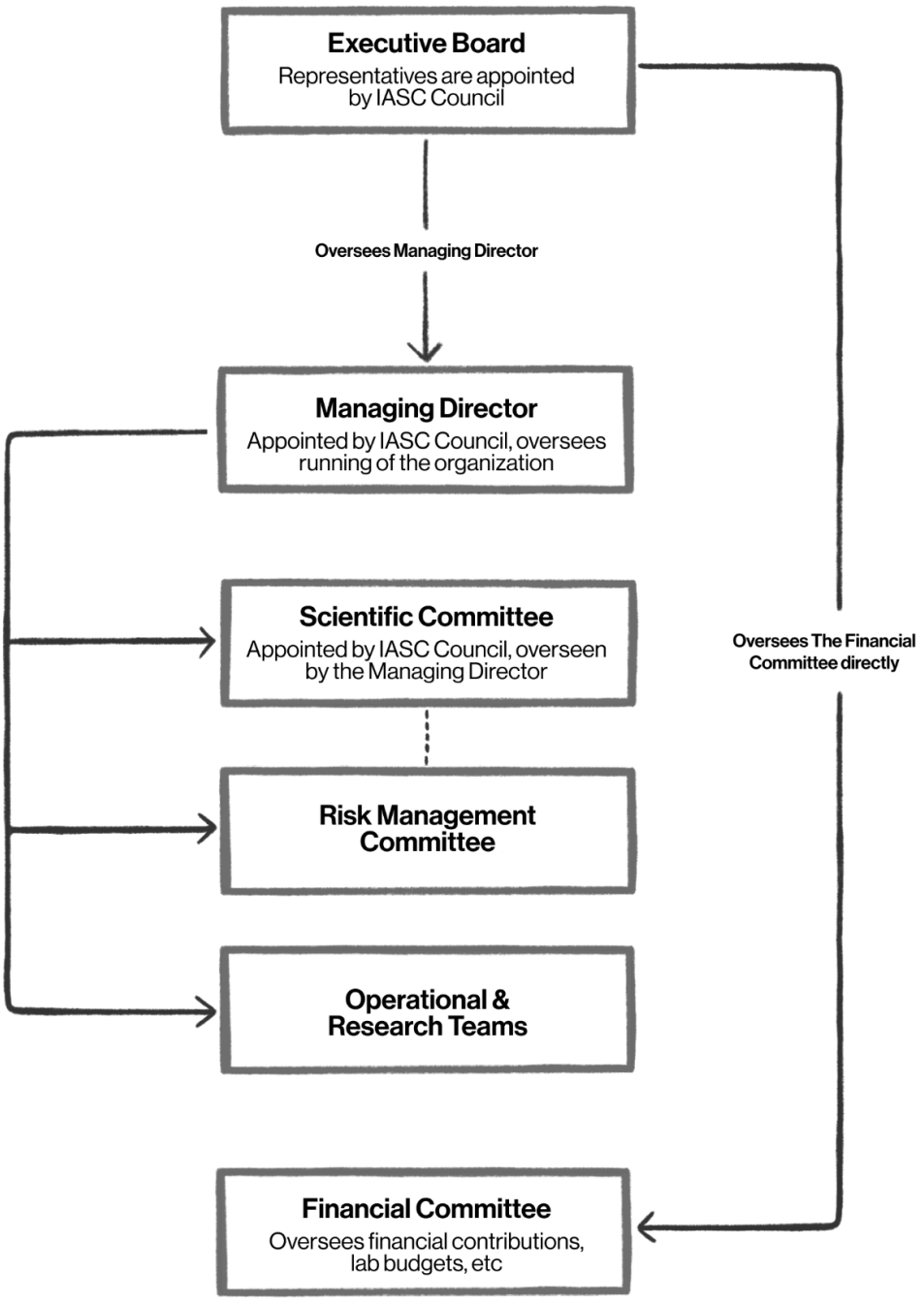
the world. In practice this will mean that GUARD should share the results of its work, breakthroughs, and best practices unless doing so would pose a hazard, but not release underlying models until the state of the art has advanced to be able to know that it is doing so in a controlled and safe fashion.

The Lab would be supervised by the International AI Safety Commision and should be proactively designed to incorporate lessons learned from existing research institutions; in particular, its institutional design should minimize the risk of internal institutional capture by researchers who willfully cut corners on safety.

The Lab should be run by an Executive Board, informed by expert Advisory Committees (to whom the Executive Board could delegate some day-to-day decisions). The GUARD Lab would be subject to oversight from IASC and in particular, IASC could veto the appointment, order the removal, or order the reassignment to a less-sensitive project of any senior official within GUARD (including the Executive Director; the chair of any Committee or other major research team; any immediate subordinate reporting directly to one of the previously mentioned persons). See Annex 1 for a proposed breakdown of the institutions' structure.

