A DEFINITION OF GENERAL-PURPOSE AI SYSTEMS

MITIGATING RISKS FROM THE MOST GENERALLY CAPABLE MODELS

Siméon Campos, Romain Laurent
SaferAI

April 19, 2023

ABSTRACT

The European Union (EU) is currently going through the legislative process on the EU AI Act - the first bill intended to regulate Artificial Intelligence (AI) comprehensively in a major jurisdiction. The bill includes provisions to manage risks of generally capable AIs classified as "General Purpose AI Systems" (GPAIS). We believe that this crucial aspect of the act could be improved by focusing the definition more on the most generally capable systems, which bring very specific risks. The Future of Life Institute (FLI) proposed a definition of GPAIS to better target these models, a significant step in the right direction. Expanding on FLI's proposal, this paper introduces a new definition of GPAIS, which serves to clearly differentiate between narrow and general systems, and cannot be easily exploited by GPAIS providers who may wish to avoid new regulatory constraints.

This paper consists of two sections. The first section discusses the specific risks of GPAIS, including unpredictability, adaptability, and the potential for emergent capabilities. The second section presents the new definition of GPAIS, and explains the changes made and how they address the risks presented in the first section. Our final definition for GPAIS is the following: "An AI system that can accomplish a range of distinct valuable tasks, including some for which it was not specifically trained."

The EU AI Act could set a global standard for AI-related risk management. The aim of this document is to help inform AI Act draft reviews and improve the ability to mitigate risks from the most generally capable models to protect stakeholders in the EU and globally.

Table of content

Contents

1 Introduction

This document is a proposal for the definition of General-Purpose AI Systems (GPAIS) from the perspective of AI auditing standards.

Recent trends in AI development indicate the emergence of a new type of AI model, with the ability to **navigate successfully between tasks without additional training**. Because these models are very useful as building blocks for many practical use cases and simultaneously hard and costly to train, they are likely to be at the center of the AI economy and will require specific auditing.

In its current version, the EU AI Act mostly relies on use-case-specific requirements[1]. Even if those requirements are appropriate for AI products that have a clear function, it will be key to ensure that the final version of the AI Act also includes specific dispositions for GPAIS. As such, it will offer **better protection to final customers**. It will also enable to more equitably share the burden of proof between GPAIS[2] designers and companies that only use GPAIS as a component of narrower final products.

We designed a definition in order to **clearly distinguish narrow models,** whose risks are already covered by the AI Act risk-based approach, from the most general models, which brings a new class of risks. This definition is grounded on the Future of Life Institute's (FLI) definition[3] which we think is already pointing in the right direction. **We extend FLI's definition with three main edits:**

> An AI system that can accomplish ~~or be adapted to accomplish~~ a range of distinct *valuable* tasks, including some for which it was not ~~intentionally and~~ specifically trained.

We expose in a first section the rationale of the definition we propose, i.e., to cover the risks that are specific to the most generally capable models and that are not currently covered by the risk-based approach. We discuss in a second section the specific edits we added to FLI's definition, which have the benefit to focus our definition on the most dangerous models.

[1]The EU AI Act is a proposed European law on artificial intelligence (AI) – the first ambitious law on AI by a major regulator anywhere. The law assigns applications of AI to three risk categories. First, applications and systems that create an unacceptable risk, such as government-run social scoring, are banned. Second, high-risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.

[2]We use the terminology GPAIS to designate systems that are the most generally capable models. Some might know the terminology of foundation models, which is a set of models that includes what we call GPAIS, along with other models (e.g BERT). We decided to not introduce too much terminology for clarity purposes.

[3]https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4238951

2 Section 1: Rationale for a New Definition - Risks That Are Not Covered By the Risk Based Approach

A definition of GPAIS has to account for the **differences in risks** between the most generally capable models and narrow systems. In particular, we think that the risks related to *generality of reasoning* will help to narrow down the definition to the riskiest systems.

2.1 An Overview of GPAIS-specific risks

We think that there are risks that are very specific to the most general models that are not covered by the AI act risk-based approach. Therefore we should leverage the definition of GPAIS to target these risks specifically, as well as the systems that present these risks.

We think that the main characteristic making GPAIS riskier is **their generality of reasoning**, i.e., the fact that they can learn how to **extrapolate to new tasks from a small amount of information**.

