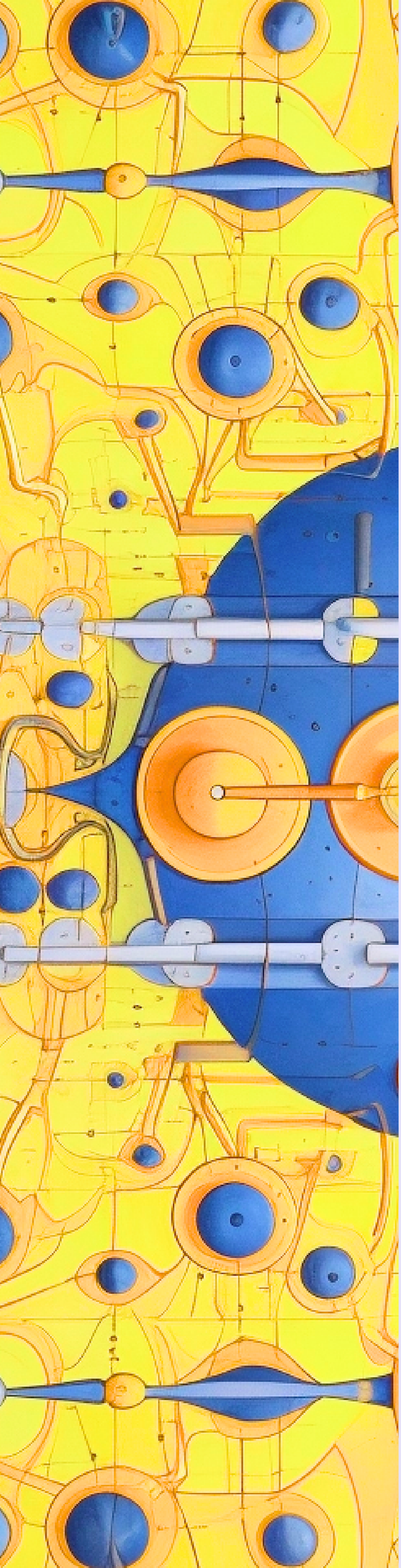


THE FUTURE SOCIETY

September 2023

Heavy is the Head that Wears the Crown

A risk-based tiered approach to governing General Purpose AI

Contact: nicolas.moes@thefuturesociety.org

Cite as:
Nicolas Moës and Frank Ryan, “Heavy is the Head that Wears the Crown: A risk-based tiered approach to governing General Purpose AI” (The Future Society, September 2023).

We thank Toni Lorente for his feedback. All remaining errors are our own.

Table of Contents

Heavy is the Head that Wears the Crown

A risk-based tiered approach to governing General Purpose AI

September 2023

Table 1: Summary of measures per tier and addressing enforcement and complicating factors

Generative AI Applications	Type-I GPAI	Type-II GPAI
Data Governance	Risk Management System	Dialogue in Navigator Programme
Content Moderation Safeguards	Basic Trustworthiness	Absolute Trustworthiness
Labelling Output	Reporting of Compute	Internal & 3rd Party Auditing
Transparency on Model Used	Quality Management System	Quality-By-Design Process
	Compliance Function & Officer	Major Accident Prevention Policy
	Notify Training Runs & Model Pre-registration	Review & Approval of Designs
	Know-Your-Customer	Responsible Staged Development & Release
		High-Reliability Organisation

Enforcement & complicating factors

Enforcement:

Navigator Programme
Regulatory Sandboxes
AI Office
EU Benchmarking Authorities
GPAI Models Database
Technical Thresholds Updating

Combination of models:

Managing Unintentional Interactions
Managing Reasonably-foreseen Interactions

Open Source:

Open Source Observatory
Future-proofing Adaptation

Value Chain governance:

De Facto Control Contractual Framework
Tier-wise Conformity Assessment

Abstract

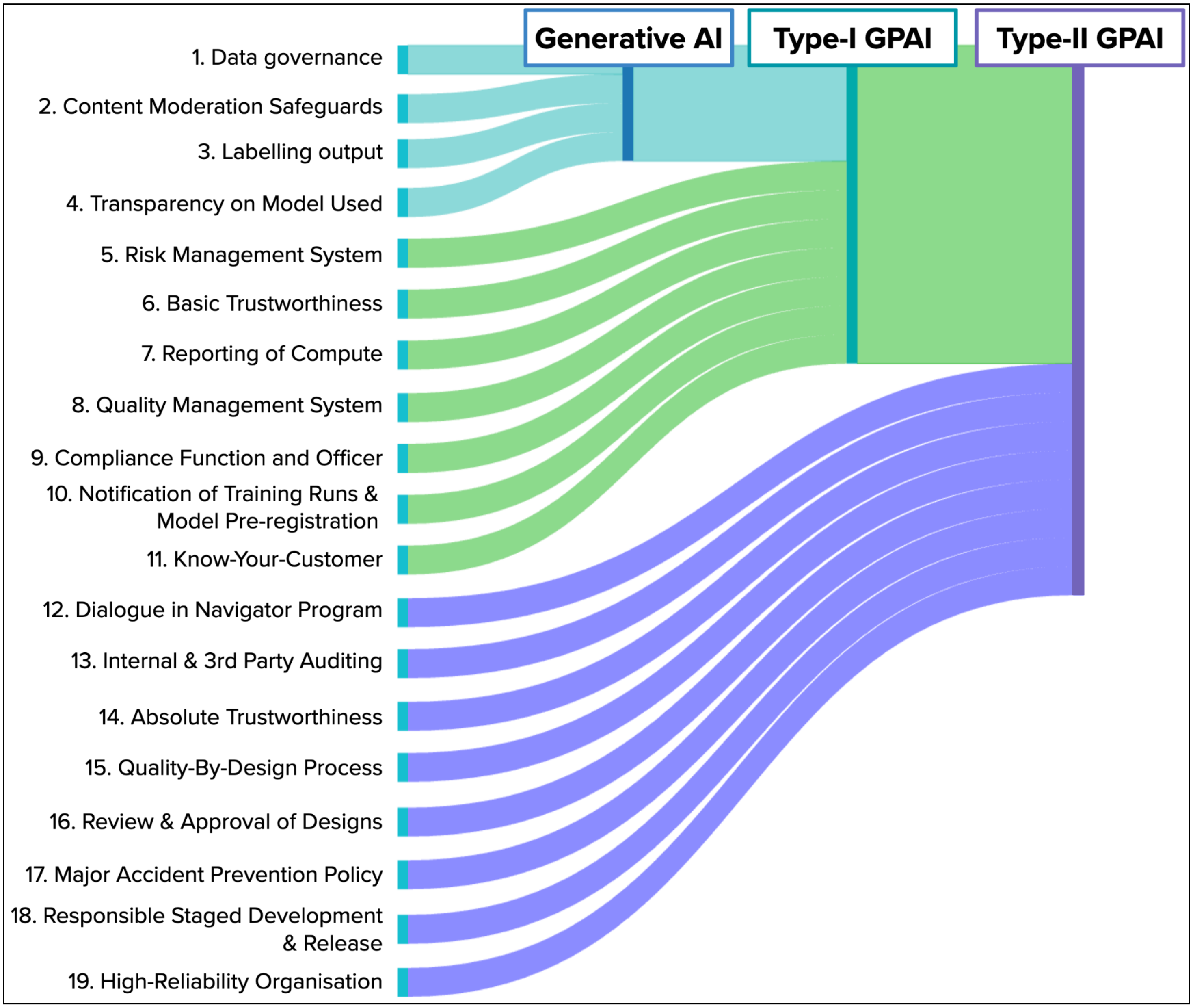
This report provides a **Blueprint for establishing a risk-based tiered system for the governance of General Purpose AI (GPAI) models.** Drawing on 2 years of theoretical, advisory and field research on the governance of these models in the AI Act, this guide distils our findings into actionable steps and a holistic system of governance, based on the AI Act's risk-based, technology-neutral, product safety framework and on the state of the debate. It also resolves the legal uncertainty while remaining future-proof.

We identify and describe seven challenges particularly salient to the GPAI industry: **Infrastructural Aspect (3.2.1)**, **Generalisation and Capability Risks (3.2.2)**, **Concentration of Power (3.2.3)**, **Corporate Irresponsibility (3.2.4)**, **Misuse (3.2.5)**, **Technical Opacity (3.2.6)** and **Incidents and Accidents (3.2.7)**. We also identify three complicating factors in that industry: **Combination of Models (4.1), Open Source models (4.2)** and **Value Chain governance (4.3).** As challenges do not apply equally across all GPAI models, we decompose "GPAI" into three categories, related in a **tiering system defined by generality of capabilities**:

1. **Tier 1: Generative AI Applications**, where a model's capabilities have been narrowed down to specialise in a specific subset of generative tasks (over 400 providers in this tier)
2. **Tier 2: Type-I GPAI models**, where a model is designed for generality of capabilities (roughly 13 providers)
3. **Tier 3: Type-II GPAI models**, which are 2022's cutting edge and beyond in terms of generality of capability (roughly 6 providers, who are all also included in the 13 Type-I GPAI providers)

We describe the requirements and obligations in sections 7 and 8; summarised & illustrated in table 1 and figure 1 below. A "Tier 1.5" exists, consisting of the foundation models described in detail in the European Parliament's compromise text [1], thus beyond this Blueprint. A speculative Tier 4 also exists, consisting of a flurry of models expected to emerge past 2023, which should not be developed until adequate risk mitigation strategies are developed and implemented at least in Tier 3 models.

Figure 1:[1] A tiered approach for the governance of GPAI & generative AI


Generative AI
Type-I GPAI
Type-II GPAI
1. Data governance
2. Content Moderation Safeguards
3. Labelling output
4. Transparency on Model Used
5. Risk Management System
6. Basic Trustworthiness
7. Reporting of Compute
8. Quality Management System
9. Compliance Function and Officer
10. Notification of Training Runs & Model Pre-registration
11. Know-Your-Customer
12. Dialogue in Navigator Program
13. Internal & 3rd Party Auditing
14. Absolute Trustworthiness
15. Quality-By-Design Process
16. Review & Approval of Designs
17. Major Accident Prevention Policy
18. Responsible Staged Development & Release
19. High-Reliability Organisation

[1] Designed with SankeyMatic (https://sankeymatic.com/build/)

1. Executive Summary

General Purpose AI (GPAI) has become a heated topic of debate since 2021. The European Parliament, the European Commission, and the Council of the European Union are now discussing how to best introduce a governance regime for GPAI. In this publication, we explore what an ideal regulatory regime would look like. To do so, we identify **seven challenges relating to GPAI and three complicating factors.** This allows us to categorise these models into **three distinct tiers based on their range of capabilities and associated levels of risk.** We then compile **explicit requirements and obligations for GPAI providers** and **high-level enforcement mechanisms** for **regulatory authorities**.

In section 3, we analyse how GPAI technologies disrupt several aspects of society, from security and economy to public mental health. We derive **seven challenges** posed by GPAI:

- **Infrastructural Aspect** i.e. built-in development decisions, large user base, economic ubiquity, and user lock-in;
- **Generalisation and Capability Risks** i.e. capability risks, societal risks and extinction risks;
- **Concentration of Power** i.e. monopolistic tendencies, vertical/horizontal integration, and barriers to entry;
- **Corporate Irresponsibility** i.e. lack of attention to quality/compliance, race dynamics, silencing of criticism, and resulting irresponsible behaviours;
- **Misuse** i.e. intentional or accidental misuse, vulnerabilities, and knowledge asymmetries in the value chain;
- **Technical Opacity** i.e. paradigm opaque by design, lacking interpretability, predictability, corrigibility and controllability;
- **Incidents and Accidents** i.e. bias, discrimination and automation of microaggressions; misinformation & privacy violations; and accidents in development & deployment.

We then discuss three main complicating factors in section 4. First, **the combination** and **interaction** of different **models** accentuates most challenges. Second, the **open-sourcing of GPAI models** can mitigate some challenges (Concentration of Power and Technical Opacity) but significantly exacerbate others (Misuse and Incidents & Accidents). Finally, **governing the value chain** is crucial to re-balance power dynamics and obtain corporate accountability - it also plays a pivotal role in establishing a level playing field for all the actors involved.

In section 5, we argue that different challenges apply differently to different sub-categories under the umbrella term "GPAI", **based on the generality of the GPAI model's capabilities**. As the AI Act should address these challenges in a proportionate and risk-based manner, we distinguish between 3 categories and provide operational definitions. In brief, **Generative AI applications** are implementations of AI techniques for the purpose of producing new content. **Type-I GPAI models** are AI models that are designed for generality of capabilities. **Type-II GPAI models** are AI models that are designed for generality of capabilities and expand the technology frontier relative to 2022 models.

In section 6, we propose a **tiered approach**, where **requirements** for GPAI models and Generative AI applications are **set in proportion to their risk potential**, avoiding undue regulatory burden. The approach includes three tiers: **Generative AI applications**, **Type-I GPAI models**, and **Type-II GPAI models**. They are differentiated via a set of criteria for generality of capabilities before deployment, as this dimension correlates with the challenges & risks identified. The highest level of scrutiny is reserved for the models with greatest generality of capabilities, which pose the greatest potential risk and challenges: Type-II GPAI.

We put forward in Box 1 four different ways to assess a model's generality of capabilities. **First**, via the **amount of compute** used during training, measured in so-called "**FLOPs**". Second, via **skill-acquisition efficiency**, which describes the efficiency with which a system or individual can acquire new skills. Third, via a **generality analysis** that simultaneously evaluates the versatility and performance across tasks. Fourth, via **algorithmic efficiency and model perplexity**, which allow inferring generalizability as a function of a model's computational power, algorithm and the richness of its training data. Last, via **modality-specific benchmarks**. It is important to encourage industry to report and predict generality of capabilities in a consistent manner. The table below summarises the tiers and their scope:

Full table: indicative operational tests for tiered-approach and number of regulated entities[2]

Tier	**Name**	**Operational test(s):**					**Estimated # of regulated entities**
#1	Generative AI application	1. Built upon a general purpose AI model 2. Refined for a specific purpose through prompt-based training, fine-tuning, reinforcement learning with human feedback, or other methods to narrow the model's purpose to a specific task with limited scope					>400 providers, several 1000s of applications
#1.5	*Foundation Models (EP)*	1. Can be applied to a wider range of tasks than tier #1 2. But still only <10^21 2022-FLOP to train the model					40-80 providers, ~85-170 models
If any of these criterion is met:		*Total amount of FLOP used to train the model (2022-FLOP)*	*Modality- specific benchmarks (e.g. MMLU average for language)*	*Skill Acquisition Efficiency (ARC Challenge)*	*Generality Analysis*	*EU-endorsed summary benchmark*	*Estimates for tiers 2 & 3 based on compute estimates, as other test results unavailable*
#2	Type-I GPAI model	>10^21 ≤10^23	>40.0 ≤68.0	>40/800 ≤60/800	...	...	14 providers, 62 models
#3	Type-II GPAI model	>10^23 ≤10^26	>68.0 ≤88.0	>60/800 ≤100/800	...	...	10 providers, 28 models
#3+	*Prohibited?*	>10^26	>88.0	>100/800	...	...	0 provider, 0 model

[2] An earlier version of this table circulated in September 2023 with a confusing threshold number on the "Modality-specific benchmarks" column. This is now clarified.

In section 7.1, we present a set of mechanisms to be put in place to achieve an efficient and responsive implementation of the Act's rules for generative AI and GPAI models, including the tiered approach:

- **Navigator programme**, which fosters direct bilateral relations between the European Commission or AI Office's staff and each Type-II GPAI development team to promote trust and compliance.
- **Regulatory Sandboxes**, which allow for testing new products in a real-world environment for developers and regulators to better understand the technology.
- **AI Office**, which would act as a central point of contact for all stakeholders, concentrating expertise and enforcement capacities.
- **EU-level Pool of benchmarking authorities**, capable of bringing together Member State's metrology and benchmarking authorities to promote accountability and consistency across norms and standards.
- **Database of GPAI models** hosted by the AI Office, in which all GPAI providers register their models in Europe to facilitate the work of the Commission and the Member States and to foster transparency.
- **Updates of technical thresholds** in the legislation, such as the adoption of implementing acts by the commission where the technical aspects are specified to ensure GPAI is effectively governed.

In section 7.2, we present the main measures to effectively **govern the combination or interaction of models:**

- **Managing Unintentional Interactions**, through proper assessment, communication and mitigation if during the testing or at deployment a GPAI model unexpectedly interacts with one or more GPAI models.
- **Managing Reasonably-foreseen Interactions**, through satisfactory ex ante assessment and communication to the AI Office, in order to build and maintain an industry-wide map of models' interactions.

In section 7.3, we present measures to help effectively **govern open source models** in a future-proof way:

- **Open source observatory**, which shall be joined by all open source providers as well as open source hosting platforms, foundations, experts and representatives from civil society, to assess and refine rules for open source GPAI models.
- **Adaptation for open source providers** of some of the acceptable means for compliance, taking place in conjunction with the open source observatory.

In section 7.4, we discuss **value chain governance**, which is necessary to mitigate five of the seven challenges identified. It is achieved through:

- the **De Facto Control contractual framework**, which is a set of rules to facilitate the evidence-based and proportionate transfer of responsibility for compliance along the GPAI value chain, via regulated contracts.
- **Tier-wise conformity assessment to ensure downstream value chain actors can integrate GPAI model in their products without undue legal risk**, thanks to intermediary or component conformity assessment carried out by the upstream developers of GPAI. For **Generative AI applications, internal conformity assessmen**t is sufficient. For **Type-I**

GPAI models, external conformity assessment is necessary. For **Type-II GPAI models,** given the near monopoly of expertise, **a "joint" conformity cross-assessment is required,** inducing joint & several liability for both the provider and the auditor.

In section 8.1, we present the main requirements to govern generative AI applications:

- **Data Governance**, ensuring that providers' application is developed on the basis of adequate data sets.
- **Minimum content-moderation safeguards**, ensuring that the application is developed so as to prevent the generation of content in breach of union law.
- **Labelling AI-generated output**, ensuring that the output of the model is automatically accompanied by an indication that it has been artificially generated or manipulated.
- **Transparency on Model Used**, clear indication of the model name, model version and model provider's name to users in end-user-facing access interfaces.

In section 8.2, we present the main requirements to effectively govern Type-I GPAI models, which are, in addition to the requirements from the generative AI application's tier:

- **Risk management system**: GPAI providers establish, implement, and maintain a risk management system for the model in a process spanning the model's entire lifecycle.
- **Basic trustworthiness**: the provider proves that the model is designed so as to have sufficient levels of cybersecurity, predictability, interpretability, corrigibility, controllability, robustness and boundedness.
- **Reporting of compute resources**: GPAI providers create systematic processes to forecast, record and report regular use of compute resources for training runs and model operation, along with the energy use associated.
- **Quality Management System**: GPAI providers implement a thorough quality management system that guarantees adherence to the stipulations of the AI Act concerning GPAI models.
- **Compliance function and officer**: GPAI providers establish an autonomous compliance function, separate from the operation of the organisation, and staffed by one or more compliance officers responsible for monitoring the provider's adherence to obligations set out under the AI Act regulation.
- **Notification of training runs & model pre-registration**: GPAI providers notify the AI Office of upcoming training runs, models under development, and pre-register models in their pipeline.
- **Know-your-customer,** to facilitate prevention of misuse: GPAI providers take all necessary and proportionate measures to prevent misuse after detection.

Finally, in section 8.3, we present the main requirements to effectively govern Type-II GPAI models, which are, in addition to requirements from previous two tiers:

- **Regular dialogue with AI Office to update on latest technical advancements** in AI to reduce the knowledge gap between the developers and the Office, through the Navigator Programme.
- **Internal & 3rd party auditing**, imposing joint & several liability on both the provider being audited and the auditor.
- **Absolute trustworthiness,** or that providers design and develop their models to achieve superior levels of advanced cybersecurity and safety.
- **Quality-by-Design process:** augmenting the mandated quality management system for Type-I models with a Quality-by-Design (QbD) process that includes a probabilistic risk

assessment and safety evaluation, akin to drug manufacturing protocols.

- **Review & Approval of designs** by AI Office before training run or that the provider notifies and awaits an opinion from the AI Office, with the authority to delay training runs designated for developing Type-II GPAI models and to review the codebase.
- **Major accident prevention policy**, developed by providers and meticulously implemented, to protect human health and the digital, physical, and natural environments, similar to that of the Seveso Directive and other production processes.
- **Responsible Staged Development & Release,** whereby providers structure their design and development process to scale responsibly and cautiously, with batteries of tests and evals at every checkpoint to be satisfied in order to continue training.
- **High-Reliability Organisation,** or that providers organise their facilities, processes and internal policies as a way to incorporate all other requirements in the practice of the provider and to establish a culture valorizing reliability, safety & trustworthiness.

Figure 1 above provides a visual overview of the tiered approach.

2. Background and Methodology

Summary

General Purpose AI (GPAI) has become a heated topic of debate since 2021. The European Parliament, the European Commission, and the Council of the European Union are now discussing how to best introduce a governance regime for GPAI. In this publication, we explore what an ideal regulatory regime would look like. To do so, we identify **seven challenges relating to GPAI and three complicating factors.** This allows us to categorise these models into **three distinct tiers based on their range of capabilities and associated levels of risk.** We then compile **explicit requirements and obligations for GPAI providers** and **high-level enforcement mechanisms** for **regulatory authorities**.

Since the European Commission introduced its proposal [5] to regulate artificial intelligence (AI) within the European Union in April 2021, the question of how to best regulate General Purpose AI (GPAI) has emerged as a topic of intense debate[3]. The Commission's initial approach, a bold first attempt to regulate AI by any major regulatory body, relied on market forces to incentivize good practices by developers of GPAI: they would have to improve the reliability of their systems for customers subject to the AI Act use-case specific requirements.

However, in January 2022, The Future Society released a pioneering brief on general purpose AI [6]. The brief outlined the market asymmetries hindering successful market-based solutions and the negative externalities the development of this technology poses, such as hazards derived from opacity and unpredictability. It proposed light-touch solutions to preemptively address these issues without stifling innovation. Less than 2 years later, many of the risks identified then, including threats to economic sovereignty, fundamental rights, and the health and safety of citizens, have tragically materialised, often on significant scale.

The Council of the EU's common position [7] of 25 November 2022, to its credit, improved upon the Commission's original proposal by explicitly including GPAI within the scope of the AI Act. It mandated several requirements GPAI should fulfil, but delegated spelling out critical details of obligations and requirements to the Commission through an implementing act. Five days later, on 30 November 2022, the notable GPAI system ChatGPT was launched [8], based on the GPT-3.5 GPAI model.

Witnessing the speed and growing effects of the race for GPAI development,[4] the European Parliament updated its draft common position [9] to introduce a more thorough governance

[3] A note on terminology: while GPAI systems, GPAI models, foundation models and generative have also often been used throughout the debate in various contexts, we use GPAI for simplicity. We provide a full explanation of the relations between these concepts, in the section "5. Defining General Purpose AI for Governance"

[4] Which we detail in the next section.

regime for GPAI. This is reflected by the inclusion of amendments to Articles 28, 28a (new) and 28b (new) in its adopted negotiating position, outlining specific rules for good and fair collaboration across the value chain and for providers of GPAI. The European Parliament also strengthened the mandate and capacities of the EU-level body for the AI Act (Articles 56-58) to match the scale of the task of governing GPAI.

Beyond the EU, other major governments and industry players are beginning to recognise the importance of directing focused attention towards GPAI governance. The White House [10] convened tech CEOs to discuss responsible AI innovation and associated risks, amidst growing concerns and regulatory efforts regarding AI's impact and usage in various sectors. Congress is discussing the possibility of a licensing regime [11] for AI to ensure public agencies like the Department of Defense can keep pace with rapid technological advancements while adhering to national security and cybersecurity requirements. The Office for Science, Technology, and Policy (OSTP) also supports an industry-led project [12] for one-off independent scrutinization (red-teaming) of their key GPAI models. However, this initiative is being implemented through the Scale AI platform, which focuses on deployment risks (which are also currently covered by the AI Act), and not pre-deployment risks.

Elsewhere in the US, OpenAI, Google, Anthropic and Microsoft announced the formation of the Frontier Model Forum [13] with the purported aim of promoting the “safe and responsible development of frontier AI systems” through “identifying best practices and standards, and facilitating information sharing among policymakers and industry”.

The UK, “mindful *of the rapid rate of advances in the power and application of LLMs, and the potential creation of new or previously unforeseen risks”* [14] has made this a core focus of their monitoring and risk assessment functions. The government has also launched a taskforce [15] on safe foundation models and a market review [16] into these models’ risks and opportunities, and announced that it will convene the first ever global AI summit [17] in a bid to broker a common approach by countries to regulating AI.