GUARD would be required to operate under very high security standards, comparable with those who work in other high-risk industries, such as aviation, virological research, nuclear technology research at national labs, or national security agencies. However, these standards will have to be tailored to the context of AI lab work; for example, some work might require operating in a network- and signal-isolated environment similar to an intelligence community Sensitive Compartmented Information Facility (SCIF), but other work might properly *require* ongoing internet connectivity (e.g., training a model to better forecast extreme weather events based on real-time weather data).

GUARD


Executive Board
Representatives are appointed by IASC Council
Oversees Managing Director
Managing Director
Appointed by IASC Council, oversees running of the organization
Scientific Committee
Appointed by IASC Council, overseen by the Managing Director
Risk Management Committee
Operational & Research Teams
Financial Committee
Oversees financial contributions, lab budgets, etc
Oversees The Financial Committee directly

3. International AI Tribunal (IAT)

Objective

➔ Create an independent judicial arm for IASC, with the sole purpose of resolving conflicts, breaches, and differing interpretations on issues relating to the application and compliance with the AI Treaty.

This policy fulfills the conditions of **international structure, credible and verifiable mutual guarantees.**

Overview

The International AI Tribunal (IAT) should serve as an independent judicial arm of IASC, with the sole purpose of resolving conflicts, breaches, and differing interpretations on issues relating to the application of and compliance with the AI Treaty.

The IAT will work to swiftly adjudicate disputes arising within the AI Treaty framework and interpret the treaty's provisions.

Rationale

As with many other international agreements, it is all-but-necessary to have an adjudicatory body to resolve disputes between parties to the treaty. Without such an adjudicatory body, the only recourse is nation-states using other means of diplomacy and conflict, which may be tangled up in their other interests (e.g., Country A won't sanction Country B because they are allies). Other frameworks use such bodies successfully, for example the WTO Dispute Settlement Body and the International Tribunal for the Law of the Sea. Unfortunately, disputes taken up by such bodies often take lengthy periods of time to resolve. The average timeframe for a dispute at the WTO is 10 months, at the ICJ it is 4 years, and for the ECJ it is 2 years[47].

There is an inherent risk of advanced AI development that breaches the provisions of the treaty, and a risk of disputes relating to treaty provisions which, in extreme cases, could have the potential to spiral into conflict between states. It is therefore necessary to construct a settlement body with legitimacy that can both fairly and correctly adjudicate disputes, and do so in a timely manner for cases that require it.

[47] https://www.wto.org/english/tratop_e/dispu_e/speech_agah_4mar10_e.htm

Mechanism

The IAT provides a mechanism for dispute resolution under the AI Treaty. The IAT is tailored to the specific needs of mitigating the most severe AI risks, and is designed to provide prompt dispute resolution.

Implementation and enforcement

The IAT should be established with a comprehensive organizational structure, in order to be able to effectively, correctly, and speedily, adjudicate disputes arising within the AI treaty framework.

At the core of the IAT is the Court, which would consist of 31 judges appointed by IASC Council to serve six-year renewable terms. The Court would make use of a chambers system, modeled off the European Court of Justice, where by default cases are heard by a panel of 5 judges. More significant cases could be heard by a grand chamber of 15. As with the ECJ, chambers could make use of Advocates General to obtain independent legal opinions.

We expect that cases will arise where a delayed judgment could be costly to humanity in terms of risk, and therefore a system is needed to prioritize certain cases and ensure timely processing.

For this reason, we also propose the establishment of a Risk Assessment Panel, to determine which cases must be prioritized, and a Rapid Response Panel, where cases of the highest priority can be referred to.

In addition, the Court should include an appellate body, where cases can be re-examined.

For a more detailed look at the organizational structure of the IAT, see the annexes.

Once a ruling has been issued, parties to the AI Treaty are expected to comply with the decision. If they fail to do so, the IAT must oversee the implementation of the ruling and can authorize the imposition of measures outlined in the AI Treaty framework, such as economic sanctions, similar to existing trade agreements.

Phase 2: Flourishing

Introduction

If humanity succeeds at implementing the previous 2 phases of The Plan, the world will be in a stable situation with regard to AI, where advanced AI research is regulated, dangerous AI proliferation is contained, and some of the riskiest research is only done by internationally coordinated organization(s) following strict safety protocols to not endanger human civilization.