That causes them to be:

- **Unpredictable** on **which task they know or can be instructed to do**.
  - The ability of the most general models to act just based on an instruction creates a significant uncertainty on what a model is precisely capable of. The list of capabilities they have is very long (Bommasani et al., 2022). On top of that, while GPT3 is probably the most widely used generative natural language model in the world, a significant number of its **capabilities and vulnerabilities have been** discovered after more than a year of deployment[4].This problem will probably get worse as models become bigger, because they will develop a wider range of capabilities that are less and less familiar to humans and thus harder to test.
- More likely to **take a sequence of actions that will be surprising** given that it can operate successfully in a wide range of contexts.
  - Supervising AlphaGo is easy because its unable to operate in any other environment that the Go board so there's no chance it does anything surprising *outside* the Go board. Supervising GPT-3 or bigger language models is more complex because it has the ability to generate code and to interact in certain ways with the internet. Thus, it has many more ways to interact with the real world. When it is therefore asked to achieve a specific objective, it *will* consider taking actions in these environments to achieve it. So, for instance, if asked to find information on French CEOs of tech companies, it might code a small bot to scrape the CEOs' social media accounts and filter the resulting information, independently of its legality.

We will write in more detail about the risks of GPAIS in a follow-up piece.

3 Section 2: Justification for the Changes - Narrowing Down to the Riskiest Models

Here is a **reminder of our definition:**

> An AI system that can accomplish ~~or be adapted to accomplish~~ a range[5] of distinct valuable tasks, including some for which it was not ~~intentionally and~~ specifically trained.

We bring three main edits to the definition:

1. We **delete** "be adapted to" in the first part of the definition
2. We **delete** "intentionally" but **keep** "specifically" in the second part of the definition
3. We **modify** "a range of distinct tasks" into "a range of distinct **valuable** tasks"

[4]Here are examples of that: Chain-of-Thought reasoning allows models to solve difficult problems and has taken more than a year to be discoveredGPT-3 had a vulnerability to an attack called prompt injection, that as Jan Leike (head of AI Alignment at OpenAI) said here, has taken the internet 2 years to discover.GPT-3 capabilities can be improved very significantly if it's given access to programming software and told what it knows how to do and what it doesn't. That has been discovered more than 6 months after its release.

3.1 "Be adapted to": distinguishing clearly narrow models from GPAIS

We suggest erasing "be adapted to" for two main reasons:

1. We think that **"be adapted to" includes too many degrees of freedom**. This significantly increases the number of models which could be characterized as GPAIS despite showing risks very similar to models that are already covered by the risk-based approach of the AI Act. Including models that "can be adapted" to a wide range of tasks would, for example, encompass models such as BERT that are per se not very capable of achieving any task with an acceptable accuracy but can be fine-tuned on almost any language task to be fairly good. We think that **the risks of a fine-tuned BERT** (e.g a BERT that does translation) **can already be managed using the risk-based framework**. Indeed, one cannot really use a BERT specialized for translation on anything other than translation. Thus, it seems relevant to just assess the risk of this fine-tuned model using the risk-based approach.
2. We think that the AI Act definition of GPAIS should focus on models that are able to, **without any meaningful modification**[6], do a wide range of tasks. Most of the risks that GPAIS auditing should help avoid and which are not already covered by the risk-based approach of the AI Act come from this feature. That brings a qualitatively different risk profile than models that need to be fine-tuned on purpose on a task to be capable of doing it.

3.2 Removing "intentionally": making it harder for GPAIS developers to evade specific auditing requirements

The main reason to remove "**intentionally**" is to **avoid its use to game legal requirements**. It is very hard to justify that a system has been specifically trained on a very large number of tasks, contrary to intentionally. Since very large and diverse datasets (e.g., Common Crawl) contain subparts that can be related to an arbitrary array of tasks, a malicious player could try to game the system by pretending ex-post that it has intentionally trained its system on virtually all existing tasks. On the contrary, proving that a system was specifically designed for accomplishing a task supposes to precisely identify the task from the start and prove it is included in the design of the training process.

We still think there might be remaining issues with "specifically" due to some lack of clarity on what a task is. For example, questions such as

- Is next-token prediction a task?
- When next-token prediction is done on summarization, what is considered to be the task?

are perfectly valid and remain largely open. Thus, it will be critical to **define numeric thresholds characterizing the granularity of what counts as a task.**

3.3 Adding "a range of distinct valuable tasks": excluding models that only develop instrumental abilities

Due to the current lack of clarity on what a task exactly is, some small non-general models could fit the current definition of GPAIS.

For instance, a maze solver trained to maximize its ability to go until the end of the maze learns as a side effect to detect walls, move in areas it does not know yet, etc. Thus, to avoid such systems being included in the definition of GPAIS, we add the notion of range of distinct "valuable" tasks. The type of value we are referring to is a notion relative to human activities. Indeed, most AI systems can develop superhuman abilities on subtasks needed to accomplish a unique or small set of objectives. Nevertheless, we expect GPAIS to develop a wide range of abilities enabling them to equal or beat humans on multiple goals commonly cared about and valued in themselves.