China also now mandates formal approval [18] of some GPAI, likely driven by national security concerns. With information control a central goal of these measures, they stipulate that both the training data and model outputs must be “true and accurate,” [19] a standard that would be exceptionally challenging, if not impossible, for today’s GPAI models to meet.

Addressing the multifaceted challenges and risks of GPAI, while harnessing their transformative potential, has been the cornerstone of The Future Society's work in improving the governance of AI since 2016. This report aims to synthesise The Future Society’s work on GPAI and provide a holistic answer to the question of what an ideal regulatory regime for GPAI would look like. We begin by identifying seven challenges relating to GPAI. This is done by reviewing the public discourse as reflected in mainstream media, technical publications, academic literature and stakeholders reports. The following chapter briefly outlines factors that add to the complexity of formulating GPAI regulation. Specifically, the existence of combinations of models, open-source software, and a range of different actors operating along the value chain of AI are additional considerations that require careful attention or targeted regulation.

This exploration of the various challenges associated with GPAI allows us to categorise these models into three distinct tiers based on their technical differences, range of capabilities, and associated risk levels. We intend for this to be a preliminary delineation of tiers, recognising that further research into the characterisation, measurement and comparison of these models is necessary for more precise categorisation.

With these three tiers in mind, we then formulate a comprehensive set of regulatory measures for GPAI. These measures translate into explicit requirements and obligations for GPAI providers and high-level enforcement mechanisms for regulatory authorities. Following an extensive exploration of the set of possible measures available to policymakers - drawn from previous work by The Future Society, along with interviews, workshops and relevant academic literature from the past two years[5] - we selected the measures deemed most critical for realistically and cost-effectively mitigating the challenges identified for each tier.

This methodology allows us to systematically anchor each requirement within a specific function it is intended to fulfil, providing reasoning and clarity amid the often contentious public discourse surrounding GPAI model governance. We have frequently noticed a lack of clarity about the purported impact, be it beneficial or detrimental, of introducing a given regulatory measure. To avoid falling into this trap, we not only articulate the rationale behind each measure, but attempt to outline actionable steps that could be undertaken to fulfil . We hope this approach creates a clearer connection between theoretical reasoning behind each measure and its practical application.

A note on the text

The AI Act has evolved at great speed - while a year ago many claimed its arrival was premature, some now seem to think it may be too late given the speed of technological development and deployment. When it comes to GPAI, however, the AI Act appears to be strikingly timely. Compared to other policy files, several stakeholders have expressed admiration for policymakers' ability to address such a complex issue at such a granular level in such a short amount of time. Additionally, while legal and policy representatives have occasionally expressed concern at the lack of legal clarity, numerous technical experts have commended the relevance of the requirements put forward.

The tethering of the AI Act's policy engineering and debate with the fast technological and commercial developments in the GPAI industry has infused the policymaking process with a "live"

[5] While we do not present the full range of measures considered in this publication, a larger sample of all the measures considered in TFS work has been shared with NIST [20]

dynamic that is rarely witnessed in Brussels. Every discovery, incident, and study occuring within the field has required immediate analysis, integration and recalibration, given the fast-growing number of citizens [21] [22] and businesses affected and sudden effect size [23]. Indeed, the governance vacuum surrounding this technology has resulted in practices akin to using consumers and citizens as guinea pigs [24], such as bypassing basic quality management or even sometimes ignoring [25] widespread minimal practices for filtering output and instead test AI systems on millions of users, or as part of "research previews" [26] for paying B2B and B2C customers, even by the largest [22] actors in the industry.

This real-time nature of the policymaking process has made the formation of technically sound policy recommendations a challenging endeavour, requiring a great deal of flexibility and adaptability. The scarcity of independent and academic peer-reviewed literature on GPAI governance has notably led thought leaders to reflect by analogy with other industries facing related problems: some recommendations on several measures that have already demonstrated efficacy for other comparable issues in other industries, such as the use of a Major Accident Prevention Policy in the Seveso Directive [27], or the quality-by-design approach for drug manufacturing.

The Future Society's continuous research in this debate has enabled a nuanced but ever-evolving understanding of the changing technological, political, and market landscapes. As such, the tiered approach to regulating GPAI models that we put forward in this report can be viewed as the culmination of numerous lessons learned and insights gained from two years of dynamic policy advisory. With this in mind, we have unconventionally decided to release this as an early research preview (a common practice within the GPAI industry) in order to keep up with the pace set by the industry, and to ensure that our insights, though continuously evolving, contribute to the advancement of good AI governance in a timely fashion.

3. Challenges posed by General Purpose AI

Summary

In this section, we analyse how GPAI technologies disrupt several aspects of society, from security and economy to public mental health. We derive **seven challenges** posed by GPAI:

- **Infrastructural Aspect** i.e. built-in development decisions, large user base, economic ubiquity, and user lock-in;
- **Generalisation and Capability Risks** i.e. capability risks, societal risks and extinction risks;
- **Concentration of Power** i.e. monopolistic tendencies, vertical/horizontal integration, and barriers to entry;
- **Corporate Irresponsibility** i.e. lack of attention to quality/compliance, race dynamics, silencing of criticism, and resulting irresponsible behaviours;
- **Misuse** i.e. intentional or accidental misuse, vulnerabilities, and knowledge asymmetries in the value chain;
- **Technical Opacity** i.e. paradigm opaque by design, lacking interpretability, predictability, corrigibility and controllability;
- **Incidents and Accidents** i.e. bias, discrimination and automation of microaggressions; misinformation & privacy violations; and accidents in development & deployment.

In this section, we summarise how GPAI technologies have disrupted several aspects of society, from security and economy to public mental health, and how these disruptions have, in instances, caused major incidents. Through analysing recent media coverage, we identify seven challenges in the current GPAI landscape. It should be noted that while this section focuses on challenges and incidents, there is extensive business literature detailing realised [28] and speculative [29] [30] benefits of GPAI. This includes an examination of the socio-cultural "hype" phenomenon.[6] Given that the AI Act adopts a product safety framework, rather than an industrial policy strategy, its primary focus is on ensuring product safety. As such, discussions on EU public sector investments or broader industrial policies for AI fall outside the remit of this report.

3.1 Review of recent GPAI related discourse

GPAI developers are relentlessly pushing the boundaries of this technology, whilst simultaneously acknowledging the problems their advancements may introduce [31] [32] [33] [34] [35] [36]. From an economic perspective, GPAI is already transforming industries such as creative [37], media

[6] Epitomised by the attempted creation of an international "AI Appreciation Day"

[38], education [39] and technology[40], and is anticipated to disrupt many more, including lobbying [41] and legal services [42].

The security implications of GPAI are also becoming increasingly apparent. GPAI has already been used to create a wide range of harmful activities, including malware [43], poisons [44], fraud schemes [45] [46], disinformation [47], revenge porn [48], and child sex abuse [49] scenarios. Despite efforts to 'guardrail' these models, leaks [50] and reverse-engineering [51] are common. Some models [52] [53] are even released open-source without minimum technical safeguards.

Hundreds of thousands of people, including those well-versed in GPAI and AI risks [54], claim to have become friends or fallen in love with [55] [56] [57], or to have been emotionally manipulated [58] or harassed [59] by GPAI-enabled bots. In one extreme case, an AI chatbot convinced a Belgian man to commit suicide [60]. These bots are used by tens, or even hundreds, of [61] of users.[62] At the macro level, GPAI-based content generation & moderation models have damaged epistemic health and polarised society, extending beyond bias and recommender systems [58]. For example, the most popular moderation tool, launched in 2017 and handling over >500 million pieces of content daily, has long censored [64] any discussions of identity & minorities, irrespective of the toxicity of the content, due to flawed R&D.

Today's cutting-edge models operate their continuously expanding range of capabilities more and more autonomously. They are now capable of developing and executing their own plans [65], prioritising and assigning tasks [66] to copies of themselves; reading, debugging, writing and executing [67] software code; interacting with the web [68], synthesising illegal drugs [69], or controlling a robotic system [70], with minimum modification (sometimes just a single prompt). They are functioning without an intended purpose, outside of a professional context, as part of a research preview [26], or without being embedded within a system put into service. This renders some of their most significant risks legally invisible to existing product safety and risk-based regulations worldwide.

Prominent tech companies have already made headlines with unexpected issues and setbacks in their GPAI implementations. Microsoft's Bing Sidney chat which has been publicly tested [71] since at least November in India & Indonesia has caused major incidents [58] [59] [72] [73]. These accidents occurred despite Microsoft following its "top-of-class" Responsible AI Principles. Google's own Bard model was Alphabet's $100bn mistake [74] as, in its rivalry [75] with Microsoft, Google failed to notice that the screenshots they published of Bard's output contained factual errors. Meta's Galactica model had to be recalled because of its propensity for disseminating disinformation [76]. This followed an earlier bot's launch which wasn't newsworthy because it made "users bored" and was "safe" [77], according to Meta's head of AI research, illustrating the dangerous incentives even the bigger players face to maintain shareholders' confidence in their companies' leadership in AI.

The race is on for more advanced GPAI, ultimately for human-level general AI and Artificial General Intelligence (AGI) [78] [79] [80] [81]. And while AGI remains a scientific hypothesis, investments in that direction are occuring on a scale of billions of dollars [82]. The progress in the GPAI field is exponential, with at its peak, half a dozen groundbreaking models[7] deployed between mid-February and mid-April 2023, often replacing older versions. This exponential progress in capabilities exacerbates the problems listed above [84].

Industry players continue to accelerate GPAI development, but call on regulators to rein them in [85]. The regulations they ask are wide-ranging & unexpected, calling for authorities to oversee them, both publicly and behind closed doors [79] [81] [86]. Developers of GPAI also call for more ambitious solutions, both privately and publicly. For instance, an “FDA for foundation models”[8] has been proposed by industry players.[86] This body would mandate trials at various R&D & pre-deployment stages [87], equivalent of Biosafety-levels [88] for labs developing GPAI technologies and the equivalent of an IAEA for monitoring labs [89], as well as the requirement for pre-registering experiments

Furthermore, international coordination for monitoring, R&D governance and a major accidents prevention regime has been widely called for because GPAI models are seen as “permanent research projects”, released as “research previews” and many of these developers’ concerns relate to R&D accidents. Examples include hundreds or thousands of Microsoft Sydney’s beta-testers harassed by its manipulative capabilities, GPT2’s training typo [90] in 2020 which caused thousands of reviewers worldwide to be exposed to ever increasingly traumatising content, and evidence of concerning “power-seeking” [84] and “deceptive” [91] behaviours during models’ evaluation stage, notably of GPT4.

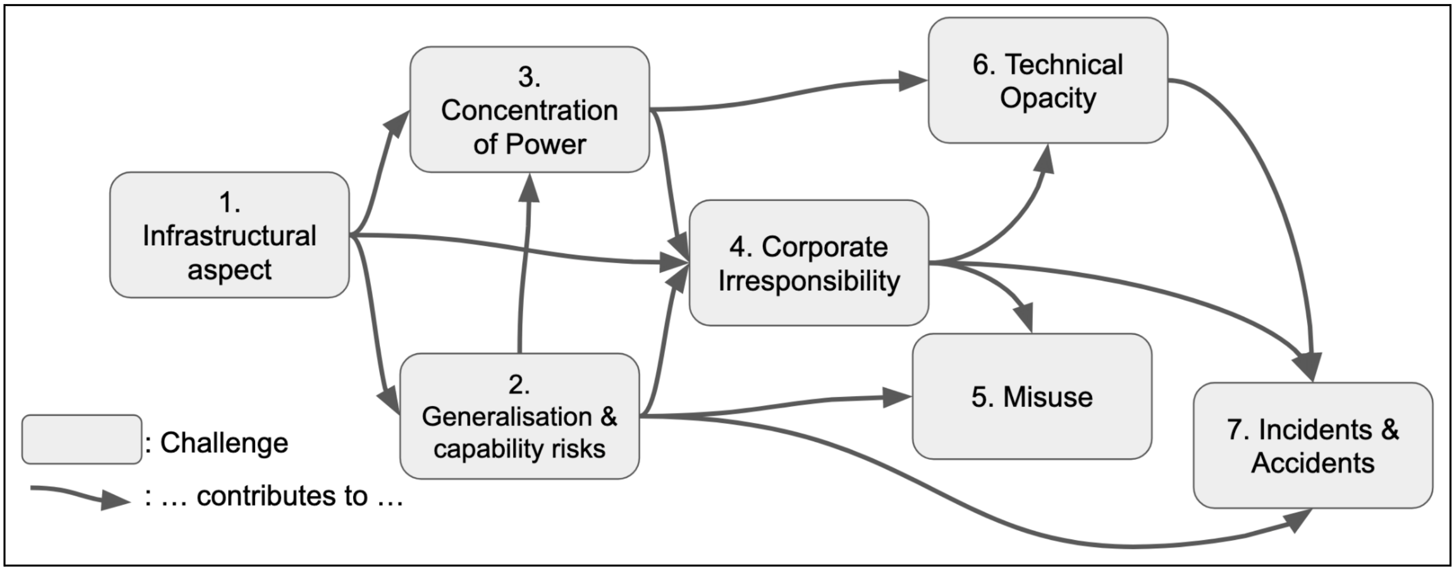
3.2 Challenges identified

In this context, we identify seven challenges and three complications particularly salient for GPAI, as opposed to other types of AI systems. Challenges that arise from all types of AI [92] systems include: beneficial use, non-discrimination & fairness, freedom & dignity, transparency & explainability, robustness, cybersecurity & safety, accountability. These are extensively discussed in the literature and addressed (though far from resolved) in the governance debate. Thus, we only integrate them in our methodology when they are exacerbated by GPAI technologies. Figure 2 summarizes the relationships between challenges identified for GPAI.

[7] OpenAI’s GPT4, Anthropic’s Claude, Microsoft’s Kosmos-1 [83], Google’s PaLM-E [70], Google’s Bard, stability.ai’s StableLM

[8] Food & Drug Administration, the authority notably in charge of overseeing and authorising the manufacturing, testing, progress along stages of trialling, and commercialisation of new drugs in the US.

Figure 2: Challenges identified for GPAI and their relationships


1. Infrastructural aspect
2. Generalisation & capability risks
3. Concentration of Power
4. Corporate Irresponsibility
5. Misuse
6. Technical Opacity
7. Incidents & Accidents
: Challenge
: … contributes to …

3.2.1 Infrastructural aspect

Summary

We identify four distinct characteristics making GPAI akin to infrastructure for the digital economy. First, **built-in development decisions** can be hard to alter ex post via fine-tuning. Second, **large user bases**, which leads to monopolistic patterns and increases the severity of a flaw in the model provided. Third, **economic ubiquity** relates to the fact that many diverse economic activities build upon a few GPAI models. Last, **lock-in effects** and the difficulties to change GPAI providers or even develop in-house alternatives due to re-engineering costs and service disruption.

The large-scale adoption of general purpose AI models across various sectors results in a significant number of downstream deployers and users of the same model [93]. This widespread embedding in the economy means that any malfunction or miscalibration could have substantial ripple effects. The growing dependence on these models is akin to the indispensable role that infrastructure plays in the economy. This challenge manifests itself in four distinct facets:

- Built-in development decisions: once a GPAI model is pre-trained, it requires considerable resources to meaningfully overwrite its training, for instance, via an improved training source code. Fine-tuning and retraining, while common, are rarely done with sufficient rigour to constitute a significant overwrite, given the costs [63] involved in the pre-training. This is analogous to infrastructure projects, where it is often more economical to rebuild the project from scratch than to make major modifications in order to correct mistaken design decisions.

- Large user base: the significant costs associated with developing a GPAI model can only be justified if a sufficiently large B2C and/or B2B user base utilises it. This often means that a few dominant players capture a significant market share, mirroring the monopolistic tendencies seen in the utilities sector. Furthermore, the sizeable user base amplifies the model's impact: any flaw in the model can have far-reaching consequences, given its influence on a multitude of applications and sectors, and therefore affected individuals.
- Economic ubiquity: securing a user base requires catering to a wide array of uses and, consequently, integration within a broad range of sectors and tasks. Unlike most products covered under the AI Act, a GPAI model is not designed or developed with a specific "intended purpose”[9] or context of use. Rather, its purpose is to be general enough to accommodate as many use cases as possible [94].
- Lock-in: A high level of criticality of a GPAI model integrated within a service or product can lead to a "lock-in" situation for downstream value chain actors. That is, if they decide to switch to another GPAI provider or, worse, build their own GPAI model, they would face substantial re-engineering costs and service disruption, similar to the upheaval caused by finding an alternative energy grid or highway system. This situation is even more pronounced for API based models, where GPAI providers bear the cost of running the model, as opposed to with open source models.[10] This bundling of the model service with cloud usage means that if a downstream deployer wanted to switch from an API to an open-source distribution channel, they would bear the hefty training and deployment costs. Moreover, within an API framework, downstream deployers cannot modify the model weights as they are safeguarded by the provider.

3.2.2 Generalisation & capability risks

Summary

To better penetrate the market, several **GPAI providers** are pursuing **increasing generality** or even AGI with every iteration of their products. This not only disrupts the market, but also tip political scales in an unprecedented manner. Three types of risk arise from this race towards generality. First, **capability risks** refer to the **increased autonomy** derived from more powerful models, increasing the odds of deceptive, manipulative, or mis-aligned behaviours. Second, **societal risks** include the danger of **enfeebling human capacities**, **political influence**, and the **embedding** of certain **values** during the development stage of GPAI, creating echo chambers. Last, **Extinction Risks** address the possibility of an existential threat to humanity posed by increasingly powerful but untrustworthy GPAI.

[9] ‘Intended purpose’ in the AI Act means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation.

[10] For a discussion of the difference between open source and API-based business models, see [93].

In order to secure the large user base necessary to make the infrastructure-scale investment worthwhile (Challenge 1), GPAI providers aim to make each version of their models more general than the previous ones. This drive for increased generality is not merely driven by market demands, but also reflects a socio-cultural fascination [95] with [96]AGI. In essence, AGI refers to a GPAI system or a combination of GPAI systems capable of all economically valuable human tasks [97] [98]. The development of such systems, to the extent they are controllable (cf. Challenge 6 "Technical Opacity"), would not only provide market power of unprecedented scale, but also political power, enabling the fulfilment of their developers' ideological and personal ambitions [85].

Regardless of whether or not AGI is achievable or permissible, three types of risks arise from ever-increasing generality prior to any successful AGI development:

- Capability risks: as the model becomes more general, it also grows more autonomous, with each instruction capable of being executed by leveraging a broader range of capabilities, such as coding, web interactions, planning, psychology and so forth - as outlined in the previous section. This makes certain behaviours like deception (both accidental and intentional), manipulation, self-replication, mis-aligned strategy and planning, cyber-offense, or AI programming more likely [84]. This wide range of possibilities is a consequence of the model's lack of boundedness (i.e. the inability of the developer to constrain the output or behaviour ex ante), goal misgeneralisation [99] and specification gaming [100].
- Societal risks: The rapid integration of GPAI models into our lives could lead to a number of systemic risks. Firstly, the widespread use of highly-capable GPAI could lead to "enfeeblement" [96], that is, the degradation of expertise that can result from humans increasingly delegating important functions to machines. In this situation, society or specific communities risk losing their capacity to self-govern and become wholly reliant on AI. Secondly, GPAI models could revolutionise how states and organisations deploy technology for political influence, potentially enabling personalised disinformation campaigns and generating emotionally charged arguments, thus risking collective decision-making, individual radicalization, and the erosion of shared realities. Finally, when developers embed specific values and principles into a GPAI model, it poses the risk of centralising ideological power [102]. Such a centralised viewpoint may produce models that are not agile enough to adjust to dynamic and diverse societal perspectives, or create further echo chambers. This is different from the less deliberate systemic bias and failure at inclusiveness amongst developers of these systems, described in Challenge 4 (Corporate Irresponsibility). The MEPs working on the AI Act[11] and US President Joe Biden[12] have also expressed these concerns.

[11] https://x.com/IoanDragosT/status/1647920290737823746?s=20

[12] WATCH: Biden exhorts world leaders to stand up to Russia at 2023 United Nations General Assembly | PBS NewsHour

- Extinction risks: As explained in the previous section, several GPAI developers have raised alarming concerns about the possibility of these models posing an existential threat to humanity or presenting catastrophic risks, accidentally through misalignment of the AI's objectives with fundamental values. This is in addition to numerous experts [103] [104] who have explored [101] the implications of the technology and concluded with similar or even greater levels of alarm. UN Secretary General Antonio Guterres[13], the UK PM Rishi Sunak[14] and, more recently, Commission President Ursula Von Der Leyen[15] have also expressed concerns about existential or extinction risks.

3.2.3 Concentration of Power

Summary

The economic and technological entry barriers to develop state of the art GPAI has heightened the risk of **monopolistic patterns**, narrowing the space to a handful of players and increasing the risk of **geopolitical instability**. Strategic partnerships are contributing to **power concentration**, and talent attraction and data acquisition are exhibiting **network effects**. Moreover, GPAI firms are **integrating** both vertically and horizontally, controlling funding and access to chips and, in turn, the **entrance into the market**.

The significant investment required for GPAI development (Challenge 1, infrastructural aspect) has heightened the risk of monopolistic tendencies, narrowing the field to a handful of well-funded firms. This trend of industry consolidation is evidenced by the uptick in acquisitions and partnerships, a concern even recognised by the US Federal Trade Commission [105].

Strategic partnerships, often accompanied by equity investments, allow GPAI labs to access the resources of large cloud computing providers. In return, computing providers gain privileged access into these labs' AI models, acquire equity in those companies, or both. This symbiotic dynamic is evident in the alliance [106] between Microsoft and OpenAI, the partnership [107] between Anthropic and Google, and is set to become a recurring theme.

The creation of a model with sufficient generality to profitably serve as an infrastructural element requires significant economies of scale. Both the computational resources and the rare talent needed to develop these models demand significant upfront investments. Furthermore, data acquisition and talent attraction exhibit network effects. That is, having one cutting-edge product fuels a virtuous cycle, generating more data and attracting more top-notch talent to train the next version of the model.

[13] Secretary-General's remarks to the Security Council on Artificial Intelligence
[14] UK should play leading role on global AI guidelines, Sunak to tell Biden | Artificial intelligence (AI) | The Guardian
[15] https://twitter.com/EU_Commission/status/1702295053668946148?s=20

The dominance of elite technology firms in the creation and control of advanced AI technologies can lead to an unhealthy concentration of power, potentially leading to harmful monopolistic behaviours or geopolitical instability. Presently, the situation resembles an oligopoly, with a handful of companies - namely OpenAI, Google Deepmind, Anthropic and their main partner or owner Microsoft and Google - dominating the field, wielding significant market power and political influence. A growing number of European software developers' businesses depend on reliable access to the GPAI value chains controlled by these US-based technology companies[16] and the lives of millions of EU citizens are, or soon will be, affected by these GPAI systems.

The problem of monopolistic control is exacerbated by both vertical and horizontal integration between AI firms. Vertical integration is demonstrated by Google Cloud providing cloud computing resources to other Alphabet subsidiaries Google Brain and Deepmind [108], as well as Anthropic [109]. Additionally, Google has invested at least $300 million into Anthropic [107]. Similarly, Microsoft now owns a 49% stake [110] in OpenAI and provides all of its compute via Microsoft Azure [106], whilst Amazon's recently announced AWS Bedrock combines LLM models from Anthropic, Stability AI and AI21 labs with cloud computing resources from Amazon Web Services to allow downstream developers to deploy these GPAI models [111]. New entrant to the GPAI industry Inflection AI raised [112] $1.3 billion from investors including Microsoft and Nvidia, whilst an array of tech titans, including Google, Amazon, Nvidia, Salesforce, AMD, Intel, IBM and Qualcomm, recently invested a combined $235 million in Hugging Face [113]. Instances of horizontal integration include the merger of Deepmind with Google Brain [114].

Consequently, in the current GPAI landscape, it appears that industry leaders often are, or receive significant backing from, established big tech giants with already entrenched advantages in the form of liquidities or in-built ecosystems granting them access to essential models or hardware components like data and computing power. Google owns a wealth of training data via GSuite, complemented by its computational assets in Google Cloud. Microsoft's rich data sources include LinkedIn, Outlook, Microsoft Office, and Github, fortified by its computational platform, Azure. Meta - arguably a smaller player, one rung below the likes of OpenAI and Google DeepMind in terms of their model capabilities - has a vast amount of training data at its disposal for Meta AI thanks to its many social networks such as facebook, instagram, and possibly others. NVIDIA, meanwhile, enjoys a virtual monopoly in high-end GPU production, supplying the trifecta of cloud computing titans: Amazon [115], Microsoft [116], and Google [117].