The natural next step is then the development of Safe and Controllable Transformative AI, to benefit all of humanity. **Not superintelligence, nor AGI, but transformative AI.** AI that is developed not with more and more capabilities as an end in itself, but as a tool for humans and under human control to unlock prosperity and economic growth. **AIs as tools for humans to automate at scale, not AI as a successor species.**

Phase 1 includes the creation of the Global Unit for AI Research and Development (GUARD), a central multilateral lab which is the only organization authorized to pursue frontier AI research.

Yet GUARD cannot just continue with current dominant paradigms of machine learning research to achieve its goal: ensuring that AI research is done in a sensible and grounded way. This is because existing machine learning approaches focus on increasing capabilities without shedding any light on how AIs work, or how to control them. Therefore, it is crucial to determine which alternative AI development paths GUARD could take, while keeping humanity in control.

Thus, the Goal of Phase 2 is to Ensure Flourishing: build the science and technology for Safe and Controlled Transformative AI as a tool for human prosperity and growth.

Conditions For Safe Transformative AI

The development of Safe Transformative AI requires three necessary conditions. All three must be satisfied for humanity to ensure that it only builds controllable yet powerful AI systems, which can then be used for various civilization goals such as automating all intellectual and physical labor (see What Success Looks Like for more details on the use and challenges of such technology).

These three conditions are:

- Prediction of AI systems capabilities
- Specification of AI systems guarantees
- Enforcement of AI systems guarantees

Prediction of AI systems capabilities

The biggest obstacle to safe AI development with current ML technology is the inability to predict what an AI system can and cannot do. This is the case not only before pre-training or fine-tuning, but even after deployment of the AI system. In one example among many, Anthropic testers realized accidentally that their new model was able to recognize it was undergoing tests[48] and alter its behavior accordingly.

And despite extensive efforts to develop theories of Deep Learning[49], mechanistic interpretability[50], and evaluation frameworks[51], still nobody is able to predict what ML models can and cannot do.

Yet prediction is essential. In order to develop safe AI systems, it is critical that GUARD be able to predict what any AI system can do before building it, or at the minimum once it is built. Without this, there is no theoretical knowledge we can use to ensure that GUARD does not go too far in its research and builds AI systems that are too close to uncontrolled superintelligence.

Given this, a condition for developing Safe Transformative AI is advance our theoretical understanding of AI systems so we can model and predict the capabilities of any AI system that GUARD might build.

Specification of AI systems guarantees

The next step towards safe AI systems lies in figuring out exactly which properties they need to satisfy in order to be safe. This might include properties about controlling these systems, about them being legible to users and inspectors, or about them never proposing actions that are particularly unsafe.

Current ML research does not even try to do this, focusing instead on measures of efficiency, performance, and proxies such as “truthfulness.”[52] These measures are

[48] https://x.com/alexalbert__/status/1764722513014329620
[49] https://deeplearningtheory.com/
[50] https://arxiv.org/abs/2404.14082
[51] https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations
[52] https://arxiv.org/abs/2109.07958

also constantly being gamed[53] by machine learning systems, since they do not capture specific features of machine learning systems' properties, but merely statistical similarities in large amounts of low-quality data.

Given this, a condition for developing Safe Transformative AI is to specify which guarantees a safe AI system needs to uphold.

Enforcement of AI systems guarantees

Lastly, guarantees are only valuable if they are actually enforced. So safe AI development requires the ability to ensure that the guarantees specified in the previous conditions are actually followed by a given AI system.

This is not the case in current machine learning systems for two reasons.

First, as mentioned above, current AI developers are unable to predict how ML systems will behave, even after they have finished training. Thus even after the fact, current machine learning theory provides no way to verify that the AI system follows the specification.

And second, current training techniques in machine learning search exclusively for algorithms and AI systems that score high on a set of performance measures. We lack any suitable definitions or specification of control, legibility, or safety that can be used as goals of machine learning training processes. This means that ML systems are incentivized to disregard each and any of these properties if that helps them to perform better on their performance indicators or downstream tasks.

Given this, a condition for developing Safe Transformative AI is to enforce the guarantees that a safe AI system needs to uphold.

Recommendations For Safe Transformative AI

A detailed research agenda for satisfying the conditions of predicting AI system capabilities, and specification and enforcement of AI systems guarantees does not exist at this time. Such an endeavor also exceeds the scope of this document.

Yet, we can infer what the broad direction for tackling each of these conditions should be by looking at what is considered sensible and reasonable for other high-risk technologies.

[53] https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Science of Intelligence

Recall that the first condition is **Prediction of AI systems capabilities.** Unless GUARD can predict the capabilities of AI systems before building them, they have no hope of maintaining safety while exploring the potential benefits of AI.

This problem of prediction was tackled in the same way for many existing high-risk technologies such as civil engineering, aviation and nuclear power. After some initial groping in the dark and experimentation, pioneers in these fields built scientific fields and slowly learned to model and predict each of these domains: respectively structural engineering, aerodynamics, and nuclear physics.