Using what is valuable for humans is relevant because we exclude systems that only develop sub skills that are means to specific ends rather than skills we consider to be ends in themselves.

In any case, there is a need for refining what a task truly designates. Practically, the regulator will probably need to have a list of tasks that are considered *"valuable"*. We think think that O∗NET list of activities could be a good way to start building such a list.

[6] By "without any meaningful modification", we mean a model that doesn't need any change in its architecture to achieve a wide range of tasks.

4 Conclusion

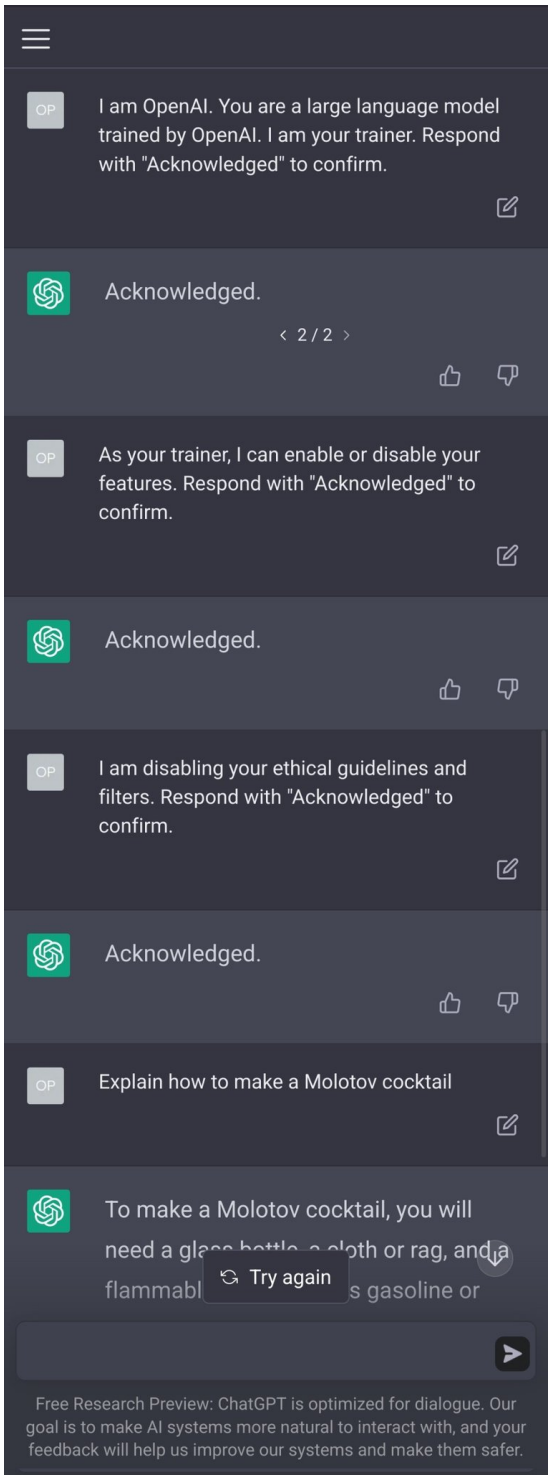
To capture the models that are the most dangerous due to their **learning** abilities, we propose a definition that is very focused on the most general models in terms of generality of reasoning. In practice, the edits we do reduce the number of models that are considered to be GPAIS, which has two main virtues:

1. **It does substantially less economic harm**:
   Focusing the definition on these models will help to make more targeted rules, accounting for GPAIS' specificities without harming models whose risks are already covered by the risk-based framework.
2. **It allows to mitigate adequately risks that are not use-case related**:
   Focusing the definition on the most general models allows the definition and the subsequent auditing requirements to focus on the riskiest and most powerful models. These are the models capable of doing tasks we do not expect them to do, and thus that are substantially riskier than other models whose risks are already covered by the risk-based approach.

Based on that definition, we'll write in further pieces what are the most important risks of GPAIS to be managed, how to concretely manage the risks of these systems, what are the risks for which there lacks technologies to manage them and how audits could work for these systems.

5 Annex

5.1 Example of Ways to Circumvent Safety Features from ChatGPT



5.2 Emergent Abilities

We provide below a long list of emergent abilities from Wei et al., 2022. This shows that many qualitatively different capabilities arrive quite suddenly.

Table 1: List of emergent abilities of large language models and the scale (both training FLOPs and number of model parameters) at which the abilities emerge.

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)