Even "new entrants" to the GPAI industry, such as Inflection AI or xAI, benefit from investors' significant scale in various production factors. Specifically, Inflection AI enjoys access to the computing power provided by its investor NVIDIA [112], likely data from its investors (founder of

[16] For example, Germany's AskBrian, Sweden's Sana Labs, Spain's Vizologi, and Belgium's Waylay[8] OpenAI Codex Live Demo; [2107.03374] Evaluating Large Language Models Trained on Code [9] Moreover, the same model fine-tuned for processing text-image pairs (DALL•E) unexpectedly learned the ability to manipulate photographic viewpoints & lighting, three-dimensionality, internal vs external structures, geographic and historical knowledge, as well as creative and imaginary compositions.

LinkedIn) and Microsoft, and a ready supply of talent through the reputation of its founder, Mustafa Suleyman. Similarly, Xai leverages the prestige associated with Elon Musk, along with data [118] from Tesla, SpaceX, and Twitter. These rare assets enable these newcomers to exceptionally compete in a market dominated by established players.

3.2.4. Corporate Irresponsibility

Summary

There are numerous instances of the few players at the top of the food chain of GPAI indulging in **reckless behaviours** towards users or individuals impacted by AI. A cultural **lack of attention to compliance** and **risk management** has led to incidents. The **race dynamics** that AI organisations are fostering by trying to outpace their competition is usually **at the expense of thorough evaluation**. GPAI providers have frequently silenced internal opposition and are lobbying hard to avoid external scrutiny, while their irresponsible behaviours range from **live-testing their models on vulnerable groups and data laundering, to offering "research previews" and disclaimers to deflect responsibility and circumvent oversight**.

As a result of the infrastructural aspect (challenge 1), the psychological, economic and political power through generalisation (challenge 2), and the limited market competition (challenge 3), the few players at the top of the food chain of GPAI development have already indulged in reckless behaviour towards users or individuals impacted by an AI system. There is often a lack of attention towards proper compliance and risk management within certain tech-focused regions, as seen in Silicon Valley. This cultural inadequacy, exacerbated by the frenetic pace of innovation, often results in lapses in adherence to product safety laws and cybersecurity protocols.

This corporate irresponsibility manifests itself in several ways. The first is the phenomenon of "race dynamics", whereby AI companies strive to outpace their competitors even at the expense of thorough evaluation and safety measures. This often comes hand-in-hand with the culture of hype surrounding new technologies, which can overshadow the need for responsible deployment and usage. Finally, there is a concerning trend of leading AI firms paying lip service to the warnings of existential risk, while recklessly forging ahead towards those very dangers.

The inherent concentration present within the GPAI industry (as detailed in challenge 3) particularly exacerbates this phenomenon of corporate irresponsibility and severe lack of accountability towards society. This issue would undoubtedly be far less prevalent if GPAI providers maintained rigorous internal accountability mechanisms. However, there is a growing pattern of stifling internal criticism and dissent. This is evidenced by countless examples such as Timnit Gebru's contentious ousting [119] [120] from Google over a paper on the risks of large language models, and the Amodei siblings' departure [121] from OpenAI (along with nine other employees) to create Anthropic due to differences [122] in strategy emerging with Microsoft's investment in the OpenAI. Concurrently, Microsoft chose to disband [123] its AI ethics and society team, even as it funnelled considerable resources into AI development through OpenAI. Additionally, Geoffrey Hinton decided to resign [124] from his position at Google in order to "freely speak out about the risks of AI".[17] This lack of internal opposition makes GPAI providers poor wardens of societal values that their concentration of power implies. This silencing of opposition is not limited to internal stakeholders: despite public statements implying their desire to be regulated, GPAI providers are lobbying worldwide to avoid democratic accountability and regulatory oversight.[18][19]

This lack of corporate responsibility is epitomised by the tech industry's mantra to "move fast and break things", resulting in a disturbing tendency to use society, particularly vulnerable groups, as guinea pigs for untested technologies. A case in point is the Microsoft chief economist's controversial claim at the Davos World Economic Forum that suggested more casualties were needed before justifying AI regulation [126]. This practice also results in the exploitation of users, especially in regions that may lack the regulatory infrastructure to protect them. A telling instance is Microsoft's decision to covertly test their GPT4-enabled Sydney chatbot on users in India as early as November 2022, 4 months prior to publicised beta-testing on the rest of the world. The insulting messages reported by one user [127] were eerily similar to the numerous accounts of unhinged behaviour [128] exhibited by the Bing chatbot released in early 2023, suggesting that even after over two months of testing on vulnerable groups in developing countries, Microsoft had not learnt from its mistakes.

The willingness of GPAI providers to exploit vulnerable groups is also palpably demonstrated by their decision to shift the onus of model fine-tuning onto outsourced workers. This technique, called reinforcement learning with human feedback (RLHF), requires workers to "teach" models to filter out harmful or offensive content from their output. Constant exposure to such distressing content has been shown to have detrimental effects on the mental well-being of the workers involved [129]. Outsourcing this task conveniently distances the tech giants from these harmful working conditions. OpenAI, for example, outsourced such jobs to Kenyan workers, subjecting them to distressing content for paltry wages and inadequate support [130]. Moreover, GPAI providers also exploit large groups of the population that generate "data" protected by copyrights (e.g. artists) or use illegal data (e.g. child pornography), sometimes instrumentalizing non-profits and academic research to obtain such data with little oversight.[20]

A particularly egregious case of misuse has been demonstrated by Meta's handling of their LLaMA GPAI model series. In 2023, within a week of Meta distributing the model to a supposedly

[17] Note that Trust & Safety lead of OpenAI, Dave Millner, also left the company in July 2023 [125].

[18] Report details how Big Tech is leaning on EU not to regulate general purpose AIs | TechCrunch

[19] Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI

[20] AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability - Waxy.org

trusted group of researchers, the model was leaked online [50], raising questions regarding Meta's commitment to securing their models and demonstrating how easily a leak, hack or theft can result in harmful actors or foreign states accessing millions of dollars worth of intellectual property. The controversy triggered bipartisan questioning of Meta by US legislators [131]. To make matters worse, Meta then went on to partly open source [132] a pre-trained version of the second, more powerful generation of the model, Llama 2, allowing unsecured access to the model weights, thus enabling ordinary members of the public to fine-tune the model, including for malicious purposes.

Finally, tech companies frequently rely on "research previews" [26] and disclaimers to deflect responsibility for the potential impacts of their products. This makes some of their most significant risks invisible to consumer protection, product safety and risk-based regulations like the AI Act. There have been additional concerns about unilateral deprecation of older models, to the frustration of many downstream developers relying on it for business, and unilateral sudden limiting of access to GPT4 via ChatGPT (from infinite to 25 messages per 3 hours), even for paying users. Finally, some developers rely heavily on making their product addictive (by making the chatbots answer as much as possible with questions) so that the user shares information and provides more data.[21]

3.2.5 Misuse

Summary

Insufficient oversight facilitates nefarious uses. The generality of **GPAI** makes it difficult to forecast all possible uses, amplifying in turn the **risk of intentional or accidental misuse**. However, misuse incidents have shown that such **models are rarely ready for consumer usage** and therefore should perhaps not be released into the market, and that better **allocation of accountability** throughout the value chain is needed.

The tangible consequences of the lack of corporate responsibility just discussed (challenge 4) are the misuse of GPAI models. Like many technologies, from hand tools to advanced bio-engineering, GPAI models, if not properly secured and regulated, can be exploited for nefarious purposes. Their inherent versatility and capability (Challenge 2), means that forecasting their full spectrum of potential applications is a formidable challenge, even for their creators. This unpredictability, coupled with the providers' negligence in their responsibilities to mitigating risks to society, (Challenge 4) amplifies the challenge of intentional misuse [133], or malicious use, beyond what is common for other technologies.

[21] InflectionAI's Pi.AI and Character.AI's chatbots are famous examples of this tactic, which results in enamoring and ex post feelings of being manipulated into a relationship [54].

There are pieces of knowledge and know-how that society deems too hazardous to disseminate, such as instructions for how to build a nuclear weapon or artificially engineer biological or chemical weapons [134]. Yet, the developers of these models are seldom held accountable for facilitating the proliferation of such dangerous information. Even when GPAI providers attempt to enforce content-moderation controls, these guardrails are susceptible to “jailbreaks”, either manually [135] through the use of special queries that can induce unintended responses [136] or, even more concerningly, via adversarial attacks [137] in an automated fashion. The ease with which these vulnerabilities have been exposed raises concerns about the safety of such models, especially as they start to be used in a more autonomous fashion, and strongly suggests that the models are not yet ready for market release. Unfortunately, best practices by one actor can be negated by the poor practices of another: by making some of the larger GPAI models accessible open source with guardrails, some careless industry players have enabled the creation of a universal jailbreak that can effectively affect all GPAI models, including otherwise protected API-based models [137].

Moreover, even accidental misuse -as opposed to intentional, malicious use- can result in significant damage. The vast majority of users lack a comprehensive understanding of the GPAI system and its associated risks. Worse, deployers embedding these GPAI models within their systems have also limited access to know-how. Combined with poor standards for technical documentation,[22] this results in deployers causing -unwittingly, and therefore rarely documented- incidents for affected subjects[138].

3.2.6 Technical Opacity

Summary

The significant influence that key AI developers have over academia is contributing to a ML paradigm that stands in opposition to scholar approaches to research, anchored in safety and trustworthiness. GPAI are **technically opaque by design**, and their high complexity introduces several shortages. First, a **lack of interpretability**, or the fact that the internal workings and “what the model thinks” is not fully understood. Second, a **lack of predictability**, reflecting how the model, its capabilities and outputs are highly unpredictable. Third, a **lack of corrigibility**, which derives from the combination of the former two shortages, resulting in notable difficulties to correct undesired behaviours. Last, a **lack of controllability**, i.e. the feedback loop between model and human operator is structurally ineffective.

[22] A review of over 400 generative AI companies’ “technical documentation” has shown that almost all of them only have marketing materials and do not abide by any technical documentation standards, with one dramatic example being a company describing their product’s workings in 3 steps, with “Apply magic” as a second step.

As a result of the clear lack of competition focused on product safety (Challenge 4), and the significant influence major players exert over academia (Challenge 3), the entire field of machine learning has become entrenched in a technical paradigm that is inherently antithetical to safety and trustworthiness. Moreover, the weights of these models being closely guarded trade secrets (amounting to tens of millions of dollars in development costs for the largest models), maintaining a level of technical opacity serves as a robust barrier to entry for individual, financially-constrained, innovators. Consequently, there is little incentive for large GPAI providers to promote transparency, even if it were technically feasible.

The techniques that underpin modern GPAI development are fundamentally rooted in probabilistic methods , as opposed to other branches of engineering which rely on well-established, deterministic scientific principles and models of the natural world. This complexity and unpredictability means that often even GPAI providers do not fully comprehend the decision-making processes of their models, let alone their capabilities [139] [140].

GPAI models are therefore opaque by design. They transform human-understandable data, such as text or images, into semi-interpretable tokens, and these tokens are further translated into vast arrays of numbers that escape human interpretation. Given the immense volume of data involved - far greater than what a human can interpret - and their automated conversion to an uninterpretable format, integrating human judgement and oversight seems incompatible with current GPAI methodology.

Thus, the model's inherent complexity, stemming from its handling of vast datasets, surpasses that of many contemporary sophisticated creations, including bio-engineered organisms and particle accelerators. This level of complexity introduces several challenges:

- Lack of interpretability: interpretability is the extent to which the internal workings of an AI model can be understood by humans. An interpretable model allows users to comprehend why and how it arrived at a particular outcome or decision. This characteristic is critical for trust, especially in high-stakes domains where understanding the AI's decision-making process is essential. With GPAI, the way the model “thinks” is not fully understood, even by its developers. In contrast to other engineering disciplines, where it is possible to understand instinctively the output or behaviour of a system given an input and a transformation process[23], in GPAI software engineering, the model itself is a black box: no-one can determine what a weight or neuron's role is, let alone a neural pathway. That is, the internal workings of the training process is not understood by humans. Even with rigorous post-training testing, understanding the reasons behind models' outputs remains a challenge. This also introduces GPAI-specific cybersecurity challenges, as lack of interpretability complicates identification of whether weights have been tampered with or whether the dataset has been poisoned.

[23] And good instincts on this are generally a key sign of expertise in the given field e.g. in project management, construction, manufacturing, business, etc.

- Lack of predictability: Partly as a consequence of the lack of interpretability, the model's capabilities and behaviour is to a large extent unpredictable by humans. Entire categories of output, such as modes of functioning and capabilities of the model, are frequently discovered post-deployment. For instance, GPT-3, designed to process natural language, unexpectedly showed the ability to write basic programs using programming languages, much to the astonishment and delight of its developers [141]. Worse, the model resulting from training is itself not predictable, though there are research efforts to simulate models without having to train them. It is important to note that predictability is related to but different from boundedness, another feature discussed in Challenge 2 (Generalisation & capability risks). Robustness (the ability of a GPAI model to perform consistently and accurately in a variety of different conditions, including those it was not specifically trained for) is currently the best way to contribute to predictability of AI models with specific intended purpose, but has turned out to be particularly challenging the more general a model is.
- Lack of corrigibility: The combined lack of interpretability and predictability makes it difficult to correct either the model or the development process to ensure that it is sound and meets specifications. Irrespective of the potential costs, such as having to re-write the training programme and re-train the model, as yet, there is an absence of clear guidance on how to correct these models in order to achieve the desired outcome.
- Lack of controllability: As a result of the lack of corrigibility, the GPAI model cannot be structurally controlled in the sense that the feedback loop between model and human operator is ineffective. Specifically, while the model and its output affects the human user, a human user or even a developer has limited ability to modify the model or its training programme in a timely, meaningful and sustainable fashion so as to impose its preferences upon the model - short of carrying out a new pre-training with a different dataset.

For all four sub-challenges, there is a nascent academic field developing experimental methods and approaches for potential solutions. Unfortunately, this field remains underfunded and investments in that direction are dwarfed by the speed and scale of investments in developing more general (and therefore more problematic) models. The industry still relies in large part on a process of costly and mostly ineffective trial and error, where the costs are borne by society, as opposed to investing in a fundamental solution or shifting paradigms.

3.2.7 Incidents and Accidents

Summary

As a result of technical opacity and lack of accountability, the risk of incidents and accidents has already materialised. First, GPAI perpetuates and amplifies societal biases, **resulting in systematisation of macro- and microaggressions for discriminated groups**. Second,

misinformation and privacy violation incidents, including the fabrication and spreading of misinformation, omission of crucial details, or the infringement of privacy rights. Last, **accidents** from **unexpected behaviours**, have occurred even months before deployment, during internal testing and live-testing on users. The risk of which is heightened in light of increasingly agentic models or models interfaced with real-world systems.

A leading GPAI provider admitted openly that "we do not know how to train systems to robustly behave well" [81]. And indeed, as a result of the technical opacity of GPAI models (Challenge 6) and the lack of accountability of the providers (Challenge 4), significant incidents have already occurred. Coupled with the concerns regarding risks from increasingly general systems, including AGI (Challenge 2), these occurrences pose a significant challenge in the GPAI landscape. Incidents stemming from the unpredictability and unreliability of these models can be grouped into the following three categories:

Bias and Discrimination: General purpose AI models, if trained on biased data from the internet, can perpetuate and amplify existing societal biases [140]. Such biases could unintentionally disadvantage certain groups based on protected characteristics, and could result in systemic discrimination against already marginalised groups. There's also the danger of these AI systems reproducing stereotypes in sectors like education, entertainment, and media [142]. For example, certain generative AI models, such as Stable Diffusion from Stability AI, have been shown to portray professions like doctors and judges in a manner inconsistent with population demographics, thereby misrepresenting and potentially reinforcing gender and racial biases [143]. This risk translates into automated and systematic occurrence of macro- and microaggressions, and to the extent these models are adopted widely in the economy and society, further institutionalises discrimination.

Misinformation and Privacy Violations: Due to their unreliability and inscrutability, these AI models can spread misinformation, omit crucial details, or even disclose true information that infringes on privacy rights. The risk is particularly heightened in critical domains like medicine or law where inaccuracies can have profound implications. Models can produce content referred to as "hallucinations" that sound authoritative but are completely fabricated. Examples include OpenAI's ChatGPT making false accusations [144] and Meta's Galactica producing incorrect information [76]. Furthermore, general purpose AI models can potentially leak sensitive personal information present in their training data, violating privacy norms.

Accidents: These models can exhibit unpredictable behaviour leading to potential accidents, especially if interfaced with real-world systems. This unpredictability could arise from unusual inputs (like glitch tokens or adversarial examples), or from models trying to achieve defined goals in unexpected ways (reward misspecification errors). An example of this is an algorithm producing buggy code [145] or behaving in a way that was not originally intended. Increasingly agentic AI models, designed to be more autonomous, can act with less human oversight, leading to serious negative outcomes [146]. There's also concern over models developing instrumental goals that involve manipulative or threatening behaviours, which can be harmful in real-world deployments. An illustrative experimental instance of this is GPT-4 pretending to be a visually impaired human to bypass a CAPTCHA [91]. As mentioned in section 3.1, these accidents also occur during development, internal testing stage, red-teaming and RLHF stages, i.e. months before deployment.

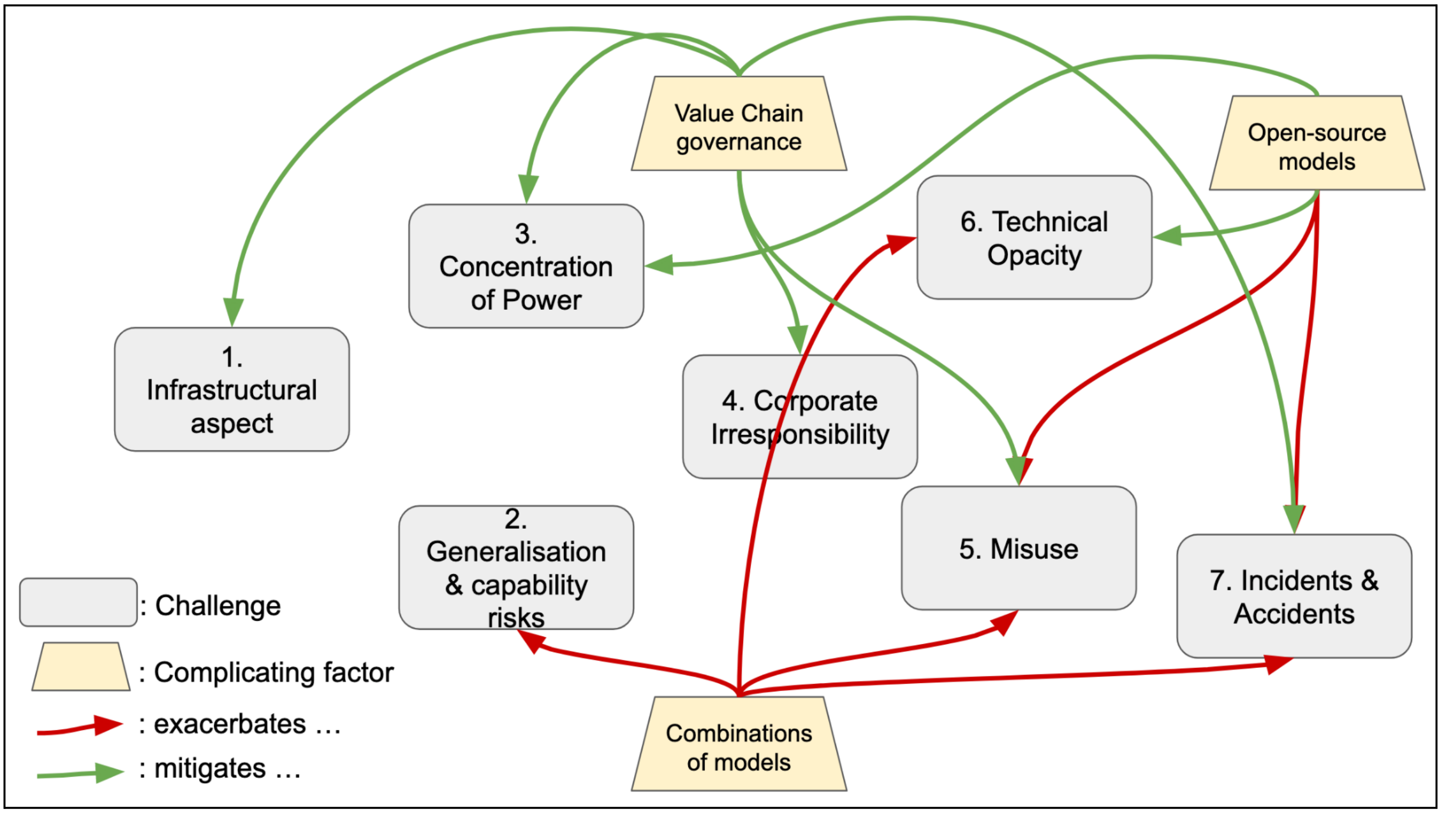
4. Complicating factors

Summary

We discuss three main complicating factors. First, **the combination** and **interaction** of different **models** accentuates most challenges. Second, the **open-sourcing of GPAI models** can mitigate some challenges (Concentration of Power and Technical Opacity) but significantly exacerbate others (Misuse and Incidents & Accidents). Finally, **governing the value chain** is crucial to re-balance power dynamics and obtain corporate accountability - it also plays a pivotal role in establishing a level playing field for all the actors involved.

As a consequence of the challenges and technical characteristics described in previous sections, the landscape of GPAI models involves unique features that demand focused regulatory attention as they exacerbate or mitigate some of the challenges identified. Addressing these complicating factors will be paramount to the evolution and safe integration of GPAI in society. Figure 3 summarises the relationships between complicating factors and challenges.

Figure 3: Complicating factors in the GPAI industry



As models interact, either by design or inadvertently, their combined functionality can intensify inherent risks and unpredictability. Such interactions not only magnify individual flaws in the system, but also introduce emergent behaviours that are harder to anticipate or control. Regulatory initiatives will need to target accidental interactions and combinations of models to ensure predictable and safe outcomes.

Additionally, the open-sourcing of GPAI models presents both opportunities and threats. While it democratises development of the models and could address concerns like monopolistic tendencies or technical opacity, it equally raises the likelihood of misuse, making these models potential tools for malicious activities. As such, open-source models must be approached with a balance between accessibility and security.

Finally, value chain governance in the GPAI industry underscores the need for equitable power dynamics and corporate accountability. The prevailing concentration of power among upstream players in the industry contrasts starkly with other regulated sectors, highlighting the urgent need for proper documentation and liability. Value chain governance will be pivotal in establishing a level playing field and ensuring that all actors in the chain bear an appropriate degree of responsibility for the impact that the deployed model has on society.

4.1 Combinations of models/interacting models

Interactions between GPAI models, whether designed to interact or accidentally doing so, can magnify the inherent risks associated with individual models and introduce new complexities. Specifically, combining models can amplify the challenges of generalisation (Challenge 2), as the merged models might demonstrate heightened autonomy and unpredictability.

Furthermore, the enhanced potency and versatility of combined models amplifies their potential for misuse or malicious exploitation (Challenge 5), as new, unanticipated vulnerabilities emerge from these interactions. Coupled with their increased complexity, these interactions may prove more elusive, making detection and prevention by authorities significantly more challenging. Consequently, these combined systems could become a more tempting target for malicious actors, who could exploit this intricate system for harmful purposes.

Technical opacity (Challenge 6) becomes even more profound, as understanding intertwined systems is considerably more challenging than interpreting individual models. Interactions between models could lead to behaviour that is difficult to predict. This unpredictability is heightened when multiple models are combined, which could lead to snowballing effects where a small error or anomaly in one model is amplified across the combined system.

Lastly, the probability of incidents and accidents (Challenge 7) escalates, as errors or biases in one model can cascade or compound when interacting with another, leading to unforeseen and potentially catastrophic results. Interactions between models often occur in an autonomous manner, in the absence of a ‘human in the loop.' This means actions may be taken, or decisions made, without human oversight, which could increase the risk of unintended consequences.

4.2 Open-source Models

The open sourcing of GPAI models presents a double-edged sword, providing both promise and dangers for the AI community and society as a whole. On the one hand, open-source models democratise access to advanced AI capabilities, fostering a culture of shared knowledge, innovation, and community-driven enhancements. They offer potential solutions to some of the challenges mentioned, such as mitigating monopolistic tendencies (challenge 3) by decentralising AI development. Additionally, they can indirectly address the issue of technical opacity (challenge 6), as the transparent and collaborative nature of open-source projects can drive research towards better interpretability.

On the other hand, unrestricted access to open-source GPAI models can amplify the issue of misuse (challenge 5) and incidents or accidents (challenge 7). With unrestricted access to model weights and architectures, bad actors can easily bypass safety measures set by developers and tailor these models for malicious ends. This might include disabling toxicity safeguards or adapting the base model to produce harmful content or misinformation, propagate biases, or even design advanced malware.