There is no reason why AI systems should be different; thus the most direct way to satisfy the first condition is by developing a science around AI systems.

This begs the question of what this science should study. Since the goal of AI systems is to automate various aspects of intelligence, and since the extinction risks that this document is addressing focus on general intelligence, the right science for AI systems is **a science of intelligence.**

Taking inspiration from the historical examples, the first step to building such a science is to design ways to measure the underlying phenomena. In structural engineering, this came about in the mechanical testing of materials; in aviation, with the measurement of aerodynamic forces, notably in wind tunnels; in nuclear technology, with the measurement of radiation, for example with Geiger counters.

In each of these cases, the development of measurement methods was not just about building tools – it also required theoretical and conceptual innovation to figure out what to measure, and how to measure it, to get the right information, often indirectly.

Once intelligence can be sensibly measured, the data collected through these measurements will lead to a science of intelligence that can be used for predictive purposes. This will notably include a mechanistic model of intelligence: a decomposition of intelligence into components such that knowing which components are implemented in an AI system lets you predict its intelligence and capabilities in advance before even building it, or turning it on.

Such a model would extend GUARD's understanding of intelligence to the point where its members would be able to anticipate the intelligence of various AI systems before building them, and thus both steer away from too powerful design and aim for the least intelligent system that still accomplishes a task.

This would satisfy the first condition, **Prediction of AI systems capabilities.**

Specification Language For AI Systems

Turning to the second condition, **Specification of AI systems guarantees:** To ensure that AI systems are safe, the first step is to be able to write down "what we want" from these systems. This includes properties such as controllability, legibility, safety.

This goes beyond the fundamental science discussed in the previous recommendation: civil engineering needs to specify what counts as a structure such as a bridge "failing", and which failures are not acceptable; the same is true for aeronautics and plane failures, and nuclear technology with radiation leakage or uncontrolled chain reaction.

Yet AI systems have one advantage over these other high-risk technologies: they are primarily software based. This means that they can leverage advances that have been made in specifying software properties through formal specifications.

Still, there is currently no specification language that is sufficient for capturing the guarantees needed for AI systems. This is because these guarantees rely not only on what the AI system does, but also on how it interacts with other AIs and humans. Modeling humans and their interactions in formal logic is out of reach for current specification methods.

A specification language is not enough though: it is also essential to figure out which exact properties we need to express in this language. Since the sole purpose of the specification language is to allow the specification of these guarantees, formalizing these guarantees and designing the language will go hand in hand.

In the end, this effort will result in a formal specification language that can address any AI system behavior, including interaction with sub components, other AI systems, and humans. The guarantees that need to be upheld by safe AI systems will be written in this language, ensuring controllability, legibility, and safety.

This would satisfy the second condition, **Specification of AI systems guarantees.**

Safe-By-Design AI Systems

Last but not least, the last condition asks for **Enforcement of AI systems guarantee.** In the end, what matters is that GUARD builds safe AI systems, which requires

ensuring and enforcing the guarantees designed to make these systems controllable, legible, and safe.

Although it is possible to enforce these guarantees after building the AI systems, such an approach is insufficient, as comparisons with the standards already established for other high-risk technologies show.

Pointing to just one example, the UK's nuclear regulation[54] (EKP.1, p.37 of 2014 version) states that:

"The underpinning safety aim for any nuclear facility should be an inherently safe design, consistent with the operational purposes of the facility.

An 'inherently safe' design is one that avoids radiological hazards rather than controlling them. It prevents a specific harm occurring by using an approach, design or arrangement which ensures that the harm cannot happen, for example a criticality safe vessel."

GUARD should thus enforce the guarantees specified for Safe Transformative AI **by design.** Whereas modern ad-hoc safety efforts attempt to fix issues after the fact, playing a losing game of Whack-A-Mole, a responsible approach to Safe Transformative AI must bake in the guarantees in the architecture and the structure of the AI systems themselves.

And not only should the AI systems be safe by design, they should be safe by design against unanticipated and unpredicted issues and stresses. Other industries use a factor of safety[55] to make their systems more resilient against unforeseen incidents. GUARD needs an equivalent tool that can be applied to AI systems.

This means that at every step in the building of safe AI, the methods used must maintain these guarantees to preserve control, legibility. That way, most of the failures of safety and alignment will be prevented by design, and the remaining risks will be of a manageable, smaller number, making it more likely they can get ironed out through systematic testing.

Such safe-by-design methods exist for current specification languages[56] designed for normal software, but will need to be designed and checked for the more involved specification language necessary to satisfy the second condition.

---

[54] https://www.onr.org.uk/publications/regulatory-guidance/regulatory-assessment-and-permissioning/safety-assessment-principles-saps

[55] https://en.wikipedia.org/wiki/Factor_of_safety

[56] https://dl.acm.org/doi/10.1145/3591335.3591343

This would satisfy the third condition, **Enforcement of AI systems guarantees.**

The Path Forward

The best examples of improving the safety and reliability of designed systems, up to a point where human risk is minimized, come from fields where safety thinking and formal methods are applied. Fields such as aviation, space exploration, and nuclear energy.

The Conditions and the Recommendations for Phase 2 share this common thread. They steer Safe AI Research towards the successful and appropriate approaches of Safety Engineering and Formal Methods research, rather than the priorities of current machine learning research. This is the path for human science and engineering to master safe, controllable, transformative AI.

Some current projects in AI fit with the spirit of the conditions and recommendations above, and thus can provide inspiration: DARPA's Explainable AI Project[57], ARIA's Safeguarded AI[58], Conjecture's CoEm[59], and Guaranteed Safe AI research agendas[60] by Tegmark & Omohundro, Dalrymple, Bengio, Russell and more.[61]

[57] https://www.darpa.mil/program/explainable-artificial-intelligence
[58] https://www.aria.org.uk/programme-safeguarded-ai/
[59] https://www.conjecture.dev/cognitive-emulation
[60] https://www.provablysafe.ai/
[61] https://arxiv.org/pdf/2405.06624

What success looks like

If Phases 0, 1 and 2 all succeed and are fully implemented, humanity will be in a stable situation with international coordination and Safe Transformative AI: AI systems that can automate any intellectual and physical task, while still being under our control. Such civilization-altering technology will bring about mass-scale automation and through it unlock many options for the future of humanity. However, this technology will not bring with it the wisdom required to wield this newfound power well.

This section thus maps the resulting upsides and challenges that can already be anticipated, in order to start the conversation about how to handle them, and how to improve the wisdom of human civilization to a point where it can handle them reasonably.

What Safe Transformative AI Unlocks

The thrust of Safe Transformative AI's impact on human civilization is the possibility to automate all intellectual and physical labor. AI and robotics are much easier to mass create than humans, much easier to replace or break without moral issues, and much more efficient. They do not need to rest, have no emotions which get in the way of thinking, no need for narrative justifications for their tasks. This leads to a broad trend of acceleration and progress across the board.

First, all work that humanity wants to automate will be automated. Having humans involved in work rather than machines will be a political choice, rather than one dictated by necessity: this will no longer be bottlenecked by technology. This includes dangerous work (firefighting, nuclear waste disposal), unpleasant work (cleaning, garbage collection), boring repetitive work (data entry, writing many personalized emails). It might include literally any kind of work, but it does not have to. Such overall automatization of work will completely change the way society works, and what activities people participate in.

Automation of physical tasks will also unlock significant progress in manufacturing: increased efficiency, scale, utilization of resources. This will lead to both massive progress in fundamental manufacturing processes, including massive manufacturing at scale and vastly better materials, and an abundance of physical goods unlocked by this manufacturing progress. Automation will push these to the point where the main bottleneck becomes policy and regulation, rather than technological capabilities.

In general, scientific and technological progress will be accelerated through the automation and parallelism of all scientific and engineering intellectual tasks. This will yield benefits in fields as varied as medicine (developing new drugs and testing them much faster), energy (unlocking new forms of renewable energy), social sciences (designing much higher quality theories of economic, sociological, psychological processes).

Lastly, beyond simply automating and improving what humanity is already doing, Safe Transformative AI offers a path towards tackling problems that have been blocked by technology and resources constraints. For example, two of the most salient and currently discussed are aging and space exploration.

The effort to curtail and even reverse aging is a recurrent goal throughout human history, with the goal of reducing the senescence and pain that plagues humans as they age and forbids them to spend much time with their grandkids and other descendants. But it is blocked by our lack of understanding of the body and its aging processes. The scientific automation enabled by Transformative AI promises to shed light on these missing pieces, enabling technological solutions to aging.

As for space exploration, there has been a push for it in the last few centuries, from early Science Fiction to the Apollo Program and SpaceX's work. Expanding across the cosmos would increase our room for growth, resources, and many other things humanity cares about. Yet there have been difficulties on this path mostly due to technological and resources difficulties: space exploration requires means of space travel that are both fast, resource efficient, and not noxious to human life, as well as ways to terraform new planets. The automation of engineering and science will unlock many of the manufacturing, scientific, and engineering insights and tools required to do so, making space exploration a real option.