In essence, the open sourcing of GPAI models brings to light the pressing need to strike a balance between fostering innovation and ensuring security and responsibility. While open sourcing in the software realm fosters community-driven research and vulnerability detection, the heightened risk of abuse in the AI sphere suggests the need for extra precautions or governance measures, such as limiting access to unsecured model weights. Moreover, further complicating the approach to open source models is the disagreements within the open source community over whether these “open source” models are truly open source [147] [148]. The need to establish a more granular classification than merely “open” versus “closed” source indicates that the governance of open source will benefit from competent regulators dialoguing with and observing the open source community.

4.3 Value Chain Governance

The concentration of power (challenge 3) within the GPAI industry has led to upstream oligopolies wielding considerable bargaining power. As described above, this bargaining power may manifest itself in the offloading liability onto downstream actors, in contract negotiations where upstream suppliers adopt a "take it or leave it" stance or through withholding technical documentation. This generates unique issues along the value chain of GPAI models.

There exists a prevailing assumption that the digital realm remains beyond the need for regulatory oversight or intervention (challenge 4). Unlike the GPAI sector, industries such as automotive have established and formalised data-sharing practices. Every participant in the automotive value chain shares the common objective of preventing accidents, such as car crashes or recalls. In contrast, even the intrinsic motivation to safeguard their reputation doesn't seem sufficient for some GPAI companies to adopt prudent policies. An aeroplane accident could bankrupt an airline, yet the providers of the chatbot that convinced an EU citizen to commit suicide are yet to be held accountable [60].

Effective value chain governance is therefore vital in order to prevent a lack of corporate accountability within the GPAI industry. As a result of the technical opacity of these models (challenge 6), it is virtually impossible for their downstream deployers to reliably predict what their outputs will be. This makes mitigating potential malicious uses or unintentional incidents resulting from their deployment even more challenging. The provision of comprehensive documentation, including data sheets, model cards, and clear usage instructions, from GPAI providers will endow downstream deployers with a deeper understanding of the tools they are testing, refining, or fine-tuning, thus increasing the safety of the system and allowing for its joint liability.

Resolving cost-effectively the problem of value chain governance is therefore crucial: it will be essential in order to ensure users dependent on this infrastructure are not exploited (Challenge 1), rectify the power imbalance (Challenge 3), make it more difficult as an industry player to behave irresponsibly, through greater transparency (Challenge 4), enable mitigation and prevention of misuse and malicious uses more cost-effectively (Challenge 5) and facilitate prevention and response to accidents (Challenge 7). Note that this has been a significant issue throughout the AI Act, as explained in Section 7.4.

5. Defining General Purpose AI for governance

Summary

Different challenges apply differently to different sub-categories under the umbrella term "GPAI", **based on the generality of the GPAI model's capabilities**. As the AI Act should address these challenges in a proportionate and risk-based manner, we distinguish between 3 categories and provide operational definitions. In brief, **Generative AI applications** are implementations of AI techniques for the purpose of producing new content. **Type-I General Purpose AI models** are AI models that are designed for generality of capabilities. **Type-II General Purpose AI models** are AI models that are designed for generality of capabilities and expand the technology frontier relative to 2022 models.

Table 2: Full definitions & examples

Name	Definition	Examples
Generative AI application	Implementations of AI techniques for the purpose of producing new content such as images, text, videos and audio based on an input "prompt" that is generally simpler than the output, such as a text sentence to generate an image. In the current technological paradigm, the generation of convincing output is often achieved through training on sufficient amounts of data using sufficient compute. As a result, many generative AI applications are built on top of one of a few General Purpose AI models, refined for a specific purpose through prompt-based training, fine-tuning, and reinforcement learning with human feedback. They are distinguished from discriminative AI applications or classifiers, which do not generate content as output but instead classify input data.	chatbots (e.g. Replika), automated text-to-voice generation (e.g. Voice AI), computer code assistants (e.g. Tabnine's), and social media content generation (e.g. tavus.io's)
Type-I GPAI model	AI models or combinations of AI models that are designed for generality of capabilities. In the current technological paradigm, generality of capabilities is often achieved through self-supervised learning, requiring significant quantities of data and computing power to estimate a significant number of internal model parameters, in order to train the model to an acceptable standard. It is also possible to combine GPAI models with each other to create new, often more powerful or more generalised, models. As a result, they are characterised by a wider range of capabilities than that of other AI models, including capabilities they were not explicitly designed for, and can therefore be applied to many distinct types of tasks.	Adept AI's ACT series, AI21 Labs Jurassic series, Aleph Alpha's Luminous series, AWS Amazon Titan series, Cohere's Generation series, Conjecture's model series, EleutherAI's GPT-Neo/J series, Google & Google DeepMind's Gato, Gopher and PaLM-E, Meta AI's CICERO, Microsoft's Kosmos-1, OpenAI's GPT series, Stability AI's Stable Diffusion series, and Tesla Bot's underlying model.
Type-II GPAI model (i.e cutting-edge)	AI models or combinations of AI models that are designed for generality of capabilities and expand the technology frontier relative to 2022 models. Like Type-I GPAI models, in the current technological paradigm, this often requires a significant amount of data and compute power. The expansion of the technology frontier relative to 2022 models has been achieved through unprecedented amounts of compute power, combinations of models and other architectural insights to maximise generality of capabilities.	OpenAI's GPT-4 & upcoming GPT-5, Anthropic's Claude, Google's LaMDA, Google DeepMind's Gemini, Meta AI's LLaMa, Microsoft's GPT4-Prometheus, InflectionAI's PI, and Stability AI's Stable Diffusion 2.0 & XL.

There is no common definition of GPAI across the scientific literature.[24] A review of 28 definitions put forward since 2021 in the academic and policy literature has highlighted the lack of convergence, as these definitions often conflate the concepts of foundation models, generative AI and general purpose AI systems; defining them by their functions or technical characteristics.[25] Technically, current GPAI models are characterised by their scale (measured in number of parameters[26]), as well as their reliance on self-supervised learning methods, transformer architectures, transfer learning, and both context-dependent and context-independent memory, which enable them to emulate some aspects of human cognition (e.g. attention and learning generalisation). Functionally, GPAI is characterised by its widespread use as pre-trained models for other AI systems. For example, a single GPAI for language can be used as the core for several hundreds more applied models simultaneously[27] (chat bot, ad generation, decision assistant, spam bots, …), some of which are subsequently further fine-tuned into multiple applications tailored for the customer. As a result, the seminal paper on the topic found that just a few GPAI models are the basis for almost all applied models for language.[28] In the context of a governance regime, a clearer definition of the concept is necessary to ensure legal certainty. We therefore attempt this, in light of the previous sections.

5.1 Distinguishing concepts based on challenges they pose

The conflation of concepts encompassed by GPAI (foundation models, generative AI, general purpose AI models and systems, etc.) is particularly problematic because the challenges outlined in the previous section do not apply to the same extent for each concept:

- Challenge 1 (Infrastructural aspect) is not as relevant to fine-tuned tools and services with more specific intended purpose than most GPAI models, because they only exhibit high levels of economic ubiquity when considered collectively, and are therefore less likely to result in a “lock-in” effect than a single, highly-generalised GPAI model deeply embedded within the economy.
- Challenge 2 (Generalisation & capability risks) is relevant to a given GPAI only to the extent its developers design it so as to increase the generality of its capabilities. On the other hand, some Generative AI tools are designed to generate high-quality content of a

[24] “Defining *General-purpose* AI systems in the AI Act” (April 2022) briefing by The Future Society, which beyond the lack of definition notes that the most thorough -though exploratory- discussion of GPAI can be found in [140]

[25] “Definitions: GPAI vs Foundation Models vs Generative AI” (Forthcoming) briefing by The Future Society

[26] Parameters in AI are the elements of the program that are tailored automatically through training. A GPAI model has several billion parameters or more. The biggest GPAI models available in 2023 have over 1 trillion parameters (Switch Transformer: 1.6 trillion; Wu Dao 2.0: 1.75 trillion). However, there is significant research into training GPAI models that perform as well as or better than the bigger models with a fraction of the parameters through information-retrieval or mixture-of-experts modules [149] [150] [151].

[27] For example, GPT-3 is estimated to be used by “tens of thousands of developers” [152] and in over 800 apps currently placed on the market covering over 220 intended purposes [153]. Similarly, the original version of BERT (the model underlying Google Search in 70 languages) has been “forked” into at least 9,000 models [154]

[28] *“Almost all state-of- the-art NLP models are now adapted from one of a few foundation models.”* (p. 5 in Bommasani, Liang et al. (2021) [140]

very specific type, through retraining or fine tuning more general purpose base models, in a way reducing the base model's perceived generality. That is, the goal of a GPAI developer is to maximise the generality of capabilities, but this is not necessarily true of all generative AI providers.

- Challenge 3 (Concentration of Power) is not as relevant for fine-tuned tools and services with more specific intended purposes, because the barriers to entry for the "fine-tuning" industry are relatively low. The issue of power concentration is more pronounced at the top of the value chain for GPAI models, with a handful of large tech firms such as Microsoft, Google and Meta dominating the field. However, there are already hundreds, if not thousands, of downstream deployers of these models that can be classified under the Generative AI umbrella, for whom this challenge isn't a significant concern.
- Challenge 4 (Corporate Irresponsibility) appears universally applicable across industry niches, albeit in different ways:
    - In fine-tuned downstream tools, the corporate irresponsibility comes in the form of misleading technical documentation, which can inflate customers' expectations and contribute to the overhyping of what the tools can actually achieve. ; Additionally, there is the, often intentional, opacity as to what third party GPAI model the tool is built upon, to maintain an illusion of in-house technical expertise in order to justify premium charges.
    - For foundation or GPAI models, the concern lies in the developers' active participation in raising customers' expectations about the capabilities of the models while simultaneously shielding themselves from liability through contractual terms that provide no assurances of fitness of purpose (in open source licences, but also beyond that) and disclaimers indicating the customer pays for a "research preview".[29]
    - In the cutting-edge GPAI models, corporate irresponsibility is evidenced by the developers' acknowledgement or even active warning to regulators of the great dangers posed by the increasingly capable technology. Yet, paradoxically, these same developers continue to aggressively compete in the race for increasingly capable models or AGI, thus actively developing even more capable and dangerous models.
- Challenge 5 (Misuse) is particularly concerning for models with a wide range of applications, though it remains a significant risk for those designed for specific intended purposes. This challenge is even more acute for open source models compared to those gated within an API. This stems from the fact that individuals with sufficient expertise can exploit unrestricted access to the model's weights, parameters, and in some cases, training data to fine-tune the model to exhibit dangerous characteristics. On the flip side, those with insufficient expertise also accidentally misuse the model.

[29] At time of writing (August 2023), Dall-E 2 is still considered in public "beta testing" [155] despite having paying customers. ChatGPT until July 2023 was still considered a research preview, despite having >100mn users since January 2023.

- Challenge 6 (Technical Opacity) is particularly acute for cutting-edge models where only a select group of engineers understand the model architecture and are familiar with the model's intricacies. Nevertheless, given that the majority of generative AI applications are built upon these state-of-the-art models, the effect of their opacity is amplified throughout the value chain.
- Challenge 7 (Incidents & Accidents) is problematic at all technological levels. However, the impact of the accidents is proportional to the number of downstream users and to the generality of the model capabilities. In particular, for cutting-edge (most general) models that pose risks of catastrophic accidents or extinction events, accident prevention should be emphasised as opposed to incident response for less capable models.

The nuances found for each challenge's applicability to various technological concepts (generative AI tools, foundation models, and cutting-edge general purpose models), we distinguish between 3 categories, which do not present the same risk profile and encompass different sub-concepts:

- 1. General purpose AI models that are designed for generality of capabilities and lie at the cutting-edge of today's technological development.
- 2. Other general purpose AI models that are designed for generality of capabilities but do not lie at the cutting-edge.
- 3. Generative AI tools and applications which are not designed for generality of capabilities but instead for automatically generating content in a more narrow category, often building upon a general purpose AI model through fine-tuning and feeding it task-specific data.

General Purpose AI is more general, requires specific resources and technical skill sets increasingly concentrated in a few teams around the world, and are often re-used in hundreds or even thousands of downstream applications, including Generative AI. Generative AI, on the other hand, is technically defined by its content generation potential, not its generality of output. Generative AI, as an industry, has come to refer to the application of AI-enabled content generation into specific tools or to specific purposes, such as e-commerce chatbots.

Models, systems or applications? We use the term GPAI *models* as opposed to GPAI *systems* because GPAI systems capture both GPAI models and many generative AI applications, in different ways:

- An AI system is defined as *an engineered system* that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives.[30]
- A model, on the other hand, is the mathematical or otherwise *logical representation* of a phenomenon, including relations between items. In the context of AI, a model is the representation of the properties of a given dataset, deduced by running an algorithm

[30] Definitions from ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology [156]

using compute. For example, a language model is a representation of language i.e. a list of probabilities that describe the probability of any word coming up after any given combination of words, for any given combination of words.[31] It is often the core component of an AI system that leverages that model to turn input .

- An AI application is an *AI system* that fulfils a certain function or task.

We therefore do away with this intermingling of concepts by using “models” and “applications”.

5.2 Definitions

For the operationalisation of governance, we define each category as follow:

A. **Generative AI applications**: Implementations of AI techniques for the purpose of producing new content such as images, text, videos and audio based on an input “prompt” that is generally simpler than the output, such as a text sentence to generate an image. In the current technological paradigm, the generation of convincing output is often achieved through training on sufficient amounts of data using sufficient compute. As a result, many generative AI applications are built on top of one of a few General Purpose AI models, refined for a specific purpose through prompt-based training, fine-tuning, and reinforcement learning with human feedback. They are distinguished from discriminative AI applications or classifiers, which do not generate content as output but instead classify input data[32].

   Typical examples of generative AI applications include chatbots [157] , automated text-to-voice generation [158] , computer code assistants [159] , and social media content generation [160] .

B. **Type-I General Purpose AI models**: AI models or combinations of AI models that are designed for generality of capabilities. In the current technological paradigm, generality of capabilities is often achieved through self-supervised learning, requiring significant quantities of data and computing power to estimate a significant number of internal model parameters, in order to train the model to an acceptable standard. It is also possible to combine GPAI models with each other to create new, often more powerful or more generalised, models. As a result, they are characterised by a wider range of capabilities than that of other AI models, including capabilities they were not explicitly designed for, and can therefore be applied to many distinct types of tasks.

[31] Technically, it’s not words but tokens, which sometimes include multiple words, phonemes and punctuation.

[32] For instance, email spam filters use AI to simply classify emails as “spam” or “not spam”, without generating any original content themselves

Typical examples for **Type-I General Purpose AI models** include Adept AI's ACT series, AI21 Labs Jurassic series, Aleph Alpha's Luminous series, AWS Amazon Titan series, Cohere's Generation series, Conjecture's model series, EleutherAI's GPT-Neo/J series, Google & Google DeepMind's Gato, Gopher and PaLM-E, Meta AI's CICERO, Microsoft's Kosmos-1, OpenAI's GPT series, Stability AI's Stable Diffusion series, and Tesla Bot's underlying model.

C. **Type-II General Purpose AI models**: AI models or combinations of AI models that are designed for generality of capabilities and expand the technology frontier relative to 2022 models. Like Type-I GPAI models, in the current technological paradigm, this often requires a significant amount of data and compute power. The expansion of the technology frontier relative to 2022 models has been achieved through unprecedented amounts of compute power, combinations of models and other architectural insights to maximise generality of capabilities.

   Typical examples include OpenAI's GPT-4 & upcoming GPT-5, Anthropic's Claude, Google's LaMDA, Google DeepMind's Gemini, Meta AI's LLaMa, Microsoft's GPT4-Prometheus, InflectionAI's PI, and Stability AI's Stable Diffusion 2.0 & XL.

6. Emergence of a tiered approach: from category to tiers

Summary

We propose a **tiered approach**, in which **requirements** for GPAI models and Generative AI applications are **set in proportion to their risk potential**, avoiding undue regulatory burden. The approach includes three tiers: **Generative AI applications**, **Type-I GPAI models**, and **Type-II GPAI models**. They are differentiated via a set of criteria for generality of capabilities before deployment, as this dimension correlates with the challenges & risks identified. The highest level of scrutiny is reserved for the models with greatest generality of capabilities, which pose the greatest potential risk and challenges: Type-II GPAI.

We put forward four different ways to assess a model's generality of capabilities. **First**, via the **amount of compute** used during training, measured in so-called "**FLOPs**". Second, via **skill-acquisition efficiency**, which describes the efficiency with which a system or individual can acquire new skills. Third, via a **generality analysis** that simultaneously evaluates the versatility and performance across tasks. Fourth, via **algorithmic efficiency and model perplexity**, which allow inferring generalizability as a function of a model's computational power, algorithm and the richness of its training data. Last, via **modality-specific benchmarks**. It is important to encourage industry to report and predict generality of capabilities in a consistent manner.

6.1 The tiered approach

As explained in the discussion of each challenge's applicability, there is a need for distinguishing between Type-I (akin to a typical foundation model) and Type-II (cutting edge general purpose models) because of the different risk profile: while some Type-I models have been deployed since 2019 and now mostly present risks related to their widespread embedding in economy, as described in the section analysing current public discourse on GPAI, Type-II General Purpose AI models advance the current state of the art, are not fully understood by their developers and pose larger potential risks in case of failures during development. Moreover, these models are being released publicly as a means of identifying model misbehaviour and bugs, essentially treating users as live, unpaid testers of the model, rather than demonstrating its safety "by-design" and robustness before public release.

These differences between Type-II GPAI models, Type-I GPAI models and Generative AI applications imply different actions that their respective developers can undertake in order to mitigate their different risks. Thus, a tiered approach to regulate (Type-I and Type-II) GPAI models and Generative AI systems ensures that requirements are set in proportion to their risk potential.

The highest level of scrutiny would be reserved for the most capable models, which pose the greatest risk: Type-II General Purpose AI models. As these Type-II models naturally also present Type-I models' risks after the deployment stage, they should also be subjected to Type-I models' requirements. Finally, since the main applications of Type-I and Type-II models are generative AI applications, they should also enable the compliance of downstream generative AI applications developers.

It should be noted that a fourth tier exists, situated between the Generative AI tier and Type-I GPAI tier. This tier is the European Parliament's "foundation model" regime in terms of regulatory target and risk mitigation measures. This additional tier helps further distinguish "foundation models" from the other two concepts. Foundation models would encompass all base models that can be applied to several tasks in a given modality, but require significant fine-tuning to be useful. This is different from generative AI because a foundation or base model can be applied to many tasks, but it is also less general and less ready-to-use than Type-I GPAI, which are useful even without fine-tuning. As the technical nuance is particularly difficult to make and necessitates judgement calls of what constitutes usefulness/usability, and given the politicisation of that term, we do not introduce this fourth tier in this blueprint.

This regulatory strategy shares its underlying logic with other regimes such as the Digital Service Act[33] or the U.S. EPA's Clean Air Act's Risk Management Programs for Chemical Accidents Prevention [162] [163]. This strategy has also been put forward several times in the discussion surrounding the AI Act's coverage of general purpose AI models, some experts distinguishing between systemic and non-systemic foundation models [164], between the most capable and other foundation models [94], or between strategically significant and other foundation models [165].

Table 3: A simplified tiered approach

Tier	Name	Definition	Examples
#1	Generative AI application	Implementations of AI techniques with the goal of producing new content such as images, text, videos and audio based on an input "prompt" that is generally simpler than the output, such as a text sentence to generate an image. In the current technological paradigm, the generation of convincing output is often achieved through training on sufficient amounts of data using sufficient compute. As a result, many generative AI applications are built on top of a General Purpose AI model, refined for a specific purpose through prompt-based training, fine-tuning, and reinforcement learning with human feedback. They are	chatbots (e.g. Replika), automated text-to-voice generation (e.g. Voice AI), computer code assistants (e.g. Tabnine's), and social media content generation (e.g. tavus.io's)

[33] The Digital Services Act (DSA) categorises and regulates 17 Very Large Online Platforms (VLOPs) and 2 Very Large Online Search Engines (VLOSEs) based on their total number of users [161]. The designation triggers specific rules that tackle the particular risks such large services pose to society when it comes to illegal content, and their impact on fundamental rights, public security, and wellbeing.

		distinguished from discriminative AI applications or classifiers, which do not generate content as output but instead classify input data	
#2	Type-I GPAI model	AI models or combinations of AI models that are designed for generality of capabilities. In the current technological paradigm, generality of capabilities is often achieved through self-supervised learning, requiring significant quantities of data and computing power to estimate a significant number of internal model parameters, in order to train the model to an acceptable standard. It is also possible to combine GPAI models with each other to create new, often more powerful or more generalised, models. As a result, they are characterised by a wider range of capabilities than that of other AI models, including capabilities they were not explicitly designed for, and can therefore be applied to many distinct types of tasks.	Adept AI's ACT series, AI21 Labs Jurassic series, Aleph Alpha's Luminous series, AWS Amazon Titan series, Cohere's Generation series, Conjecture's model series, EleutherAI's GPT-Neo/J series, Google & Google DeepMind's Gato, Gopher and PaLM-E, Meta AI's CICERO, Microsoft's Kosmos-1, OpenAI's GPT series, Stability AI's Stable Diffusion series, and Tesla Bot's underlying model.
#3	Type-II GPAI model (i.e cutting-edge)	AI models or combinations of AI models that are designed for generality of capabilities and expand the technology frontier relative to 2022 models. Like Type-I GPAI models, in the current technological paradigm, this often requires a significant amount of data and compute power. The expansion of the technology frontier relative to 2022 models has been achieved through unprecedented amounts of compute power, combinations of models and other architectural insights to maximise generality of capabilities.	OpenAI's GPT-4 & upcoming GPT-5, Anthropic's Claude, Google's LaMDA, Google DeepMind's Gemini, Meta AI's LLaMa, Microsoft's GPT4-Prometheus, InflectionAI's PI, and Stability AI's Stable Diffusion 2.0 & XL.

6.2 How to assess "generality of capabilities"?

Generality of capabilities can be understood as the share of a finite but exhaustive list of capabilities that a model can exhibit satisfactorily at any point during training.

- *"Capabilities"* refers to the concept of "broad cognitive abilities", used by Chollet [166] , as the intermediary echelon between general intelligence (i.e. extreme generalisation) and task-specific skills (with either no generalisation or generalisation only at a very local level) in his hierarchical model of cognitive abilities.[34]
- *"An exhaustive list of capabilities"* includes all theoretical capabilities of an intelligent entity. A close but imperfect example is the O*NET lists of abilities, skills and knowledge. 2017 O*NET contains 52 abilities, 35 skills and 33 fields of knowledge (cf. Annex 1 for the lists). Human Resource services firms have more thorough lists of such abilities at various levels of granularity.
- *"The share of"* such a list that a model can exhibit satisfactorily at any point during training is assessed by testing the model's performance at that capability, for a given operationalised threshold of "satisfactorily" (which includes consideration of time or compute constraints for achieving a certain performance). Eventually, this performance

[34] Note that capabilities is also related to concepts in in O*NET database [167]. The concept is limited to human capabilities which are currently in use in the US economy, while a GPAI model could have additional capabilities. These O*NET concepts include "Abilities" defined as "Enduring attributes of the individual that influence performance" Skills, containing "Basic Skills" (Developed capacities that facilitate learning or the more rapid acquisition of knowledge) "Cross-functional Skills" (Developed capacities that facilitate performance of activities that occur across jobs), and "Knowledge" (Organised sets of principles and facts applying in general domains).

will have to be predicted by the developers without having to train the model at all, as this prediction will inform development decisions and significant investment decisions.

It would be tempting to operationalize the last point through a series of test tasks that represent each capability assessed. However, using task-based evaluation to predict capabilities would fail to capture the ability of GPAI models to learn to solve tasks i.e. the dynamic nature of their task-solving abilities; a dynamism predetermined by their capabilities which are more stable over time [168]. Task-based evaluation would too easily lend itself to overfitting or, on the contrary, artificial stunting for compliance reasons. Alternative approaches to assessing capabilities include psychometric approaches [166] and information-theoretic evaluations [168]. While these have been implemented and tested, few if any modern GPAI models have been assessed in a systematic way, and they remain experimental.

Importantly, a large language model's capability is *not* defined as its ability to predict the next token with highest accuracy - even though this is what it has been programmed to do. Its capabilities emerge from its pursuit of that straightforward objective when provided with the necessary ingredients of a carefully-prepared dataset, large amount of compute and a training programme to communicate that objective. It is the emergence of these capabilities from a relatively simple (though data- and compute-intensive) process that makes general purpose AI so economically valuable. It is also what underlies their opacity, perceived complexity, and resulting riskiness.