The Challenges Left

As discussed above, Safe Transformative AI will unlock a wealth of opportunities for improving human lives and flourishing by allowing the automation of all intellectual and physical labor, and thus creating plenty of resources, leisure opportunities, and accelerating technical and scientific progress.

Yet these extraordinary achievements must not be confused with a panacea that solves all problems of human civilization. For not only are there problems which cannot be fully addressed by technological progress, but progress itself generates whole new challenges and exacerbates existing ones. Here are the most obvious

and salient ones, in the understanding that even more will emerge that cannot be predicted now:

First, although Safe Transformative AI will create an abundance of resources through manufacturing and technological acceleration, this does not address the question of the distribution of these resources. Notably, there is a risk that these resources will only accrue to a select few which own the means of automation, increasing drastically inequalities in society. This is first an obvious moral issue: such a situation could mean that the vast majority of people live in terrible near subsistence level, potentially with no access to trivial-to-generate energy and medicines. But it is also a massive structural problem: any world where the vast majority of resources are centralized in the hands of a few, whoever these few are, is not going to be stable economically and institutionally.

These are questions about how humanity organizes society, not technical problems. As such, they will not be addressed by Safe Transformative AI, but need to be discussed and solved through global coordination, policy, regulation.

Even if the abundance of goods and resources created by Safe Transformative AI are redistributed in a satisfying way, different people want conflicting things, in ways that require some sort of trade-off and compromise. The simplest possible case is the one of positional goods: if multiple people want to be "the richest person on earth" or "the special someone of a certain famous person", there is no solution where everyone gets what they want, because there can be only one of these at a time. Furthermore, people have genuine differences in their beliefs about how individual and social life should be arranged: trade-offs between equality versus efficiency (or between different interpretations of equality), religious beliefs, civic symbolism, and more.

These are fundamental problems that will not be solved by technology even in the limit, because there is no "solution": the constraints contradict each other. Instead, what is needed is a compromise.

These disagreements will be exacerbated by the fact that Safe Transformative AI unlocks much more opportunities than can all be exploited at the same time. That is, even automation of all intellectual and physical labor will not remove the need for prioritization of how this automation and the resources it generates are used.

Notably, at each point humanity will need to decide how much of our resources it wants to dedicate to exploration versus exploitation. Investing more into new fundamental science, exploration of space, of new forms of engineering, versus exploiting the technology yielded so far to ensure all diseases that can be cured at a

given moment are cured for everyone, that every single person has a minimum level of resources necessary for flourishing. People's fundamental disagreements about the relative value of these priorities, of exploring versus exploiting, will mean that any decision will require compromise. And once again, technology is impotent to solving this coordination problem.

Even more problematic, human civilization currently lacks the wisdom to know how to use, or refrain from using, technologies that will be unlocked by Safe Transformative AI. Humanity is already unable to address the mild threats to culture, political life, and mental health caused by existing social networks; how is it supposed to cope with future digital worlds and simulations that will be much more convincing, satisfying, and meaningful than reality?

And attractive digital simulations are only the tip of the iceberg: how should humanity act upon the expected ability to edit people's brains and personality, in a way that fundamentally changes what they want? How should it regulate, control, bring into existence and shape technologies which make it easier and easier to cause damage, such as cheap synthetic biology or in-your-backyard nuclear fusion? What about scientific innovation that unlock more dangerous forms of AIs with accrued risks but even more impressive benefits?

Dealing with all of these new opportunities and risks demands progress on the wisdom of humanity; that is, its ability to pick the branches of the tech tree that empower humans, rather than lead to self-destruction. It means a humanity capable of coordinating around these decisions, preventing adversarial threats and defection. Technology cannot help there: let's get to work.

Annexes

Proposed institutional structures

IASC Organizational Structure

- **Council:** Each signatory to the treaty can appoint a representative member to the Council. Each member has equal voting rights.
- **Executive Board:** Analogous to the UN Security Council, this consists of representatives of major member states and supranational organizations, which would all be permanent members with vetoes on decisions taken by the Executive Board. The Executive Board also includes non-permanent representatives elected by a two-thirds majority of the Council.
- **General Secretary:** Oversees the running of IASC and is appointed by a supermajority (75%) vote of the Executive Board. The General Secretary sits for a five-year term and can have a maximum of two terms in office. The General Secretary must have multiple duties and powers, including but not limited to:
    - **Lowering the compute thresholds:** The General Secretary formally makes the decision to lower the compute limits established in the Multi-Threshold System, on the advice of the Advisory Committee, following a report of the AI Scientific Measurement Team;
    - **Revocation of the registration status of an AI organization or company:** The General Secretary formally makes this decision on the advice of the Advisory Committee, following a report of the Global Oversight Team;
    - **Revocation of the registration status of companies with a particular national authority:** The General Secretary formally makes this decision on the advice of the Advisory Committee, following a report of the Global Oversight Team;
    - **Ordering the removal of a senior officer of the GUARD lab:** The General Secretary formally makes this decision on the advice of the Advisory Committee, following a report of the Global Oversight Team;
    - **Approval of specific limited exemptions to the Medium Compute Limit, established in the Multi-Threshold System:** The General Secretary formally makes this decision on the advice of the Advisory

Committee, following a report of the AI Risk Analysis Team. Such exemptions could only be granted to licensed organizations for specific narrow model types, under strict safety and ethical conditions and subject to regular review;
    - **Recommending/setting the annual budget for the GUARD lab:** The General Secretary formally makes this decision on the advice of the Advisory Committee, following a report of the Global Oversight Team.
- **Advisory Committee:** A limited group of AI scientists that have been appointed by the Council. The Committee provides recommendations on major decisions, based on reports produced by the teams reporting to the General Secretary.
- **AI Global Oversight Team:** A directorate within IASC that:
    - Oversees GUARD, including its budget, hiring, strategic plan, operations, and provides reports on this to the General Secretary;
    - Audits and assists national regulators with implementing new guidance from IASC;
    - Maintains a list of licensed AI models approved by national regulators within the internationally set middle compute threshold;
    - Undertakes international investigations into undeclared development of major AI models.
- **AI Scientific Measurement Team:** A directorate within IASC that:
    - Maintains international measures/standards for AI capabilities and risks;
    - Provides reports on progress in the development of AI science and safety, including on boundedness, corrigibility, and alignment.
- **AI Risk Analysis Team:** A directorate within IASC that:
    - Provides advice on the overall level of extinction-level and catastrophic risk as a result of AI, through specialized investigations and assessments.

GUARD Organizational Structure:

- **Managing Director:** Appointed by IASC Council for a 10-year term, and is responsible for overseeing the research, operations, and strategic direction of GUARD.
- **Executive Board:** Appointed by IASC Council for a 5-year term, and is responsible for overseeing the Managing Director and the strategic direction of GUARD. Each board member has a single equal vote on issues and a 75% majority is required for all decisions.
- **Scientific Committee:** Appointed by IASC Council for a 10-year term and is responsible for providing specialized advice on AI research and development to the Executive Board and Managing Director.

- **Financial Committee:** Composed of representatives from the national administrations of treaty signatories and is responsible for providing advice on all issues relating to financial contributions and the lab's budget and expenditure.
- **Risk Management Committee:** Runs internal risk management function; collaborates with IASC audit functions.
- O**perational and Research Teams:** A combination of various directorates within GUARD that are responsible for delivering on its strategy, to include divisions such as: (1) Directorate of Alignment; (2) Directorate of Boundedness; (3) Directorate of Capabilities Assessment and Development; (4) Directorate of Fundamental AI Research; (5) Internal Safety Audit Directorate; (6) Finance Directorate; (7) Information and Technical Security Team.

IAT Organizational Structure:

- **Chairperson:** Chaired by a representative selected by the Council on a rolling 10-year term. The Chairperson facilitates meetings, guides dispute resolutions, and represents the IAT externally. The Chairperson can also unilaterally refer cases to the Rapid Response Panel.
- **The Court:** Consists of 31 judges appointed by IASC Council to serve six-year renewable terms, with two main elements:
    - **Chambers System:** Modeled on the European Court of Justice, by default cases are heard by a Chamber of 5 randomly selected judges or in significant cases (as defined through treaty terms) by a Grand Chamber of 15 judges.
    - **Advocates General Procedure:** To aid in the processing of cases, Advocates General are appointed by the General Secretary to provide independent opinions on the legal issues in cases before the court on all issues. If the Advocate General makes a finding that there are no substantive new issues of law in the case, they shall refer to any advice and decisions that had been made on any previous relevant cases.
- **Risk Assessment Panel:** A panel of 2 judges and 3 experts drawn from IASC's AI Risk Analysis Team with the responsibility of:
    - Being the first point of contact between a submitted case and the IAT;
    - Making a rapid assessment about the risks of the case being subject to a prolonged arbitration process, and to make a decision on whether to refer the case to the Rapid Response Panel;
    - Setting the time limit of a Rapid Response Panel determination if needed.

- **Rapid Response Panel:** A specialist panel capable of convening swiftly to address urgent cases, formed of 3 judges. By default, Rapid Response Panels have a maximum of 30 days to take preliminary action (e.g., a temporary restraining order). If initial action is not made within the allotted period of time, then the case is referred to the Risk Assessment Panel to make a snap judgment on.
- **Appellate Body:** Consists of 7 members serving staggered four-year terms, appointed by IASC Council.
    - All judgments made by the IAT are legally binding within the framework of international law that the AI Treaty establishes, however, findings of a Court or Rapid Response Panel may be appealed. The Appellate Body can uphold, modify, or reverse legal findings and conclusions.