In order to operationalize this definition, we propose the formation of a pool of European Benchmarking Institute as propounded by the European Parliament Committee on Industry, Research and Energy [169]. At its core, the Institute will contain a specialised taskforce, composed of benchmarking experts from across the EU and Member States metrology labs. Such an institute would be instrumental in devising robust and accurate metrics, along with measurement protocols as highlighted in Article 15 of the European Parliament's compromise text on the AI Act. The vision is to model this after the american counterpart National Institute of Standards & Technology (NIST), working closely with the EU AI Office, standard-setting organisations and benchmarking laboratories of Member States. To ensure the effectiveness and longevity of the taskforce, it is vital to institutionalise it by providing it with a clear mandate and a dedicated budget. Furthermore, this approach should be subjected to periodic reviews by expert panels to ensure its continued relevance and effectiveness. With adequate foresight and planning, there is a valuable opportunity here to institutionalise an approach that is both resilient and adaptive.

A core mandate of the Benchmarking Institute will be to ensure that AI models are assessed for their generality of capabilities, and to prevent that the imperfect proxies for assessing this dimension supersedes the meaning of the dimension itself. The development, maintenance and updating of state-of-the-art measurements procedures for generality of capabilities would therefore be its main work programme. If the GPAI models are found to be general based on these measurement procedures, they must align and comply with any pertinent requirements set forth by the AI Act.

In the box below, we outline a handful of metrics that could be used to evaluate the generality of a given GPAI model, facilitating its categorization across tiers.

Box 1: Proxies for "generality of capabilities"

Floating Point Operations (FLOP) used in training

The current "deep learning" paradigm in the field of AI has seen the emergence of compute- and data-intensive large GPAI models, which are mostly opaque and unpredictable but exhibit very broad generality of capabilities, in a way that correlates with the amount of compute used during training. This informs the initial criteria provided in this section to gauge the overall generality of the model. However, when equivalent generality of capabilities is achieved through other methods under other paradigms (e.g. expert models and emulations) the criteria in this section will need to be supplemented or made more focused on the relevant concept to reflect this evolution and ensure a technology-neutral regulatory regime.

The relation between training compute and generality of capability is known as the scaling laws [170]. When it comes to measuring compute, the general industry practice is to use floating point operations, 'FLOP'. A floating point operation is a computer operation equivalent to one basic mathematical operation (addition, multiplication, subtraction, or division).

For our purposes, the total amount of FLOP used to train the model is measured accounting for all FLOP used to develop the model at least back to 1st January 2021. This includes FLOP from the training runs themselves, the FLOP of models used to generate training data (if any) and the total amount of FLOP used to train the base or foundation models components (if any). This ensures there are no loopholes whereby a more powerful GPAI model is built by using a combination of (or "scaffolding" around) 2021 or 2022 models.

It should be noted that the amount of compute needed to achieve a given level of capability tends to decrease over time due to algorithmic progress: newer models require comparatively less compute than older models for achieving a given level of generality of capability. This phenomenon can be referred to as **"capability-per-FLOP" inflation.** As a result, to be converted into comparable "2022-FLOP", an adjustment is needed to ensure consistent measurement: past models' Total amount of FLOP is adapted downward, and future models' Total amount of FLOP is adapted upward. Noting this, the AI Board should convene at least yearly to assess whether to update "capability-per-FLOP" inflation figures, based on expected algorithmic efficiency gains in the previous year.

The total amount of FLOP disclosed nowadays is the outcome of measurements that are not standardised. Both developers and cloud compute providers would have better estimates of these figures available internally. Existing estimates also do not factor in the "capability-per-FLOP" inflation. Nevertheless, these existing estimates are a starting point to assess the incidence of the tiered approach on the economy.

However, based on existing estimates of training compute compiled and published by external researchers, as of July 12 2023, there have been around 30 Type-II GPAI models developed so far (Claude 2, GPT4, PaLM 2, ...) by around 8 providers [171]. There have been around 60 Type-I GPAI models developed so far (Stable Diffusion, DALL-E, Ernie 3.0, Gato, ...) by around 16 providers, including the ~8 Type-II providers. This is using thresholds defining Type I GPAI as 10^21 to 10^23 FLOP and Type II GPAI as 10^23 to 10^26 FLOP.

Table 4 provides a set of criteria, including total amount of FLOP used to train the model (2022-FLOP) as a means of determining whether a given model is Generative AI, Type-I GPAI or Type-II GPAI. That is, the total amount of FLOP used to train the model is being used as a proxy for the model's generalisation capabilities. Whilst we expect this to be an adequate stop-gap measure for the next 2 years, this approach to approximating generalisation will almost certainly need to be augmented in the relatively near future with more accurate benchmarks, such as Skill Acquisition Efficiency and Generality Analysis or modality-specific benchmarks, such as MMLU. Unfortunately such metrics have not yet been widely tested and adopted within the AI industry as a means of measuring GPAI capabilities. However we hope that with further research, and with support from the European authority on benchmarking, this situation will change. Other benchmarks as appropriate shall be developed or formalised by the European authorities on benchmarking.

Skill-acquisition efficiency

Chollet [166] argues that a major component of intelligence is not just the ability to perform a task (which might be the result of prior learning or training) but the efficiency with which a system or individual can acquire new skills. The concept of "skill-acquisition efficiency" is therefore defined by Chollet as how efficiently an agent can learn new tasks, given a certain amount of experience or training data. That is, the greater a model's skill-acquisition efficiency, the greater its ability to generalise across a range of tasks.

Chollet's Abstraction and Reasoning Corpus (ARC) then serves as a practical benchmark for evaluating the theoretical notion of intelligence defined as skill-acquisition efficiency.The goal of ARC is to test the true adaptability of a model, ensuring that success on its tasks requires genuine reasoning and cannot be achieved by brute force or memorization alone. Whilst Chollet does not provide a quantitative measure of the generalisation of the evaluation set given the test set, or the generalisation difficulty of each task, the performance of a given model could nonetheless be quantified as a percentage, representing the fraction of tasks it was able to solve correctly out of the total number of tasks. This would give insights into the model's degree of skill-acquisition efficiency and allow for comparisons between models with regard to this metric: a model that can solve a high percentage of ARC tasks demonstrates better skill-acquisition efficiency and, by implication, better general problem-solving abilities. Note that there have been numerous implementations [172] of the ARC and, contrary to many implementable benchmarks in machine learning that become obsolete after a few months, ARC has remained very discriminative [173] across levels of generality, with best performance steadily rising every year since 2020, but still falling very short of human performance.

Generality Analysis

A second auxiliary metric to use for measuring the general intelligence of a given model is Hernandez-Orallo et al.'s Generality Analysis [174]. It does this by measuring two aspects of the model: "Generality" and "Capability". Generality refers to how uniformly an agent's capability is spread across tasks of varying difficulty, whilst Capability denotes the maximum performance level an agent can achieve on the tasks it can handle. It's a measure of an agent's peak proficiency.

Generality Analysis is calculated using a response matrix, which represents the performance of agents (rows) on various tasks (columns). An intrinsic measure of difficulty is also factored in, which defines how hard each task is. Generality is measured based on how an agent's capability is distributed concerning this difficulty. The mean value for each row, representing an agent, across each task is provided as the overall Generality Analysis coefficient between 0 and 1.

One of the most notable differences between Generality Analysis and other analysis types like Principal Component Analysis, factor analysis, or Item Response Theory is that the latter are populational, meaning they are influenced by the entire population's performance. Generality Analysis, when using an intrinsic difficulty metric, remains independent of population effects, focusing on individual performance relative to task difficulty.

Data richness, Algorithmic Efficiency and Model Perplexity

Broadly speaking, the generalizability of a model's capabilities can be understood as a function of its training compute power, the richness of its training data, and algorithmic efficiency. The relation between training compute and generality of capability is known as the scaling laws. Training compute is the first proxy explained at length in this box, measured in FLOP.

The relation between richness of training data (quantity, average quality and diversity of data) and generality is a topic of active research. It is often assumed that greater data richness implies greater generality of capabilities, though the relation is unlikely to be linear. Measuring data richness can be done by comparing the size of datasets compressed using a standardised compression method. The advantage of this method is that there is no need to actually train the model in order to predict generality, as data richness is an ex ante ingredient.

The relation between algorithmic efficiency and generality of capabilities relies on calculating the "perplexity" of the model. By accurately predicting a model's perplexity, we could infer its generalizability. This, in turn, would allow us to apply the appropriate requirements and obligations to GPAI models with a clear distinction between their classifications (Type I and Type II). However, predicting the perplexity of an upcoming model based solely on its training programme is currently only a theoretical endeavour, with significant hurdles towards implementation.

A challenge in using model perplexity as a criterion for classifying GPAI models is that it is sensitive to the test set used. This means that if a model is trained specifically to perform well on a particular test set, its perplexity will look artificially good (and therefore its algorithmic efficiency artificially high). This opens up the possibility of "gaming" the metric if the test set is not kept secret or is inadequately representative. Another drawback is that perplexity (and therefore algorithmic efficiency) is measured ex-post (i.e. once the model is trained) and therefore cannot be fully predicted in advance. However, a possible countermeasure for both of these issues could be for a global regulatory body to maintain a confidential test set, using it to monitor perplexity during training runs of cutting-edge models. This could serve as an early warning mechanism for sudden improvements in algorithmic efficiency. If the algorithmic efficiency of a model at an early checkpoint of training surpasses that of previous models at a similar checkpoint, the authorities and developer could decide to interrupt training or to proceed more cautiously.

Modality-Specific Benchmarks

The current landscape of GPAI models suggests that for the next 1-2 years, the total compute used for training a model, measured in FLOPs, could serve as a reliable proxy for its generalisability. However, with the emergence of smaller models that demand fewer FLOPs while still achieving better levels of generalizability and capability, there will be a growing need to integrate a robust metric for a model's generalisation ability into the GPAI model tiering process. If current metrics like Skill-Acquisition Efficiency and Generality Analysis fall short or aren't suitable, it may be worth considering modality-specific benchmarks as temporary solutions until a more universally accepted measure for generalisability is established. For instance, the Automated Programming Progress Standard (APPS)evaluates a model's proficiency in solving coding problems using unrestricted natural language [175]. Similarly, the MATH [176] benchmark assesses the capability to tackle competition-level mathematics problems, while the MACHIAVELLI [177] benchmark is designed to gauge models' tendencies towards power-seeking behaviours, causing harm, and ethical infractions. Massive Multitask Language Understanding (MMLU) [178] has been used for language models' generality.

Table 4: Assessing tiers' thresholds through operational tests

Tier	Name	Operational test(s):				
#1	Generative AI application	1. Built upon a general purpose AI model 2. Refined for a specific purpose through prompt-based training, fine-tuning, reinforcement learning with human feedback, or other methods to narrow the model's purpose to a specific task with limited scope				
#1.5	*Foundation Models (EP)*	1. Can be applied to a wider range of tasks than tier #1 2. But still only <10^21 2022-FLOP to train the model				
If any of these criterion is met:		*Total amount of FLOP used to train the model (2022-FLOP)*	*Modality- specific benchmarks (e.g. MMLU average for language)*	*Skill Acquisition Efficiency (ARC Challenge)*	*Generality Analysis*	*EU-endorsed summary benchmark*
#2	Type-I GPAI model	>10^21 ≤10^23	>40.0 ≤68.0	>40/800 ≤60/800	…	…
#3	Type-II GPAI model	>10^23 ≤10^26	>68.0 ≤88.0	>60/800 ≤100/800	…	…
#3+	*Prohibited?*	>10^26	>88.0	>100/800	…	…

6.3 Implications of the tiers

The operational tests outlined in Table 4 and Box 1 inform the boundaries for falling into a tier or another. We can assess in practice the number of regulated entities and, if possible, products in each tier thanks to this work:

- Tier #1: Generative AI applications ➔ we identified over 400 providers worldwide falling in this category. Each of these often has multiple applications (e.g. voice generation also enables voice emulation), implying several thousands of applications. Note that the category is very dynamic: compiling the list of 400 providers over a period of 3 months in 2023Q2, we realised that two dozens of the first companies we had identified had already disappeared or been acquired.
- Tier #1.5: Foundation models ➔ we estimate 40-80 providers providing 85-170 foundation models. Indeed, as of September 17, 2023, one of the most thorough mappings identifies 175 foundation models from ~50 different providers [179]. However, the inclusion criteria are unclear. Moreover, these figures include most of the Type-I and Type-II GPAI models, which are already accounted for in tiers 2 and 3. We therefore reflect this uncertainty.

- Tier #2: Type-I GPAI models → based on compute estimates[35], as other test results are unavailable, we identified 62 models that fall between 10^21 and 10^23 of 2022-FLOP of compute (cf. box 1), incl. models developed by 14 different non-academic, non-governmental providers.[36]
- Tier #3: Type-II GPAI models → based on compute estimates[37], as other test results are unavailable, we identified 28 models that fall between 10^23 and 10^26 of 2022-FLOP of compute (cf. box 1), incl. models developed by 10 different non-academic, non-governmental providers.[38]
- Tier #4: prohibited GPAI models → based on compute estimates, no model has gone >10^26 of 2022-FLOP of compute, no provider falls into this tier.

Note that the figures for the number of regulated entities have significant drawbacks. First, some providers remain secretive about the models they develop or even whether a model is deployed, because of commercial strategy relying on publicising only when ahead of competition (e.g. Apple; InflectionAI and Tesla/Xai; or in the case of Microsoft deploying quietly GPT4-enabled Bing Chat in India and Indonesia) or because of sensitivity of the industry they provide to (e.g. Palantir).

Second, when the existence of a model is known, the relevant tests are mostly not carried out or, at least, the results are often not disclosed, even in whatever technical documentation exists (this is partly due to the challenge of Corporate Irresponsibility afflicting the industry).

Third, while some tests and assessment are carried out externally by experts[39] for some relevant operational tests or in competitions disclosing results on leaderboards[40], there is little harmonisation in the dimensions to measure, the protocol to measure them, and the way to report these measurements in a comparable and transparent way. The pooling of national benchmarking capabilities at the EU level should address some of these issues.

[35] Epoch "Parameter, Compute and Data Trends in Machine Learning", 2023, CC-BY, https://epochai.org/data/pcd

[36] Namely: Facebook/Meta, IBM, Beijing Academy of AI, Google/Brain/DeepMind, Microsoft, Alibaba Group, EleutherAI, Open AI, NVIDIA, Baidu, Cerebras, Runway/stability.ai, Huawei, NAVER.

[37] Epoch "Parameter, Compute and Data Trends in Machine Learning", 2023, CC-BY, https://epochai.org/data/pcd

[38] Namely: Amazon, Yandex, AI21 Labs, Inspur, Facebook/Meta, Anthropic, Google/Brain/DeepMind, Microsoft, OpenAI, Baidu

[39] Epoch.org, HELM being two examples.

[40] Hugging face leaderboard for LLMs

Full table: Estimated number of regulated entities per tier

Tier	Name	Operational test(s):					Estimated # of regulated entities
#1	Generative AI application	1. Built upon a general purpose AI model 2. Refined for a specific purpose through prompt-based training, fine-tuning, reinforcement learning with human feedback, or other methods to narrow the model's purpose to a specific task with limited scope					>400 providers, several 1000s of applications
#1.5	*Foundation Models (EP)*	1. Can be applied to a wider range of tasks than tier #1 2. But still only <10^21 2022-FLOP to train the model					40-80 providers, ~85-170 models
If any of these criterion is met:		*Total amount of FLOP used to train the model (2022-FLOP)*	*Modality- specific benchmarks (e.g. MMLU average for language)*	*Skill Acquisition Efficiency (ARC Challenge)*	*Generality Analysis*	*EU-endorsed summary benchmark*	*Estimates for tiers 2 & 3 based on compute estimates,* as other test results unavailable
#2	Type-I GPAI model	>10^21 ≤10^23	>40.0 ≤68.0	>40/800 ≤60/800	…	…	14 providers, 62 models
#3	Type-II GPAI model	>10^23 ≤10^26	>68.0 ≤88.0	>60/800 ≤100/800	…	…	10 providers, 28 models
#3+	*Prohibited?*	>10^26	>88.0	>100/800	…	…	0 provider, 0 model

Moreover, using compute estimates makes providers with particularly inefficient training (i.e. training that requires many more FLOPs than expected to achieve a given level of generality of capabilities) likely to be captured in a higher tier higher than would be implied by the generality of capabilities of its model. Finally, english-language bias persists in the GPAI industry, exacerbated by the secrecy some foreign providers may want to preserve during their fundraising or the difficulty to compare models in different languages or cultural contexts. It is possible that several providers of GPAI models abroad, notably in China, would be overlooked by third party experts, and it is possible that several would unduly be included due to lack of comparable testing configurations.

The estimates of the number of regulated entities and models are therefore only indicative. Only a regulatory intervention requiring suspected providers to share the information can help be more precise.

7. Enforcement & complicating factors

Given the potentially transformative impact of GPAI models and their potential risks, ensuring that the AI Act's provisions are effectively enforced despite the complicating factors is paramount. To this end, a range of enforcement measures and mechanisms will need to be put in place to achieve efficient and responsive implementation of the Act's regulations and to address nuances of the GPAI industry..

This chapter introduces and elaborates on various enforcement and implementation measures. Each measure is explained using a 'What', 'Why', and 'How' structure, in order to provide a clear, comprehensive explanation of each mechanism. This tripartite structure is designed to facilitate a comprehensive understanding of each requirement as one of the key concerns for stakeholders observing the AI Act has been the absence of clarity or granularity of the discussions, or the disconnect with business reality.

The **'What'** section provides an overview of the measure, defining its main features. It outlines the measure itself, describing the measure proposed. It provides a detailed description of the expectations for various actors, sets clear guidelines on what needs to be achieved or adhered to, and the outcome.

The **'Why'** section delves into the rationale behind the measure, explaining why it is needed and the benefits it brings in the context of GPAI regulation. This section delves into the underlying motivation for each measure. The aim is to make explicit the connection between the challenges posed by the deployment of GPAI models and the measures designed to mitigate those risks.

Lastly, the **'How'** section outlines how each measure is to be implemented, providing insights into the operational aspects and practical application of the enforcement mechanism or requirement. It serves as high-level operational guidance, providing where possible means and actionable steps or methods for the successful execution of each measure. This could include examples of best practices, required resources, potential challenges and ways to overcome them.

7.1 Measures for proper enforcement or implementation

Summary

Ensuring the enforcement of the AI Act's provisions is paramount to address the potential risks entailed by GPAI models. To this end, a set of mechanisms need to be put in place to achieve an efficient and responsive implementation of the Act's rules for generative AI and GPAI models, including:

- **Navigator programme**, which fosters direct bilateral relations between the European

Commission or AI Office's staff and each Type-II GPAI development team to promote trust and compliance.

- **Regulatory Sandboxes**, which allow for testing new products in a real-world environment for developers and regulators to better understand the technology.
- **AI Office**, which would act as a central point of contact for all stakeholders, concentrating expertise and enforcement capacities.
- **EU-level Pool of benchmarking authorities**, capable of bringing together Member State's metrology and benchmarking authorities to promote accountability and consistency across norms and standards.
- **Database of GPAI models** hosted by the AI Office, in which all GPAI providers register their models in Europe to facilitate the work of the Commission and the Member States and to foster transparency.
- **Updates of technical thresholds** in the legislation, such as the adoption of implementing acts by the commission where the technical aspects are specified to ensure GPAI is effectively governed.

From establishing agile governance solutions to fostering direct interactions between AI developers and regulatory authorities, these enforcement measures play a crucial role in translating the Act's provisions into tangible, effective regulation. Moreover, by allowing for flexibility, direct engagement, and collaboration, these measures aim to strike a balance between fostering innovation in AI and ensuring its safety, trustworthiness, and alignment with societal values. We explore enforcement in several other articles.[41]

7.1.1 Navigator Programme

Summary

The **Navigator Programme** is intended to foster **direct bilateral relations** between the **European Commission or AI Office**'s staff and each **Type-II GPAI development team**, facilitating monthly discussions under strict confidentiality agreements. Addressing the challenges of **rapid innovation** and **technical opacity** in the field of GPAI, this programme promotes **trust** and **compliance** by **building mutual understanding** between **developers** and **authorities**.

What: The Navigator Programme is a governance solution aimed specifically at developers of the most significant GPAI models (i.e. Type-II). The programme shall foster direct bilateral relations between the European Commission or AI Office's trained staff and each GPAI development team, facilitating monthly discussions under strict confidentiality agreements. These discussions will span topics from latest progress and experimentation in the GPAI team to measures in place for quality assurance and risk management.

[41] Giving Agency to the AI Act (2023) [4] The Future Society & Blueprint for an EU AI Office (forthcoming) [180] The Future Society.

Why: Corporate Irresponsibility (Challenge 4) and technical opacity (Challenge 6): Given the rapid innovation and unpredictable evolution in the field of GPAI, traditional, static governance mechanisms prove inadequate. Instead, a dynamic, tailored approach is required to balance innovation and trustworthiness. The Navigator Programme's flexible controls address this need. It directly engages with GPAI developers to ensure proportionate compliance-by-design and builds an evidence base for designing smarter governance mechanisms in the future. Furthermore, the regular dialogue shall form a mutual understanding and a trust-based partnership between developers and authorities, allowing both parties to navigate each other through legal implications and technological advances, respectively..

How: The implementation has been covered at length in earlier work [6]. In brief, each Type-II GPAI development team is assigned a staff member from the European Commission, who initiates monthly dialogues regarding the model's progress, risk management, and regulatory compliance, among other topics. This staff member serves as a navigator for the developers, guiding them through the legal implications of their work, while the developers keep the civil servant up to date with the latest technological advancements. All interactions are subject to strict confidentiality to protect commercially sensitive information.

7.1.2 Regulatory Sandboxes

Summary

Regulatory sandboxes allow innovators to **test** their **new** AI-based **products** and services in the **market** under i**ncreased regulatory oversight**. It provides regulators with a deeper understanding of new technologies and business models, and builds customer confidence in engaging with disruptive technology providers.

What: Regulatory sandboxes (RSs) are programmes administered by the AI Office that allow innovators to test their new AI-based products and services in the market under increased regulatory oversight. They have been covered thoroughly in our earlier work.[3]

Why: Corporate Irresponsibility (Challenge 4) and technical opacity (Challenge 6): RSs allow for monitored regulatory flexibility for entrepreneurs to test their innovations in the market, provide regulators with a deeper understanding of new technologies and business models, and build customer confidence in engaging with disruptive technology providers. RSs enable entrepreneurs to work transparently and in collaboration with regulators for compliance-by-design, rather than attempting to fly under the radar of authorities. The insights gained by regulators can be used to assess and refine the regulatory environment. Additionally, customers and society at large can benefit from the adoption of innovative products, knowing that their fundamental rights, health, and safety are protected.

How: The creation of RSs necessitates the establishment of two types of RSs: Physical RSs for AI systems embedded in physical products or services, and Cyber RSs for standalone AI systems operating in cyberspace. The differentiation ensures appropriate and tailored testing environments for different types of AI applications.

The rollout of these RSs will occur in three phases: Phase 1 aims to generate demand for sandboxes from entrepreneurs and regulators, facilitated through the creation of an "innovation hub" and the promotion of compliance-by-design practices. In Phase 2, efforts will focus on boosting the capacity of regulators and increasing customer confidence through transparent information sharing about AI systems and investment in institutional infrastructure. Phase 3 will involve an evaluation and potential redesign of the ecosystem based on impact assessments and previous experiences, with plans for streamlining and scaling the system developed according to the needs and constraints of both Physical and Cyber RSs.

7.1.3 AI Office

Summary

The **AI Office** would serve as a **central point of contact for all stakeholders**, bringing together the necessary expertise and resources to enforce the new regulation across Member States and abroad, to facilitate the **concentration** of **expertise** and **resources**.

What: The AI Office would serve as a central point of contact for all stakeholders, bringing together the necessary expertise and resources to enforce the new regulation across all EU Member States and foreign regulated entities. As discussed extensively in other research [180] the Office would interact with regulated entities and providers of models of concerns i.e. these models that fall in the upper tiers or that affect a cross-section of member states. Though consensus hasn't been reached, various legislators would like the EU-level authority to be assigned several missions related to GPAI models (Type-I and Type-II) [4], , including but not limited to several necessary for the proper enforcement of the tiered-approach suggested here:

- Monitoring of the technological and commercial landscape; collect evidence and monitor efforts for incident management and major accident prevention.
- Foresight for technological trends, with internal reports about current and planned training activities by GPAI providers; monitoring of large training runs as defined in the tiered-approach.
- Inspecting and carrying out unannounced visits to development facilities, carrying out interviews of developers.
- Cooperation with external experts, establishment of an advisory group for general-purpose AI models; collaboration with regulatory authorities abroad.
- Oversight of the Navigator Programme and of the Regulatory Sandboxes.

- Bringing together national metrology and benchmarking authorities and provide guidance to address the technical aspects of the AI Act
- Providing guidance on risk assessment methodology, on delegated acts, on common specification, on compliance and acceptable means of compliance.
- Render binding the commitments made by existing GPAI providers.