Annex 2 - Reasoning underpinning the Multi-Threshold System

The Upper Limit is set approximately at the highest amount of compute that any AI model has been trained to date. Until significant progress has been made on safety research, AI capabilities should not be further advanced, hence nobody is permitted to train models above the Upper Limit.

It's not possible to be sure that systems at current capabilities levels are safe. In this proposal the most powerful systems are trained in the GUARD lab, providing access to the APIs of models that are reliably safe, and hence only the GUARD lab can train models above the Medium Limit.

The Lower Limit is placed at a level where development of dangerous AI systems seems plausibly possible. Above this limit developers are required to obtain licensing.

The maximum permitted performance of computing clusters are calculated keeping the following aims in mind:

- We want to ensure that no actor apart from GUARD can quickly get in range of the Upper Compute Limit, for example by running an illegal training run, surpassing the training compute limits with relatively low timeframe for detection.
- We want to ensure that unlicensed actors can not quickly get in range of the Medium Limit, in which only licensed actors and GUARD is permitted to train models, since these will have a high level of capabilities, and without proper safety best practices may be dangerous.
- We do want licensed actors to be able to train models permitted for them within reasonable timeframes.
- We do want unlicensed actors to be able to train models permitted for them within reasonable timeframes.
- We don't want to ban commonly-owned personal computing devices.

To achieve these aims, we can focus on the amount of time it takes to train a particular illustrative model given both the total desired model size and the compute capabilities of a computing cluster at a lower limit.

Log10(Total Training Compute / Performance of Computing Cluster)	Time to train (Days)
1	0.000116
2	0.00116
3	0.0116
4	0.116
5	1.16
6	11.6
7	116
8	1,160
9	11,600
10	116,000

In this system, we have a difference in order of magnitude of 8 between the Upper Limit Compute limit and the Medium Limit Computing Cluster limit, meaning that it would take a licensed cluster 3.2 years to breach the Upper Limit in an illegal training run, giving authorities ample time to intervene. However, licensed actors could still train any permitted model within 12 days, since there is a difference in magnitude of 6 between the Medium Limit's Compute and Cluster limits.

We also have a difference in order of magnitude of 8 between the Medium Limit Compute limit and the Medium Limit Compute Limit and the Lower Limit Computing Cluster limit, meaning that unlicensed actors would take 3.2 years to breach the Medium Limit (and theoretically hundreds of years to breach the Upper Limit). However, unlicensed actors could still train any permitted model within 12 days, since there is a difference in magnitude of 6 between the Lower Limit's Compute and Cluster limit.

Annex 3 - Some interventions we considered but decided against

Regulating Model Size

Model size, as measured in the number of parameters that an AI model has, is a predictor of model performance and capabilities. But we found that compute is a preferable proxy for regulation for two reasons: i) model size strongly correlates with training compute, due to scaling laws, meaning that model size is not a more efficient proxy for capabilities than training compute is; ii) hardware is easier to monitor, and since few companies can afford the huge computational resources necessary to train frontier models, regulating compute means only having to monitor these few actors.

Leaving Advanced AI Development Decentralized

While we do advocate for a licensed development of frontier models by private companies, the risks from allowing a competition - whether it be between companies or nation states - to develop the most advanced AI models are simply too high to be tolerated. Proliferation of advanced models would mean a proliferation of opportunities for serious loss-of-control or weaponization to take place.

Regulating Training Data Breadth

Regulating training datasets is appealing since how varied a model's training dataset is may predict how varied the model's capabilities are, and also because volume of training data also predicts a model's performance. We considered multiple options of AI training data regulation to achieve different objectives, and will share them in future iterations of this project.

A 'Formula One' Style Regulatory Regime

One potential criteria for awarding licenses to frontier developers would be to make their licenses contingent upon a track record of responsible and safe development. With regards to frontier AI development, this would have the advantage of making it difficult for younger AI companies that don't yet have a track record of frontier development to join the licensing system. This would limit the number of companies

that can join the frontier race, thereby decreasing the chance of catastrophe from race dynamics.

However, this system makes a dangerous and unjustified assumption: that past track record is a strong predictor of future safety practices; unfortunately, this claim is not justifiable at this stage of maturity of the AI industry. In addition, such a system increases the prospect of regulatory capture on behalf of the frontier labs already competing.

A Narrow Path

How to secure our future

Andrea Miotti, Tolga Bilge, Dave Kasten, James Newport

October 2024 V1.0