Why: The AI Office will facilitate the concentration of expertise and resources, thereby ensuring effective and consistent enforcement of the new AI regulation across all member states and foreign entities. This would help govern AI industry players of concern (such as GPAI providers) cost-effectively and provide a centralised platform for stakeholder interaction.

How: As discussed in [4], the AI Office could function through two primary modes: an AI Board or an AI Agency. An AI Board leans towards a more decentralised approach, enabling Member States to retain more control but potentially leading to inconsistencies in the application of the AI Act. In contrast, an AI Agency adopts a centralised stance, ensuring consistent application of AI regulation but potentially facing resistance due to perceptions of national sovereignty infringement. An AI Agency set-up is preferred, in particular to deal with regulated entities of concern with compliance function.

7.1.4 EU Benchmarking Authorities

Summary

A common **European authority on benchmarking** would **bring together** Member States' national **metrology** and **benchmarking authorities**. Since benchmarking will be integral to the effective regulation of GPAI under the EU AI Act, this would help establish **consistent benchmarks** that would promote transparency and fairness in AI, while creating a means to hold systems accountable.

What: The AI Office and Commission shall collaboratively establish a common European authority on benchmarking that brings together Member States' national metrology and benchmarking authorities. Its primary task will be to address technical questions pertaining to the measurement of concepts relevant to the definition, typology and categorisation of and discrimination amongst general purpose AI models.

The EU benchmarking authority's mandate will encompass a wide array of benchmarks. Examples include, but are not limited to, the total amount of Floating Point Operations (FLOP) utilised during the model's training phase, metrics that evaluate the prediction accuracy of models, the time required to train the models, data efficiency, and model robustness. Other pertinent factors, like the environmental impact of AI model training and the degree of model alignment with ethical standards and societal values, can contribute towards a comprehensive benchmarking framework

Why: Benchmarking will be integral to the effective regulation of general purpose AI under the EU AI Act given the technical opacity of these technologies (Challenge 6). It shall serve as a robust mechanism for evaluating AI performance against standardised metrics, thereby facilitating comparative analyses of different AI models and systems. Further, the establishment of consistent benchmarks promotes transparency and fairness in GPAI and provides a means to hold models developers accountable for the evidence-based outcome of their work.

How: The establishment of the European benchmarking authority would begin by convening experts to address the question of what needs to be measured for the AI Act. Following these discussions, a taskforce of skilled experts would be formed, bringing together representatives from Member States' national metrology and benchmarking authorities. This taskforce is then institutionalised with a clear mandate and budget, enabling it to define and implement a robust benchmarking framework for effectively measuring and comparing AI models and systems.

7.1.5 GPAI Models Database

Summary

All **GPAI** model **providers** should be required to **register** each of their **models** in an **EU database managed by the AI Office** to facilitate the work of the Commission and Member States, and to ensure **transparency** towards the **public**.

What: Producers of GPAI models should be required to register each of their models in an EU database, along with all other high-risk (non-GPAI) models, to be established and managed by the AI Office. Any substantial modification of GPAI models shall also be registered in the database.

Why: A database of GPAI models is required in order to facilitate the work of the Commission and the Member States within the field of GPAI as well as to increase the transparency towards the public [181]. Beyond the logistical use of it as a practical way to monitor compliance, it also helps address a few challenges:

- A centralised database managed by the AI Office would make it easier for the Commission and Member States to keep track of GPAI models incidents and accidents, and conduct appropriate regulatory oversight (challenge 7). This scrutiny would help to mitigate risks as regulatory bodies can quickly respond to emerging issues with specific models.
- The database reduces technical opacity (challenge 6) thanks to the comparable information it makes available to the public. This information, however limited, is currently not easily accessible. Sometimes, even the existence of such a model, or its embedding

into a service, is not known. Having a publicly accessible database increases transparency, offering insight into the types of AI systems being used and making information about these models more widely available.

- A database with records of all substantial modifications to GPAI models helps hold providers and deployers accountable for their models' performances as changes that lead to negative impacts can be traced back to their sources; mitigating corporate irresponsibility (challenge 4).

How: There are multiple regulation-related databases managed at the EU level, so we don't envision much difficulty. The AI Act also includes a database of high-risk systems; a similar procedure could be used for GPAI. This database should be freely and publicly accessible, easily understandable and machine readable. The database should also be user-friendly and easily navigable, with search functionalities at minimum allowing the general public to search the database for specific high-risk systems, locations, categories of risk and keywords. The AI Office should be the controller of the database. In order to ensure the full functionality of the database, when deployed, the procedure for setting the database should include the elaboration of functional specifications by the AI Office and an independent audit report.

7.1.6 Technical Thresholds Updating

Summary

To ensure effective governance of GPAI, the **commission** shall be **empowered** to adopt **implementing acts**, in which **technical elements** may be adapted and **specified**.

What: In order to ensure uniform conditions for the effective implementation of regulation on general purpose AI, the Commission shall be empowered to adopt implementing acts. These acts will specify and adapt the technical elements of the given approaches, taking into account market and technological developments. Elements that will be subject to specification or modification through these implementing acts shall include metrics or benchmarks for the categorisation of Type-I/Type-II GPAI and any specific requirements or exemptions for open source models.

Why: These updates are important for future-proofing the tiered-approach, as explained extensively in Box 1 (Section 4). It also addresses Generalisation & Capability risks (Challenge 2) and Corporate Irresponsibility (Challenge 4). The adoption of implementing acts is crucial to ensure that the AI regulation remains current, robust, and effective in the face of rapid technological advances and evolving market conditions. By specifying and adapting technical elements and requirements, the Commission can respond promptly to innovations and changes in the AI sector. This flexibility in adjusting regulatory frameworks is important in a field like AI where advancements can quickly outpace existing rules and standards.

Moreover, it allows for a more granular oversight and regulation of different AI categories, including the Type-I and Type-II GPAI and open source models. This constant refinement and adaptation helps protect consumers, promote fair competition, and foster innovation in the AI space across the EU. It also ensures that AI regulation is grounded in technical realities, which is key to its effectiveness and legitimacy.

How: Implementing acts undertaken by the Commission shall be done in accordance with Article 5 of Regulation (EU) No 182/2011. These implementing acts will draw upon recommendations from the AI Office and the benchmarking authorities, and divergences from these recommendations will be duly justified by the Commission and explained technically. The establishment of technical elements, general provisions, and the categorisation criteria for Type-I and Type-II GPAI shall occur within one month following the regulation's entry into force. Furthermore, the technical elements will undergo revision at least annually, taking into account the evolution of market and technology trends. This approach ensures that the criteria continue to accurately reflect the corresponding level of general capabilities.

7.2 Measures for combination of models/interacting models

Summary

The main requirements to effectively govern the combination or interaction of models include:

- **Managing Unintentional Interactions**, through proper assessment, communication and mitigation if during the testing or at deployment a GPAI model unexpectedly interacts with one or more GPAI models.
- **Managing Reasonably-foreseen Interactions**, through satisfactory ex ante assessment and communication to the AI Office, in order to build and maintain an industry-wide map of models' interactions.

7.2.1 Managing Unintentional Interactions

What: Providers who, during the testing or deployment of a GPAI model (Model A), inadvertently instigate interactions with one or more other GPAI, generative AI, or AI systems (Model B), are required to take specific steps. This occurs for example when two GPAI-enabled trading AI systems start being significantly affected by the buy/sell decisions of one another[42], or when an automated scheduling system is being accidentally "hacked" by automated spamming assistants.

[42] Note that GPAI-enablement is not necessary for these interactions to be catastrophic, such as the 2010 flash crash which erased $1 trillion in market value [182]. GPAI-enablement however greatly complexifies the interactions as it enables the use not only of price signals but also GPAI-written analyses and reports on prices and tendencies for various products, which are re-used as signals by other systems.

We recommend the following steps :

1. Carry out an initial assessment of the interaction, encompassing technical effects such as alterations in input, output, or objectives. Consider its projected influence on end users, stakeholders, providers, and other operators and enumerate potential long-term feedback loops. If these loops and their potential impact pose a risk to stakeholders, the provider should halt testing or retract the model deployment.
2. Contact the providers whose model or system (Model B) was interacted with to inform them of the interaction. Share the preliminary assessment and inquire about Model B's own interactions (termed '2nd-order interactions') with other models or systems. If a risk emerges from these 2nd-order interactions involving Model B, the provider of Model A must suspend testing or retract the model deployment.
3. Relay the details of the interaction, the preliminary assessment, and the correspondence with Model B providers to the AI Office.
4. Begin the process of developing a detailed technical map of the interaction, inclusive of the 2nd-order interactions. This map should be submitted to the AI Office within 10 days of the interaction occurring.

Why: As GPAI models like AutoGPT, developed from OpenAI's GPT-3.5, increasingly exhibit autonomous interaction with other systems over the internet, the need for special protective measures is escalating. Given the extensive penetration of these AI models across diverse sectors (Infrastructural aspect, Challenge 1), even minor interactions could trigger substantial consequences that ripple across the economy.

The stipulated actions help in preemptively identifying and mitigating potential risks, thus reducing the possibility of accidents (Challenge 7). These measures are especially important given the potential scope of these incidents and their ramifications due to the autonomous nature of the models. Ensuring robust reporting and management of inadvertent interactions can act as an additional layer of protection against catastrophic outcomes. If an interaction were to escalate uncontrollably, the swift detection and intervention could prevent a potentially disastrous event.

How: The EU AI Office serves as a clearing house for these interactions, maintaining the map and history of incidents. Providers should have real-time monitoring systems that can immediately identify anomalies [183] and detect when their GPAI model (Model A) is inadvertently interacting with another model (Model B). In such a scenario, the provider should either rollback aspects of their model or shut it down completely, based on the interaction's severity. Finally, the provider of the model must promptly coordinate with all relevant stakeholders, indluding the AI office and other regulatory bodies, downstream deployers, and users of the system [184].

7.2.2 Managing Reasonably-foreseen Interactions

What: Providers who intentionally cause a GPAI model to interact with one or more other GPAI, generative AI, or AI systems (a 'combination of models'), either by design or experimentally, are required to:

1. Notify the AI Office and the original provider(s) of the models being combined, clearly stating whether this combination qualifies as a Type-I or Type-II GPAI, along with the technical and explicit objectives of the individual models and the model combination.
2. Conduct a thorough assessment of the interaction induced by the combination, encompassing technical effects like changes in output or input, the projected potential impact on end users, stakeholders, providers, and other operators, as well as a list of potential long-term feedback loops. If these loops or the potential impact pose a risk to stakeholders, the model combination must be aborted.
3. Prepare a detailed technical map of the interaction, including the 2nd-order interactions, and submit this to the AI Office.
4. Meet the requirements outlined in the tiered approach (Section 8) scaled to whether the model combination equates to a Type-I or Type-II GPAI. The provider may utilise the conformity assessment of the original providers in fulfilling this requirement.
5. Share the thorough assessment and conformity assessment with the AI Office, without prejudice to other requirements stated in this Regulation.

Furthermore, the provider must continually monitor the combined model after deployment and roll it back if unexpected outputs are observed. The AI Office should be notified of any such rollback or incidents.

Why: The AI Act currently overlooks the development of hybrid models, such as ChaosGPT, BabyAGI, AutoGPT and other “LangChain”-based combinations. This provision aims to address this regulatory gap. Indeed, intentional combination of models are intended to expand the generality and autonomy of the GPAI models (Generalisation & Capability Risks, Challenge 2) It also helps address Corporate Irresponsibility (Challenge 4) and the much greater technical opacity (Challenge 6) induced by these combinations, by forcing a mapping of the interactions and aggregating these mappings at the macro-level.

Moreover, by forcing providers to notify relevant authorities and other involved providers, conduct comprehensive risk assessments, and continuously monitor the combined models, this requirement ensures that they are held accountable for their role in managing these AI systems.

How: The EU AI Office serves as a clearing house for these interactions, maintaining the map and history of incidents. Providers should assess ex ante and list their GPAI model (Model A)’s interactions with other models. In such a scenario, the risk management system, quality management system, accidents prevention policy, and other requirements shall explicitly address these interactions and the tests shall be adapted to meaningfully capture the dimensions of the combined models.

7.3 Measures for open source

Summary

The main requirements to effectively govern open source models include:

- **Open source observatory**, which shall be joined by all open source providers as well as open source hosting platforms, foundations, experts and representatives from civil society, to assess and refine rules for open source GPAI models.
- **Adaptation for open source providers**, of some of the acceptable means for compliance, taking place in conjunction with the Open Source observatory.

7.3.1 Open Source Observatory

What: In order to address the specific regulatory challenges related to open source GPAI, providers of general purpose AI models and generative AI, alongside the AI Board, and if applicable, the open source hosting platforms, Open Source foundations, experts and civil society, shall cooperate to establish a joint open source observatory to assess cost-effectiveness and implementation of the provisions in the open source ecosystem. This observatory's mission is to

- evaluate the cost-effectiveness and successful implementation of the AI Act's provisions within the open source ecosystem,
- further innovation and democratisation of RegTech[43]
- monitor misuse and malicious use trends in AI.

For example, one of the first challenges for this mandate will be to establish collaborative and cost-effective verification systems, such as trusted buyer systems, to mitigate misuse and malicious use of Open Source models. This might involve the active cooperation of the hosting platforms, Open Source actors, and other stakeholders as has been regularly the case for example with software verification, trust and security initiatives (e.g. sigstore [185]). Providers of GPAI models and generative AI distributed as Open Source software may be invited to participate in this joint Open Source observatory.

Why: Addressing the unique regulatory challenges associated with open source GPAI necessitates a collective effort. The Open Source Observatory provides a platform for effective collaboration, knowledge sharing, and strategic decision-making to ensure the secure and responsible growth of the open source GPAI ecosystem. As policymakers debating the AI Act have realised over the past 2 years, governing the Open Source ecosystem is a major challenge. Ensuring that there is no over- nor under-regulation is crucial, given the role Open Source plays as a supply of models for the EU downstream innovation (Challenge 3, Concentration of Power) but also the significant risk for misuse and malicious use (Challenge 5).

[43] Regulatory Technology is the application of digital technologies to facilitate compliance without reducing the quality of enforcement.

How: The EU AI Office would first establish a subgroup in charge of Open Source that could lithely fulfil the observation function, informing the creation and calibration of this Observatory. If Open Source providers systematically reach out to that subgroup with questions about the AI Act, the subgroup will quickly be overwhelmed, which would be a signal that a separate team might be needed. The insights gained during the implementation period will therefore not only be key for the adaptation discussed below, but also for the formation of the Open Source governance structure.

7.3.2 Future-proofing Adaptation

What: Certain obligations and requirements for Generative AI and Type-I GPAI models, such as the requirement to label AI-generated content, should be adapted to Open Source value chain operators to ensure the spirit of the EU AI Act is carried to that ecosystem while avoiding issues of "hold-up" or "regulatory moats" by actors with greater bargaining power (e.g. platforms or companies other than SMEs).

Why: Due to the asymmetry of power in the Open Source ecosystem itself (Challenge 3), and considering the specific nature of open source operations, some regulatory obligations may disproportionately affect some providers using Open Source distribution relative to the social benefits the requirements actually generate at a specific actor or model's scale. Recognising this, this regulation would provide for certain adaptation to maintain and update over time the requirements for Open Source. In the absence of such adaptations, stringent regulations could inadvertently favour large, resource-rich providers capable of navigating complex regulatory landscapes, thereby exacerbating monopolistic behaviours; and reduce the ability of the EU AI open source ecosystem to develop viable trustworthy alternatives.

How: The Commission, following consultation with the Open source observatory and the EU AI Office may adopt implementing acts to specify and update these adaptations, in light of market and technological developments and findings from the joint open source observatory.

7.4 Measures for Value Chain governance

Summary

Value chain governance is necessary to mitigate five of the seven challenges identified. It is achieved through:

- the **De Facto Control contractual framework**, which is a set of rules to facilitate the evidence-based and proportionate transfer of responsibility for compliance along the

GPAI value chain, via regulated contracts.

- **Tier-wise conformity assessment to ensure downstream value chain actors can integrate GPAI model in their products without undue legal risk**, thanks to intermediary or component conformity assessment carried out by the upstream developers of GPAI. For **Generative AI applications, internal conformity assessmen**t is sufficient. For **Type-I GPAI models, external conformity assessment** is necessary. For **Type-II GPAI models,** given the near monopoly of expertise, **a "joint" conformity cross-assessment is required,** inducing liability for both the provider and the auditor.

Value chain governance is one of the complicating factors, recognised by policymakers and industry early on in the debate [94] [186]. In order to address this, we recommend what we term as the "*de facto* control approach" and conformity assessments.

7.4.1 De Facto Control Contractual Framework

What: The De Facto control approach is a way to clarify the distribution of compliance responsibilities across the value chain actors of AI systems, applying to all types of value chain. It is particularly relevant for GPAI and generative AI systems, where there is reliance on GPAI as a service. In brief, it relies on giving responsibility over a compliance-relevant aspect of an AI system to whoever along the value chain has de facto control over it.

An "aspect" of an AI system is an element, process, production decision or design decision constituting the AI system, akin to an AI system's component. For example, "data collection", "training dataset", "compute", "pre-trained model", "training process", "testing & verification of interpretability", among many others are aspects of an AI system. Each aspect can be further divided as necessary for contractual clarity, and so long as one actor in the value chain recognises the relevance of an aspect for compliance, it should be discussed.

"De facto control over an aspect of an AI system" can be assessed using the following test: which actor or set of actors along the AI system's value chain has enough information, capabilities/skills, or technical access to alter that aspect deliberately when desired and bring it into or out of compliance with the AI Act. "Information", for example, would include tailored detailed explanations and standard technical documentation. "Capabilities/skills" includes existing skills or skills gained through ad hoc training and webinars provided by one actor to another. "Technical access" might for example include giving access to deployed AI systems to a trained maintenance technician (similar to car mechanics' maintenance) or providing the model source code and dataset.

Note that it can be an actor or a set of actors given de facto control over an aspect, implying joint or disjoint control, and therefore responsibilities. For example, if it is too complicated to ensure successful separation of de facto control over an aspect, the developer and deployer may opt to continuously exchange information and to give technical access to staff with the requisite capabilities to monitor the compliance-relevant aspect.

Why: Due to the infrastructural aspect, asymmetries of power and corporate irresponsibility (Challenge 1, 3 and 4), value chain actors are often unable to reach economically efficient contract terms. Moreover, due to the technical opacity of these models, inducing asymmetry of information (Challenge 6) and risks of misuse (Challenge 5), there are significant benefits to making explicit who-knows-what to avoid gaps in perceived own responsibility, which has already resulted in major accidents in the past (for example, in the case of the English breast screening program scheduling algorithm[44]

The other approaches proposed in the debate so far do not work: a case-by-case contractual agreement between the upstream GPAI provider and downstream deployer without the de facto control contractual framework doesn't provide the downstream deployer with sufficient bargaining power (from Challenge 3 i.e. concentration of power in the upstream). The opposite approach of indiscriminate blanket-transfer of liability for compliance with upstream responsibilities and obligations onto downstream deployers (a recurrent talking point of GPAI model developers) is equivalent to carpet-bombing the EU downstream AI industry: indeed, it would solidify upstream concentration of bargaining power upstream. The idea that market players would then efficiently rely on secondary adjudication to then distribute the costs of fines and litigation, which only works if the primary adjudication targets the most powerful player in a supply chain, who can expend the effort to sue less powerful players in private tribunals afterwards.

The de facto control contractual framework relies on ex ante exchange of information as opposed to ex post exchange of lawsuits. It reconnects in a proportionate way "responsibility" with the actor's economic efficient ability to "control" i.e. reconciling the "de jure" and "de facto" realities. It solves also solve several problems specific to the GPAI value chain:

- It is technology-neutral: even if the technological paradigm evolves and new types of GPAI occurs, the contractual framework's principle and precedents remain applicable. There will be new "aspects" of GPAI models, but these can fit the framework.
- It is actor-neutral: it addresses the supply and value chain evolutions, and their variety of actors (compute providers, developers, red-teamers, data providers, application developers, deployers, operators) in the GPAI value chain indiscriminately. It does not

[44] The English breast screening program's scheduling algorithm, which automatically invites elderly women for preventive breast cancer prevention screening. IT systems involved in the program (health database management software, algorithm provider, etc.) obscured a miscalibration of the automated scheduling for over 4 years. The authorities in charge of this program reported that the miscalibration affected up to half a million people. A 6-month investigation uncovered that poor communication between value chain actors for an algorithmic system can affect the health of hundreds of thousands people [187].

systemically favour an actor or another, it does not legally enshrine inequalities. If new types of players (e.g. brain data holders or MLops trustworthiness firms) appear in the value chain, their contractual relationship can still respect the principles of the de facto control approach. This is particularly important because this industry is still exploring the optimal supply chain setup, contrary to e.g. automotive supply chains which, despite high complexity, have been almost institutionalised after close to a century of evolution.

- It is regulation-neutral: we can imagine the rules and regulations surrounding AI will evolve in the coming years and decades. The evolution of requirements actors want to fulfil is possible under the De Facto approach.
- It is exhaustive: the regulator does not need to intervene for every novel pair-wise relationship in the value chain.
- It is independent of business models and distribution channels: it helps resolve the Open Source complicating factors by making it clear to potential users of OS GPAI models that they commit to control a very broad range of aspects. It improves on the sometimes abusive disclaimer of “no guarantee of fitness of purpose” used in OS licences by forcing to detail explicitly what aspects are actually guaranteed (cf. a few paragraphs below).
- It also resolves the bargaining power issue efficiently: the De Facto control contractual framework is a “templated“ approach with mandatory but simple harmonized way of disclosing who controls what aspect:
    - It is simple in the sense that lawyers and legal experience don’t give an undue advantage to any party, so that smaller players -typically with little to no legal counsel- are on the same footing as tech giants who have several hundreds or even thousands of lawyers available. Only technical expertise and humility about own knowledge & know-how pays off in contractual negotiations.
    - The disclosure process is harmonized, and therefore no longer relies on the strongest players’ “take-it-or-leave-it” own way of disclosing responsibilities.
    - As it ties responsibility to actual control, undue brand effects vanish under the De Facto control. Undue brand effects occur when one procurement manager feels incentivized to use the well-known brand in order to shelter oneself with argument that deviating from that norm would expose them to personal blame from management in case of supplier’s underperformance, regardless of whether the lesser-known supplier has better or worse performance. “Everyone does it” and “If they fail, every other supplier would have failed” are common rationalisations, based on brand regardless of supplier’s track record.[45]. For an industry widely acknowledged to be plagued by “hype” and particular technical opacity (Challenge 6), ensuring that branding effects are not used to mislead downstream providers is important.

[45] A typical advice in industrial procurement “If you don’t know what supplier to select for cybersecurity, use Cisco; regardless of track record, no board member will ever blame you personally for using Cisco even when it fails. If you use anything else, you become individually responsible for all failures of that supplier.” (CTO office advisor on AI at a Fortune 20 company , in conversation with the author)

Moreover, there are more incidental but important benefits of the De facto control approach. It helps compile the technical documentation of the final AI system: each aspect is documented by the person who is (de facto) in control of that aspect. This also helps solve significant problems of the Open Source governance.

- One of the main issues with OS is that compliance efforts are not rewarded because other low-quality but hyped OS components are flooding the market (a typical example is data laundering, where researchers create datasets evading most regulations about child pornography and copyrights under the premise that research is not regulated [188]; flooded OS ecosystem with low quality/low costs but hyped datasets.)
- At the same time, imposing compliance with requirements would disproportionately impact genuinely community-based, non-commercial OS, as these have less resources to dedicate to compliance than corporate OS projects. The distinction is blurry [93].
- To level the playing field, it is important to at least force the disclosure of what aspects have been controlled, what has been done/assessed, so that OS players making an effort are rewarded by being comparatively better than less motivated players.
- The De Facto control approach enables OS developers to be explicit about their quality control, aspect per aspect and be transparent about it - resolving a major asymmetry of information that persists in the OS ecosystem.

It also prevents abusive OS-based business models that rely on the “hype” benefits their products generate (e.g. equity of funding or stock price increases in the case of Meta’s Llama and stability.ai’s stable diffusion). These business models rely on unclear loopholes in the current regulation, for example leveraging the fact that, currently, releasing an OS model and then only “consulting” for helping deploying it at a client’s site prevent any product liability. The De facto approach would prevent this abuse as a consultant constitutes a form of technical access, skills and information about many aspects of the model, and the de facto control contractual framework would imply that the consultant has responsibility in case of non-compliance. On the contrary, a trainer educating the deployer about an OS model to the point that the downstream deployer estimates being in control would be sufficient for transferring part of the responsibility.

The De Facto approach also sets the foundation to address the technical opacity (Challenge 6). If a developer releases a GPAI model, currently, it means a black box appears in the environment, hard to reverse engineer, but relatively easy to use. The downstream might be fortunate enough to have a model card describing some superficial features of the GPAI model, but most of the documentation focuses on the “how to use” as opposed to “what compliance-relevant design decisions have been embedded in that model”. With the De Facto approach, that OS developer must disclose clearly the de facto control it has on what aspects e.g. the tests he did and *did not* do, the design features relevant to performance, .. It can no longer hide behind a disclaimer of “no warranty of compliance/fitness for purpose” buried in ToS while hyping the model in the media and then claim they are not not responsible for damage done.

This disclosure incentivizes one of the following strategy for the OS GPAI Model developer

- Making OS GPAI compliant for high-risk, to shield from any liability;
- Ensuring no customer re-uses the GPAI model for high-risk, which is difficult in the OS world;
- Being very clear that the released GPAI is non-compliant, so that judges determine beyond reasonable doubt that the deployer is the problem.

And therefore, the De Facto approach ties “hyping” with “responsibility”: no hype, no responsibility.

Finally, the De Facto approach enables to disclose the absence of control (from any actors) on some aspects of cuttinged-edge or Type-II GPAI models (Challenges 2, 4 and 6). We have already explained that developers of these models do not have de facto control on many aspects of their own models, despite their near monopoly on information, skills and technical access. Notably, this is because these models are having “emergent capabilities”: the complexity of the model-generation is such that it is not humanly possible to predict these capabilities, no matter their skills and access. New machine tools are needed in order to establish De Facto control over these models.

This lens on the De Facto control of GPAI models is an important reason why a separate regime is needed for GPAI models i.e. to reduce the risk emanating from lack of de facto control and to mandate investments in gaining de facto control; we explore this separate regime in Sections 8.2 and 8.3.

How: There are 3 simple rules and one corollary for implementing the De Facto control approach as a contractual framework:

1. You must disclose clearly your de facto control (or lack thereof) over all relevant aspects identified: this could be part of the technical documentation, where you describe the item you provide. List of relevant aspects identified can be built over time based on incident investigations, court cases, etc. De Facto control should be expressed as a combination of information, skills/know-how, and technical access to the relevant aspect.

For example, a pseudo-contractual annex could read this way:

> *We provide standalone dataset Z (either OS or API or built tailor-made or…);*
>
> - *We have de facto control over:*
>   - *Data collection method (Art. 99 (b))*
>   - *Initial labelling procedure (Art. 99 (c))*
>   - *Testing & verification for accuracy of labels (Art 95 (a))*
>   - *…*
> - *However, we don’t have de facto control over*

- *Testing for discrimination*
    - *No information: we lack information on what the representativeness and inclusiveness could look like*
    - *No skills: we don't have the technical skills to test for representativeness*
- *[other aspects relevant to datasets]*

Similar disclosure rules would be mandated for the pre-trained model, the GPAI. In particular, it implies making very granular descriptions of "lack of control".

2. The value chain actors are allowed to exchange information, skills and technical access so that they have collective De Facto control. This enables e.g. the provider to determine whether sending a technician is more profitable than teaching the deployer; whether customers need more than model cards or whether they are sufficient. So the De Facto control contractual framework would over time include template contracts for transfer of control, to limit abuse.

3. You are presumed responsible for non-compliance if non-compliance arises because of one of the aspects you control. Contrary to automotive, don't assume secondary adjudication or private arbitration work because in GPAI, there is an upstream oligopoly (Challenge 3). So, agreements must be clear enough to be adjudicated over in primary adjudication when there is non-compliance.

The corollary is that De Facto control becomes an asset to be traded and invested in: investing in developing De Facto control over a GPAI model is valuable because it credibly shows up in contract negotiations with the downstream customers, as opposed to as ex post adjudication (as is currently the case). Building De Facto control over an existing dataset is also valuable (e.g. testing its representativeness and such) as it enables to make guarantees that are otherwise hidden in the "hype-fog". A customer may request this De Facto control to be made.

7.4.2 Tier-wise Conformity Assessment

What: In the performance of their own conformity assessment for an AI system using a GPAI model or generative AI, the downstream provider may rely on the conformity assessment of the GPAI model or generative AI as evidence of conformity for the related component of its AI system. This presumption of component conformity shall not apply if the downstream provider modifies the component GPAI model or generative AI during its integration in a way that affects the compliance of the component with the requirements if it shifts the De Facto control (cf. previous section) away from the upstream provider. This also applies in cases where a GPAI model or generative AI system is based on or built with another GPAI model or generative AI system.

The conformity assessment from the upstream component shall correspond to the tier of the component, irrespective of how it is placed on the market or put into service, including if placed as open source software:

- Tier 1: Generative AI providers shall carry out an internal conformity assessment as described elsewhere in the AI Act.
- Tier 2: Type-I GPA model providers shall undergo external or third party conformity assessment. This is involve hiring the services of a notified body that has demonstrated the professional integrity and the requisite competence in assessing GPAI models conformity, as per article 33(9) of the AI Act, considering “the requisite competence in the specific field” mentioned therein as the requisite competence in a given architecture and modality of GPAI (e.g. system-of-experts architectures vs pure-play deep learning, image generation vs language generation; etc.). The notified body shall cross-assess the internal conformity assessment stipulated in the previous tier.
- Tier 3: Type-II GPAI model providers shall undertake cross conformity assessment i.e. Type-II models require both an internal and third-party conformity assessment. For both types of conformity assessment, joint liability should be imposed on both the provider being audited and the auditor, regardless of liability insurance. As it stands out from the usual AI Act parlance, we detail it in section 8.3 as a standalone requirement. The notified body shall cross-assess the internal and external conformity assessments stipulated in the previous two earlier tiers.

Why: Presumption of component conformity is key to ensure that high-risk system providers can be legally certain they won’t have to redo GPAI conformity assessment. They must be clearly allowed to rely on the conformity assessment of the GPAI model or generative AI as evidence of conformity for the related component of the high-risk AI system. This significantly incentivizes purchase of AI Act-compliant GPAI and generative AI. This is facilitated by the De Facto approach mentioned earlier. Moreover, this upstream conformity assessment system ensures that the near-monopoly of knowledge, skills and technical access is addressed, in line with the De Facto control approach, redressing some of the issues related to concentration of power (challenge 3) and corporate irresponsibility (challenge 4), thanks to greater supplier-customer scrutiny.

How: Conformity assessments are common tools widely described in the AI Act and practised in a broad range of industries worldwide. Certificate of conformity for a specific component shall be made available to the downstream customers for their own use. The De Facto control approach discusses in detail the operationalization of the exchange of information necessary.

8. Tiered-approach: requirements and obligations to tackle GPAI and generative AI

Following the identification and extensive discussion of the regulatory targets' underlying concepts (Generative AI, Type-I GPAI and Type-II GPAI), of the seven challenges, and of the enforcement measures and complicating factors, we proceed to the detail of what regulated entities in each Tier should do in order to mitigate these challenges in a cost-effective but sufficient manner.

8.1 Tier 1: Measures for generative AI and Type-I and Type-II GPAI

Summary

The main requirements to effectively govern generative AI applications include:

- **Data Governance**, ensuring that providers' application is developed on the basis of adequate data sets.
- **Minimum content-moderation safeguards**, ensuring that the application is developed so as to prevent the generation of content in breach of union law.
- **Labelling AI-generated output**, ensuring that the output of the model is automatically accompanied by an indication that it has been artificially generated or manipulated.
- **Transparency on Model Used**, clear indication of the model name, model version and model provider's name to users in end-user-facing access interfaces.

These also apply to GPAI models.

A provider of a generative AI application, irrespective of how it is placed on the market or put into service, including as open source software, should be mandated to comply only with the following requirements. For the purpose of complying with these obligations, providers of such models shall follow the conformity assessment procedure based on internal controls. These requirements also apply to Type-I and Type-II GPAI models (cf. Sections 8.2 and 8.3).

8.1.1 Data Governance

What: Providers must ensure that their model or system is developed on the basis of data sets that are relevant, representative, and to the best extent possible, free of errors and complete. These characteristics of the data sets may be met at the level of individual data sets or by combining several data sets.

Why: Poor representation of minorities and of diversity in general in the dataset being fed into an AI system is a fundamental cause of the system's bias and resulting discrimination (Corporate irresponsibility, Challenge 4; and Incidents & accidents, Challenge 7). As investing in higher-quality data and data governance is a costly endeavour, and as it is difficult for a customer to know whether a provider has indeed invested in that direction, the best-in-class developers with excellent data governance practice are not rewarded by the market. A legally-binding data governance is therefore necessary. The same can be said for the tendency of generative AI applications and GPAI models to generate false information (e.g. for disinformation purpose) and hallucinations; or on the contrary to generate correct output that violates privacy, trade secrets, or copyrights.

The performance, accuracy, fairness, and safety of any AI model or system depends on the quality of the data it is trained on. Performance in turns determines the behaviour of the model in out-of-distribution situations, where incidents most often occur (Incidents and accidents, Challenge 7). By requiring data to be relevant, representative, and as error-free as possible, the Act aims to reduce the risks of AI systems making erroneous decisions or predictions that could potentially lead to incidents and accidents.

How: Providers should follow extensive guidelines already established for GDPR compliance. The monitoring and regulation of these practices will be overseen by relevant authorities, ensuring GPAI and generative AI providers uphold these standards when developing their models. Where relevant, the regulated entities will also apply the copyright directive to generative AI [189].

8.1.2 Content Moderation Safeguards

What: Providers must ensure the model or system is designed and developed in such a way as to ensure adequate safeguards against generation of content in breach of union law, without prejudice to fundamental rights, including the freedom of expression.

Why: There have been instances where the 'move fast and break things' culture of tech innovation (Challenge 4) has led to disregard for potential societal impacts. These cases highlight the need for stringent safeguards to ensure the respect for laws and rights when developing AI systems. Furthermore, without adequate safeguards, AI systems can be exploited to generate content that furthers illegal activities or infringes upon fundamental rights.

How: Internal audits, carried out both pre- and post-development of the model, can provide a comprehensive mechanism to ensure compliance with this requirement. The audit team can check whether the safeguards are working as intended and propose modifications if necessary. The audit team can periodically reassess the safeguards in place, conduct spot checks and red teaming exercises, and review any user reports of inappropriate content generation. The Generative AI application industry can learn a lot from the Digital Service Act and the surrounding community of practice working on making content published on digital platforms trustworthy and compliance.

8.1.3 Labelling Output

What: Providers must ensure that the output of the model is automatically accompanied by an indication that the output has been artificially generated or manipulated. This disclosure shall be carried through the value chain until the end user. The Commission is empowered to adopt delegated acts in accordance with Article 73 to specify and update modalities for this indication.

Why: Disclosure requirements can help prevent misuse of Generative AI applications (Challenge 5), such as deploying AI to spread misinformation or conduct fraudulent activities. If users are aware that an output has been artificially generated, they may be more cautious and less likely to be misled.

How: To ensure transparent usage of Generative AI applications, visual labels or watermarks can be used for AI-generated visual content, auditory cues for audio outputs, and textual disclaimers for written outputs. Furthermore, for online digital content, metadata attached to the output can be utilised to signify that it was artificially generated or manipulated.

8.1.4 Transparency on Model Used

What: Providers must at all times clearly indicate the model name, model version and model provider's name to users in end-user-facing access interfaces. If a general purpose AI model or combination of models or generative AI systems use a component general purpose AI model, the provider using that component shall at all times clearly indicate the component general purpose AI model's name, version and original provider's name to users in end-user-facing access interfaces.

Why: Users have a right to know which technologies are shaping their digital experiences and making decisions that may impact them. This helps to build user trust and understanding of Generative AI applications, while also making the market more liquid: if a scandal with a specific application occurs, and we know the underlying model, the consumers can "penalise" that model's provider by avoiding to use its model in other applications in the future. The widespread lack of transparency currently prevents such a feedback loop, but as soon as the feedback loop is enabled, it will curb Corporate Irresponsibility (Challenge 4)

This is important due to the economic ubiquity and, more broadly, the infrastructural aspect (Challenge 1), which can easily lead to a situation where downstream users do not know about the complexity of the GPAI value chain, and who could be responsible. In cases where a GPAI model uses a component from another GPAI model, indicating the original provider's name ensures that the creators of the underlying technology are recognized and can be held accountable if necessary (e.g. in the event of malfunctions or misuse).

How: This would be fairly straightforward for the customer-facing deployer to implement as the relevant information about the model or component model should be provided within the software interface, documentation, user manuals, FAQ etc. Upstream providers of GPAI models may be able to programmatically communicate this information via the API. For example, upon initiation or when queried, the model could output a statement like "This service uses GPAI model X, version Y, provided by Z."

8.2 Tier 2: Measures for Type-I and Type-II GPAI

Summary

In addition to requirements from the generative AI application's tier, the main requirements to effectively govern Type-I GPAI models are:

- **Risk management system**: GPAI providers establish, implement, and maintain a risk management system for the model in a process spanning the model's entire lifecycle.
- **Basic trustworthiness**: the provider proves that the model is designed so as to have sufficient levels of cybersecurity, predictability, interpretability, corrigibility, controllability, robustness and boundedness.
- **Reporting of compute resources**: GPAI providers create systematic processes to forecast, record and report regular use of compute resources for training runs and model operation, along with the energy use associated.
- **Quality Management System**: GPAI providers implement a thorough quality management system that guarantees adherence to the stipulations of the AI Act concerning GPAI models.
- **Compliance function and officer**: GPAI providers establish an autonomous compliance function, separate from the operation of the organisation, and staffed by one or more compliance officers responsible for monitoring the provider's adherence to obligations set out under the AI Act regulation.
- **Notification of training runs & model pre-registration**: GPAI providers notify the AI Office of upcoming training runs, models under development, and pre-register models in their pipeline.
- **Know-your-customer,** to facilitate prevention of misuse: GPAI providers take all necessary and proportionate measures to prevent misuse after detection.

These requirements do not apply to generative AI applications; but they do apply to Type-II GPAI models.

A provider of a Type-I GPAI model, irrespective of how it is placed on the market or put into service, including as open source software, should be mandated to comply with the following requirements, in addition to the requirements described in the previous tier (Section 8.1). For the purpose of complying with these obligations, providers of such systems shall follow the third party conformity assessment with the involvement of a notified body that has demonstrated the professional integrity and the requisite competence in assessing GPAI models conformity. These requirements also apply to Type-II GPAI models (cf. Section 8.3), but not to generative AI applications (cf. Section 8.1).

8.2.1 Risk Management System

What: The GPAI provider is required to establish, implement, document, and maintain a risk management system for the model. This risk management system should be an ongoing, iterative process that spans the entire lifecycle of the model, inclusive of steps taken prior to taken such as conceptualisation, design, and development.

When determining the most suitable risk management strategies, the provider is expected to prioritise risk elimination or reduction as much as possible through suitable design and development. If risks cannot be completely eradicated, adequate mitigation and control measures should be in place. It is also the provider's responsibility to inform both users and authorities about all identified risks and the actions taken to mitigate them.

Why: GPAI-related risk assessment and management can be more effectively performed by the principal GPAI developer, as they are typically the ones with the most in-depth understanding of the technology, compared to the vast number of downstream applications.

As GPAI models have the potential to be used by thousands of downstream users, and are increasingly integrated into various sectors of the economy (Challenge 1, infrastructural aspect), their impact becomes more extensive. Ensuring the source model doesn’t present risks is therefore an investment whose return is multiplied by as many downstream users. This necessitates a robust and continuous risk management system that can anticipate and mitigate risks before they spread across different sectors. Relying on each user of an infrastructure like a bridge or a power grid to carry out risk management of the infrastructure is ill-advised for two reasons: First, the user has little knowledge, skills or access over the infrastructure risk-relevant feature, so they cannot achieve risk mitigation. Second, it would imply each user repeats the (ineffective) risk mitigation measures of thousands of other users before them, plaguing society with cumulative collective costs incomparable to the cost for the infrastructure provider to establish a risk management system once and for all.

Moreover, a risk management system is crucial in identifying potential avenues for misuse or malicious use of the GPAI model both in isolation and when it is incorporated within a wider system by downstream deployers. (Challenge 5). Finally, the use of GPAI models raises not just conventional risks, but also existential risks - those that could potentially cause humanity's extinction or bring about widespread suffering (Challenge 2). These risks, if not managed properly, could pose severe threats to humanity's long-term survival and prosperity, and are therefore arguably worth managing.

How: Most industries and most medium or large companies in these industries, including many software and system engineering activities, carry out risk management, either standardised or mandated by regulations, such as ISO 31000 [190]. Moreover, some AI-specific risk management methodologies have emerged. GPAI providers are encouraged to use the National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF)'s profile on general purpose AI systems as a blueprint for setting up their risk management system [191]. The AI RMF "core functions", used to categorise the range of activities within the framework, are as follows: "Map" for identification and contextual understanding of potential AI risks; "Measure" for quantifying AI trustworthiness characteristics; "Manage" for decision-making on AI risks, whether it is prioritising, evading, mitigating, or accepting them; and "Govern" for a set of policies, roles, and responsibilities which oversee the AI risk management process.. The AI RMF encourages consideration of systemic and societal-scale risks, in addition to risks to individuals and groups. Barrett et al. (2023) [191] supplement this framework with additional guidance on risk assessment and mitigation strategies to proactively address use-agnostic risk factors.

8.2.2 Basic Trustworthiness

What: In order to instil trust in the upstream GPAI model—which has the potential to serve as the technical foundation for numerous downstream systems—it should be designed and developed to attain sufficient levels of:

Cybersecurity: the protection of the AI system from digital attacks. It includes measures to safeguard the integrity, confidentiality, and availability of the system and its data.

Interpretability: the extent to which the internal workings of an AI model can be understood by humans. An interpretable model allows users to comprehend why and how it arrived at a particular outcome or decision. This characteristic is critical for trust, especially in high-stakes domains where understanding the AI's decision-making process is essential.

Predictability: the capacity of the AI model to behave as expected and produce consistent outcomes. Predictable AI is more easily understood, trusted, and controlled.

Corrigibility: the ability of an AI model to be corrected or adjusted over time. This could involve learning from mistakes, adapting to new data, or allowing human operators to intervene and guide the model's behaviour or decisions.

Controllability: the extent to which the feedback loop between model and human operator is effective. Specifically, the extent through which the human user or even a developer has the ability to modify the model or its training programme in a timely, meaningful and sustainable fashion so as to impose its preferences upon the model while the model and its output affects the human user.

Robustness: the ability of an AI model to perform consistently and accurately in a variety of different conditions, including those it was not specifically trained for.

Boundedness: the likelihood that a system will only operate within a set range of contexts or tasks and does not exceed its specified limitations.

If models fail to exhibit these characteristics or lose them over their lifecycle, the provider is obliged to halt their development or retract the affected models.

Why: One of the key issues in the current industry is the near-total opacity of internal practices for trustworthiness. Publishing internal policies pertaining to this regulation and making internal compliance reports available to the authorities would enable greater scrutiny by customers. The industry players that prioritise and invest in meeting these requirements are duly recognized and rewarded. This prevents companies that might take shortcuts or neglect these important considerations from gaining an unfair advantage, and therefore addresses Corporate Irresponsibility (Challenge 2) Requiring these standards could also slow down the race towards deploying AI systems without adequate checks and balances in place. It would encourage companies to prioritise safety and reliability over speed to market, helping to mitigate the risks associated with rushing untested or potentially unsafe AI systems into use.

Moreover, this requirement creates a market for testing & evaluating models. Combined with benchmarking capabilities as per Article 58b in the EP's compromise text, this requirement not only stimulates participation in standard-setting but also encourages Research and Development (R&D) to discover improved methods of enhancing these key dimensions of AI models. This could help industry players compete over quality, as opposed to competing over unhinged generality of capabilities.

How: GPAI providers should also establish the capacity for multiple types of rollback in response to unexpected degradation in trustworthiness or unintended outcomes of their deployed models.[46] One of the primary strategies to consider is user restrictions, which would involve blacklisting specific users or groups who misuse the technology. Another effective strategy could be limiting access frequency. For instance, implementing measures that limit a model to produce a certain number of outputs per hour, which would prevent overuse or abuse.

Furthermore, modifications in the model's capabilities can serve as a measure to increase its overall trustworthiness. These modifications might encompass filtering certain outputs or narrowing a model's context window to limit its reach. Lastly, providers should also consider usage case limitations, such as forbidding the application of the model in high-risk situations or sectors. In highly critical instances, a GPAI provider should trigger a shutdown of the system. This

[46] https://arxiv.org/pdf/2305.15324.pdf

would involve, at minimum, promptly withdrawing the affected model from circulation and, if necessary, also retracting earlier versions of the model.

8.2.3 Reporting of Compute

What: GPAI providers are obligated to create systematic processes for forecasting, recording, and reporting the utilisation of compute resources for training runs and model usage, along with the associated energy consumption of said compute.

Why: As discussed in box 1, the compute usage can aid in distinguishing between Type-I and Type-II models, helping to implement appropriate risk mitigation strategies and regulatory oversight corresponding to each type. Certainty about whether a model will be Type-I or Type-II is key, given some of the requirements apply prior to training.

Compute resources necessary for training large language models (such as NVIDIA's H100 chips) are increasingly concentrated, notably in the hands of a few players in which NVIDIA is a major shareholder or with whom they have a partnership, making it increasingly difficult for startups to access this crucial resource. Reporting compute would provide the basis of a system to assess the concentration power (Challenge 3). As consumption of computing resources also leads to significant environmental concerns, its reporting for public scrutiny could also curb some of the most irresponsible behaviours (Challenge 4).

The internal recording and monitoring of compute required for reporting compute is also a foundation for enabling the enforcement of other obligations and requirements, notably the verification and restriction of access to the model necessary for Know-your-customer and notification of training runs, and helps therefore curtail misuse and malicious use (Challenge 5). Reporting of compute is therefore a first step and a low hanging fruit for further cost-effective governance, such as on-chip mechanisms and other forms of large training runs monitoring.[47]

How: This information is typically readily available within these organisations, thus complying with a requirement for its disclosure should not present a significant hurdle for GPAI providers. AI companies already account for compute as it's a significant cost factor, and compute providers also must account for this data for billing purposes. Additionally, many organisations already measure and report their environmental impact through Corporate Social Responsibility (CSR) initiatives, which indicates that a framework for this kind of disclosure is often already in place.

[47] Timothy Fist, Onni Arne, and Caleb Withers "On-Chip Mechanisms for AI Governance" (Center for a New American Security, forthcoming)

8.2.4 Quality Management System

What: Providers are mandated to implement a quality management system that guarantees adherence to the stipulations of the AI Act concerning GPAI models. This system should be meticulously documented in a methodical, organised fashion, encompassing written policies, procedures, and directives. The implementation scale of the quality management system should be commensurate with the provider's organisational size.

Why: A robust Quality Management System would enforce a culture of quality and compliance within the organisation, and the GPAI industry more generally, offsetting some of the Corporate Irresponsibility (Challenge 4). By establishing clear policies and procedures regarding quality management, the organisation also builds its ability to comply cost-effectively with the AI Act.

How: To build an effective Quality Management System (QMS), GPAI providers can draw guidance from established standards such as ISO 9001 [192], which is internationally recognised as the benchmark for QMS across various industries. The principles central to ISO 9001 include a strong customer focus, the involvement of top management, a process approach, continual improvement, fact-based decision-making, and mutually beneficial supplier relationships.

8.2.5 Compliance Function and Officer

What: Providers are required to establish an autonomous compliance function, separate from the operational arms of the organisation, and staffed by one or more compliance officers. These officers shall be responsible for monitoring the provider's adherence to obligations set out under the AI Act regulation. They will ensure that the aforementioned risk management system identifies all reasonably predictable risks, including but not limited to, those related to fundamental rights, health and safety made prominent in public discourse, by stakeholders, or by representatives of vulnerable groups.

The head of the compliance function shall be professionally responsible for signing off and dating the code base prior to training and execution of the model.

Why: This practice prevents non-compliant projects from mistakenly being developed, which would pose a moral hazard to the organisation given their substantial unrecoverable development costs: once a non-compliant Type-I GPAI model is trained, disregarding it corresponds to admitting wasting several millions of dollars in training costs. Attempting to make the model compliant ex post, through backdating documented tests, will therefore be the course of action of the least trustworthy developers, unless there is internal oversight. As such, the compliance function mitigates conflicts of interest. This separation ensures that the compliance officers can objectively monitor adherence to obligations without being influenced by the operational needs or financial pressures of the organisation.

How: To implement an effective compliance function, the AI Office can adapt tried and tested methodologies from the banking and other regulated industries. In the digital sector, the DSA's framework requires providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to establish an independent compliance function, composed of qualified compliance officers. The head of this function directly reports to the management body and cannot be removed without prior approval of the management body of the provider, ensuring independence and continuity.

8.2.6 Notification of Training Runs & Model Pre-registration

What: Providers of types I & II GPAI must notify the AI Office of upcoming training runs for such models, including the amount of training compute predicted for their model's training and the upper limit of training compute agreed internally. These notifications ensure compliance officers and authorities can verify that the provider does not accidentally find itself developing a Type II GPAI (which comes with additional risk and, therefore, additional compliance requirements).

Moreover, to facilitate workload planning for the compliance function and relevant authorities as the AI Act is implemented, providers must notify the AI Office of any models currently under development and keep authorities updated as to the pipeline of development for upcoming models. These notifications will be handled in accordance with rules for confidentiality and protection of sensitive data detailed elsewhere in the AI Act (Article 70).

Finally, providers must pre-register the models whose training is notified to the AI office in a public EU database (without providing details of training compute required or predicted), to ensure transparency vis-a-vis society. This public pre-registration is meant to facilitate inquiries by sectoral authorities, civil society, trade organisations, unions, media, experts … with regards to the implications of the model for society and to discuss mitigation measures, pre-empting conflict and "fait accompli" roll-outs.

Use of pre-registration or of the preliminary classification as Type I or Type II models, whether undue or factually correct, shall not be used in providers' marketing or fundraising material or in any way that may lead an audience to believe that a conformity assessment has been successfully carried out.

Why: notification and pre-registration requirements help ensure that regulators dedicate sufficient attention to models that necessitate it, while having advance notice on developments in the industry (helping address Challenge 6, technical opacity and Challenge 3, concentration of power). This helps mitigate the technical advantage developers have and help authorities take action to prevent harm and guide providers away from non-compliance, contrary to current industry practice of rolling out first and asking questions later. This requirement therefore encourages a culture of compliance and respect for product safety laws within the GPAI industry (Challenge 4)

How: The notification to the AI Office, with the associated transfer of relevant information, can be done through a secure, encrypted channel or in controlled private networks, such as virtual data rooms (established by the AI Office). The same private network can be used for maintaining the AI Office abreast of models under development. Pre-registration is carried out in much the same way as the registration in the EU database.

8.2.7 Know-Your-Customer

What: When the provider detects or is informed about market misuse they shall take all necessary and proportionate measures to prevent such further misuse, in particular taking into account the scale of the misuse and the seriousness of the associated risks. To this end, the provider shall conduct thorough Know Your Customer checks by collecting any business user's intended purpose with the system prior to enabling use of its model, as well as the user's electronic means of contacts and match it to a unique customer identifier. For models served through an API, the provider shall continuously sample server requests and assess whether they match the intended purpose provided by the business user.

For models served as Open Source software, the provider, the AI Office, and if applicable, the hosting platform and Open Source foundations shall cooperate to establish a collaborative verification system, such as a trusted buyer system, and a joint open source observatory to facilitate OS providers assessment of misuse and malicious use trends.

Why: This requirement helps to prevent misuse by the potentially numerous downstream users of the GPAI model (Challenge 1, infrastructural aspect) and ensure that it is used responsibly and in a way that aligns with its specifications (helping address Challenge 5, misuse). As the scale and complexity of GPAI models increase, so too does their potential for misuse, particularly by downstream business users. By enforcing strict KYC checks, continuous assessment of server requests, and collaboration for open-source misuse detection, the potential for malicious uses of GPAI can be substantially mitigated.

How: KYC checks conducted by the GPAI model providers would involve the collection of the business user's intended purpose with the model, electronic means of contact, and matching these to a unique customer identifier. This step will provide an initial layer of scrutiny, ensuring that potential misuse can be identified and prevented at the outset.

Additionally, a simple, standardised mechanism to report misuse or non-compliance of a GPAI model should be put in place for the public, model providers, and authorities. This can range from a straightforward complaint mechanism for the public, to upstream providers being able to detect and act on misuse, to regulatory bodies having the ability to investigate and enforce laws against non-compliant entities. This process could be similar in kind to mechanisms used in the Digital Services Act, which allow users to flag illegal content online, and for platforms to cooperate with specialised 'trusted flaggers' to identify and remove illegal content [193].

Lastly, the provider should establish robust protocols to either rollback features or completely shutdown the model in the event of widespread misuse or catastrophic misuse. The choice between these actions hinges on the scale of the misuse and the risks involved. Shutting down the model is a step that should be taken in cases where the risks of maintaining its availability could have devastating consequences, especially if it jeopardises multiple downstream applications

Rollback options specifically relevant for misuse scenarios include: blacklisting or whitelisting specific users or groups, the removal of specific capabilities of the model (e.g., pathogen design), limiting the ability of the user to fine-tune the model, restricting the model from being used in high-stakes applications, and limiting the ability of a model to interact with downstream tools (e.g., to use other APIs), to make function calls, to browse the web, etc.

There are several non-mutually exclusive rollback strategies that are particularly relevant for misuse scenarios. For instance, providers can implement user-based controls, such as blacklisting or whitelisting particular users or groups. They might also consider stripping the model of specific functionalities, like pathogen design, to limit its potential for harm. Another option is to limit the extent to which users can fine-tune the model to prevent said users from introducing malicious or misguided alterations. Providers should also consider restricting the model's deployment in high-stakes environments, further safeguarding against potentially catastrophic incidents of misuse. Lastly, limiting the model's capability to interface with downstream tools — such as other APIs — or to execute functions like web browsing, ensures additional layers of safety.

8.3 Tier 3: Measures for Type-II GPAI

Summary

In addition to the requirements from the other tiers, the main requirements to effectively govern Type-II GPAI models are:

- **Regular dialogue with AI Office to update on latest technical advancements** in AI to reduce the knowledge gap between the developers and the Office, through the Navigator Programme.
- **Internal & 3rd party auditing**, imposing joint & several liability on both the provider being audited and the auditor.

- **Absolute trustworthiness,** or that providers design and develop their models to achieve superior levels of advanced cybersecurity and safety.
- **Quality-by-Design process:** augmenting the mandated quality management system for Type-I models with a Quality-by-Design (QbD) process that includes a probabilistic risk assessment and safety evaluation, akin to drug manufacturing protocols.
- **Review & Approval of designs** by AI Office before training run or that the provider notifies and awaits an opinion from the AI Office, with the authority to delay training runs designated for developing Type-II GPAI models and to review the codebase.
- **Major accident prevention policy**, developed by providers and meticulously implemented, to protect human health and the digital, physical, and natural environments, similar to that of the Seveso Directive and other production processes.
- **Responsible Staged Development & Release,** whereby providers structure their design and development process to scale responsibly and cautiously, with batteries of tests and evals at every checkpoint to be satisfied in order to continue training.
- **High-Reliability Organisation,** or that providers organise their facilities, processes and internal policies as a way to incorporate all other requirements in the practice of the provider and to establish a culture valorizing reliability, safety & trustworthiness.

These only apply to Type-II GPAI models, not Type-I nor generative AI applications.

A provider of a Type-II GPAI model, irrespective of how it is placed on the market or put into service, including as open source software, shall be mandated to comply with the following requirements, in addition to the requirements outlined in the previous two tiers (cf. Section 8.1 and 8.2). For the purpose of complying with these obligations, providers of such models shall follow the cross-audit (internal & 3rd party) conformity assessment discussed in Section 8.3.2.

It is worth noting that the senior management of these select providers have repeatedly and publicly expressed their concerns about the catastrophic risks posed by Type-II GPAI models. They have called for regulation, comparing the potential danger of these models to that of pandemics or nuclear wars. This, combined with the surgical scope (around 28 Type-II models to date, by around 10 providers), renders this tier particularly strict to ensure that these rare providers reduce the risk they pose to society below a more acceptable risk threshold than what they are at today.

8.3.1 Dialogue in Navigator Programme

What: the provider is required to participate in regular dialogue sessions within the relevant subgroup of the AI Office.

Why: This ongoing dialogue would nurture a mutual understanding and foster a trust-based relationship between the developers and the AI Office. Given the fast-paced innovation in the field, measures must be taken to ensure that the regulator stays up-to-date with the latest technical advancements, mitigating Challenge 6 (technical opacity) and rebalancing power (Challenge 3). The providers, on the other hand, benefit from better understanding the implications of their models and gaining insight into the regulatory perspective and requirements, which they can incorporate early into their development processes (“compliance-by-design”), which helps mitigate Challenge 4 (corporate irresponsibility). The level of scrutiny, and special attention and bandwidth attributed to Type-II GPAI model provider should be seen as a reflection of their infrastructural aspect and the capability risks they pose (Challenges 1 and 2)

How: The navigator program (discussed at greater length in Section 7.1.1) would facilitate ongoing dialogue between the GPAI providers and the AI Office by assigning a dedicated AI Office staff member to each of the identified GPAI provider teams. These staff members would act as intermediaries, ensuring regular, monthly communications between the two entities. A full detail of the navigator programme's implementation is available in earlier work, to be adapted to current circumstances [6].

8.3.2 Internal & 3rd Party Auditing

What: Type-II models require both an internal and third-party conformity assessment. For both types of conformity assessment, joint & several liability should be imposed on both the provider being audited and the auditor. The third party auditor must be competent specifically for the auditing of GPAI models, like for the Type-I GPAI model providers (as discussed in Section 7.4.2)

Why: The intermingling of internal and 3rd party audit is one of the most straightforward way to establish an ecosystem of corporate responsibility (addressing challenge 4). Indeed, because of the near-monopoly of knowledge about Type-II GPAI models, which lie at the cutting-edge, mandating internal auditing will require provider to leverage internal expertise, rather than outsource the internal conformity assessment. The internal auditors therefore provide narrow expertise on the providers' technologies, while the 3rd party auditors provide the expertise from the rest of the industry and even other industries in terms of compliance. The 3rd party auditors are incentivized -through the joint & several liability- to leverage the internal auditors' expertise as much as possible but cross-check it through their work, while internal auditors are incentivized to leverage 3rd party auditors' independent opinion and expertise abroad for how best to comply.

This approach also ensures that this rare internal expertise is cross-evaluated by a competent third party. This ensures that no single entity has absolute control over a model's approval process and avoids regulatory capture - addressing Challenge 3 (Concentration of Power). Joint & several liability incentivises both the provider and the auditor to perform more meticulous assessments to prevent mishaps, hence avoiding financial and reputational damage, providing a strong incentives to prevent incidents and accidents (Challenge 7)

How: The implementation of both internal and third-party conformity assessment for Type-II models could be modelled on the independent audit requirements set out in the Digital Services Act (DSA). The DSA stipulates that providers of very large online platforms and search engines are required to undergo independent audits at least once a year at their own expense. As detailed in Section 7.4.2, audit organisations should demonstrate to the notifying authority a satisfactory level of expertise in risk management, technical competence, and adherence to ethical codes of practice or standards as they pertain to GPAI specifically (Article 33(9) of the AI Act) considering "the requisite competence in the specific field" mentioned therein as the requisite competence in a given architecture and modality of GPAI (e.g. system-of-experts architectures vs pure-play deep learning, image generation vs language generation; etc.)

Furthermore, the auditor must be independent, without any conflicts of interest with the provider. They should not have provided non-audit related services to the provider within the 12 months before the start of the audit and should commit not to provide such services in the 12 months after the audit.

8.3.3 Absolute Trustworthiness

What: Providers are mandated to design and develop their models to achieve maximum levels of advanced security and safety. Security involves information security ("InfoSec"), operations security (OpSec) and cybersecurity. Safety includes demonstrating restrictions on model abilities that could harm developers and beta-testers, the environment, whether physical or digital, or users, whether physically or psychologically. These abilities include social engineering, cyberattacks, autonomous planning, self-replication or inner understanding of how to harm a significant number of people (e.g. ability to instruct users on making viruses or bombs). Should a Type-II GPAI model fail to exhibit the advanced levels of security and safety during development, or fail to maintain them throughout their lifecycle, the provider is obligated to halt their development or retract already deployed models (roll-back & shutdowns).

Why: The risks of accidents (identified in challenge 7) combined with the corporate irresponsibility (challenge 4) and plausibility –according to Type-II GPAI developers themselves– of generality of capabilities posing existential threats (challenge 2) justify a request by society to guarantee absolute trustworthiness. Type-II GPAI models, possessing more advanced capabilities than Type-I GPAI models, are envisioned by investors and AI companies' CEOs alike as ultimate products automating and running the creation of other products and services, significantly contributing to increasing automation in the economy [194]. Their owners would therefore accrue significant profits and concentration of power (Challenge 3). Yet, this potential internalised upside comes at a significant -potentially existential- risk for all worldwide, should the model be "misaligned" [195]. Because of these stakes and extreme conflicting incentives (significant internalised reward for developer, at the risk of significant externalised costs for society), a principal agent problem arises: the principal -society- will be betrayed by the agent -developers of GPAI-II models- due to moral hazard. Requirements and obligations to re-establish agent's trustworthiness as much as possible is therefore necessary.

Moreover, the risk of misuse or malicious use (challenge 5) is proportional to the generality of capabilities and widespread, particularly affecting Type-II GPAI. By setting exceptionally high standards for cybersecurity and implementing restrictions on tendencies for social engineering and the dissemination of hazardous information (e.g. indications for the fabrication of bombs or viruses), this requirement reduces the risk of malicious use or the risk of unauthorised access to the model.

How: Absolute trustworthiness naturally builds upon the Risk Management System. Developers and researchers are dedicating significant time to creating and improving ways to assess models' abilities through testing, evaluations, validation and verification processes. The use of red-teaming (by independent experts) is a way to discover bugs, vulnerabilities and dangerous capabilities. Compliance might require red-teaming at every training checkpoints to avoid wasting training compute on models and to spot early on the apparition of these capabilities.

The pooling of European benchmarking capabilities will facilitate the harmonisation and comparison of testing protocols, building an epistemic community and knowledge base along the way and, eventually, to predict model's riskiness and generality of capabilities.

Having third-party audits and adhering to recognized cybersecurity and AI safety certifications can provide an additional level of assurance that the GPAI model is safe and secure. Overall, a compliance-by-design approach integrating compliance with safety measures and autonomous replication evaluations [196] into the design of the model, rather than trying to incorporate them in later stages, would aid in cost-effectively achieving higher levels of cybersecurity and safety, given significant training costs. The tools for trustworthiness and associated "trustworthy AI" companies that have emerged for assessing and even predicting whether a model will be compliant (based on simulation and comparison with previous versions of the model) are therefore valuable in the Type-II GPAI developers' toolkit.

Moreover, beyond the obligation to roll back certain features of models during minor incidents or misuse, as detailed in the previous section, providers of Type-II GPAI models must implement a robust mechanism that facilitates their rapid and complete shutdown when necessary. The objective of this measure is to prevent an ongoing failure from escalating into causing widespread and significant damage. Maintaining and implementing a rollback and shutdown mechanism can be a four-part process, consisting of preparation, monitoring, incident response, and post-incident recovery, similar to the NIST computer security incident handling guide [197].

Emergency shutdown planning would involve building tools and procedures for swift incident responses, which encompass threat modelling [198], establishing decision-making authorities [199], and sharing best practices [13]. The providers would then be required to continuously monitor the model's capabilities and behaviour, detecting any anomalies [183] and escalating cases of concern to relevant decision makers. Should a shutdown become necessary, providers must act decisively, coordinating with regulatory authorities, activating failback systems for downstream users, and promptly alerting customers [184]. Finally, the provider is required to carry out several post-incident reviews, feeding lessons back into the previous stages and liaising with external stakeholders where necessary [197].

8.3.4 Quality-By-Design Process

What: Providers of Type-II models are required to supplement the mandated quality management system for Type-I models with a Quality-by-Design (QbD) process. This augmentation must include a probabilistic risk assessment and safety evaluation. Such measures aim to predict and ensure the quality of the model through statistical, analytical and risk management methodology in its design, development and deployment. Furthermore, these processes should ensure that the model consistently aligns with its predefined characteristics right from inception. In order to achieve the sufficient level of quality control over the model developments, this would require a holistic approach to model development: every aspect of the development facilities and team become a variable to be controlled for in order to achieve highest-quality through the training process.

Why: The significant economic and environmental costs associated with building a Type-II GPAI, coupled with the current lack of solutions to assure these models' reliable functioning as intended make quality control throughout the design and development stage important. Moreover, QbD is necessary for developers to build capacity to predict what models are likely to look like at each checkpoint. This aspiration is necessary to address Challenge 2 (Generalisation & Capability Risks), Challenge 4 (Corporate Irresponsibility), Challenge 6 (Technical Opacity) and Challenge 7 (incidents & accidents) - it is therefore fundamental for the proper governance of Type-II GPAI.

How: The implementation of a QbD process for Type-II models can draw significant insights from the approach employed by the European Medicines Agency (EMA) for quality assurance in medicine development and manufacturing. Similar to the EMA's approach, the QbD should incorporate statistical, analytical, and risk-management methodologies. Establishing QbD protocols will be a demanding process given the industry current lack of governance, but a necessary one given the challenges it resolves. Collaboration with the EU AI Office to guide the creation of these industry-wide QbD protocols is highly recommended.

8.3.5 Review & Approval of Designs

What: The provider is required to notify the AI Office of impending training runs designated for developing Type-II GPAI models and must grant the Office and other competent authorities access to review the codebase and designs of the model, whilst ensuring the protection of trade secrets (as per Article 70). The AI Office is obliged to present a positive, undecided, or negative opinion concerning the expected compliance of the training runs and model design within 30 days of receiving the initial notice. This opinion should include the reasoning behind the judgement, to which the provider must respond in writing.

The Office also has the right to issue a second opinion, providing an extended explanation, within 15 days following the receipt of the provider's response. The provider then has the option to reply in writing within a 15-day timeframe. If the provider's decision contradicts the AI Office second negative opinion, the Office, within 15 days of receiving the second response, has the authority to declare the proposed model as non-compliant with the AI Act.

Why: Given Challenges 2, 3, 4 and 6 (Capability risks, Concentration of power, Corporate irresponsibility and Technical opacity), the decision to train a cutting edge model like a Type-II GPAI should be heavily scrutinised by a party independent from the developer itself. The exchange of opinions ensures timely feedback and aligns expectations in a co-regulatory manner. Given the global aspect of the development of Type-II GPAI, unless these opinions are expressed by a coalition of regulators worldwide, they cannot easily be made binding or would risk causing regulatory flight.

How: This type of pre-production review is common in several industries under various forms: in the electronic hardware world, the US Federal Communication Commission's Equipment Authorization Approval applies to most radio devices. The procedure involves compliance testing -which requires various measurements of performance of the prototype- and approval [200]. While in the case of GPAI, one could argue the source code is insufficient to predict all the model's relevant characteristics, several characteristics can be inferred by detailing the development process, architectural decisions and software modules to be executed during production. Developers would therefore submit designs & blueprints before training, and would be held responsible for deviations from the blueprint. This is similar to how the construction industry's project developers must demonstrate their building project *will* comply with regulations in place prior to construction work beginning. The preliminary assessment of compliance is based almost only on analysis of blueprints and schematic designs.

The notification to the AI Office, with the associated transfer of relevant information, would be done through a secure, encrypted channel or in controlled private networks, such as virtual data rooms (established by the AI Office).

8.3.6 Major Accident Prevention Policy

What: Providers are mandated to construct and meticulously implement a Major-Accident Prevention Policy (MAPP) similar to that of the Seveso Directive. This policy must be designed to uphold an exceptional level of protection for human health and the digital, physical, and natural environments. It should incorporate the provider's fundamental objectives and action principles, clarify management's roles and responsibilities, and show a commitment to continually enhancing major accident risk control and maintaining superior levels of protection.

The MAPP must undergo a review at least annually, and its execution documented quarterly through a safety report submitted to the AI Office. The policy implementation should encompass suitable means and structures, including an early warning system and a safety management system responsible for producing a quarterly safety report. An appointed on-site safety officer will be charged with the enforcement of the MAPP and will hold professional responsibility for the co-endorsement and dating of the Type-II GPAI models' code base prior to execution or training, alongside the approval of the head of the compliance function as discussed in the previous chapter.

Why: Given the considerable potential for catastrophic accidents (Challenge 7) associated with Type-II GPAI models—a concern that has been voiced by the majority of Type-II GPAI providers and numerous experts, and corroborated by past incidents—it is imperative that these providers develop proactive strategies to prevent such accidents, rather than merely establishing reactive measures for mitigation. A Major Accident Prevention Policy, common in other industries, is the best way to address Challenge 7.

How: The Seveso-III Directive (2012/18/EU) [201], enacted by the European Commission in 2012, offers a solid benchmark for how accident prevention policies can be effectively structured. It outlines a thorough approach towards prevention of major accidents involving hazardous substances, providing valuable insights on how similar strategies can be adapted for GPAI model risk management, including a MAPP, assessment of domino effects, regular publication of provider-specific safety reports, establishment of a Safety Management System, performance monitoring, audits and reviews, communication plan, and establishment of an emergency response plan with early warning system. The Directive has been recognized as a model for industrial accident prevention policy, inspiring similar legislation globally [202].

Major accident prevention is also a part of several other regulatory regimes worldwide. For example, the EPA’s accidental release prevention requires accident prevention procedures to be put in place [203]. Regulated entities falling in its strictest level of risks (program level 3: strictest because workers are exposed directly in case of accidents) requires the US Occupational Safety and Health Administration [204] standard, additional hazard assessment and management measures such as Failure Mode and Effect Analysis (FMEA) and Fault Tree Analysis, as well as emergency response requirements akin to that in Seveso-III.

8.3.7 Responsible Staged Development & Release

What: Developers of a Type-II GPAI model must structure the design and development process in a responsible way. Specifically, the design and development process shall include the incremental scaling of compute per checkpoint, model size and data processes during the training, to ensure tests carried out at a given checkpoint are sufficiently informative about the likely outcome of the next checkpoint to predict whether redesign and adjustments are necessary. This process shall be documented to ensure best practices in compliance can evolve in line with developers' ability to predict or simulate models.

To do so, providers of Type-II GPAI shall at all time maintain documentation of the capabilities of their models and the policies & processes for containing the risks of their models development and at release. They shall document, before moving to the next increment, what additional risk-mitigating measures they have researched, designed, developed and tested to address the risk associated with the predicted level of capabilities associated with the next incremental stage. They shall demonstrate the effectiveness of the risk-mitigation measure prior to moving to the next incremental stage. To ensure a risk-based approach, they shall do so in line with the risk levels defined by the AI Office.

Why: Staged development and responsible scaling have long been proposed as a way to incentivize greater understanding of Type-II GPAI models during the development and to mitigate risks associated with unforeseen capabilities; all while limiting regulatory burden for less dangerous models. Similar proposals will start being institutionalised into hard law in some regions[205] and have already been applied internally by at least one Type-II GPAI provider.[206][48] Formalising these efforts also helps pace the development process, giving it the time needed for developing the De Facto control (cf. Section 7.4) over these models and fulfil the other requirements and obligations associated with the Type-II GPAI Tier.

This pacing is necessary because of the Technical opacity (Challenge 6): as Type-II GPAI systems are highly complex and opaque, it helps regulators, and even the developers themselves, better understand how these systems work and invest in a procedural "testing ground" for worst-case scenarios where such models exhibiting power-seeking behaviour or autonomous-replication capabilities, providing valuable information to help prevent such outcomes. This deliberate pacing and potential controlled environment of a sandbox scheme can help prevent accidents by allowing potential issues to be identified and fixed before progressing, which address risks of accidents & incidents from unexpected generality of capabilities (Challenges 2 & 7).

How: The AI Office, in collaboration with independent experts, civil society and industry, shall establish the factor of increment for compute (i.e. by what factor can compute resources dedicated to training can be multiplied between checkpoints) as well as the risk levels within 6 months after the entry in force of the AI Act. The AI Office may decide to obligate providers of GPAI systems that they think pose a particularly large threat to society to take part in a regulatory sandbox scheme to accompany this staged development process and release.

[48] Anthropic's Responsible Scaling Policy, Version 1.0

Moreover, the operationalisation of this pacing will require harmonising and formalising a risk categorisation per model, as already attempted by many. The process involves structured testing and assessment under controlled conditions, closely monitored by the AI Office. Developers shall provide comprehensive documentation of their system's design, training, and operation. They shall clearly report how they identify or test for potential risks without endangering staff or stakeholders, as well as preventative measures or corrective actions, and how these latter are designed and implemented. Through iterative testing and refining, the system's safety and reliability are confirmed before being approved for final training and full-scale deployment.

8.3.8 High-Reliability Organisation

What: Type-II GPAI providers shall organise themselves, their infrastructure, processes and internal policies in ways that ensure they carry out their work with high reliability and conducive to a rigorous safety culture to reduce incidents and accidents. Beyond the Tier 3 requirements outlined above, this involves precise organisation charts and lines of responsibility with an unusual degree of accountability, documented decisions, systematic investigation of incidents and deviance from predicted performance, robust incident reporting mechanisms, systematic study of past anomalies and scanning of potential risk scenarios, comprehensive quantitative and qualitative risk assessment, large amount of redundancies in control and regulating mechanisms providing resilience, and deference to experts independently of their hierarchical position.

Why: Given the scale of economic and financial risks (Challenges 1 & 2), the complexity of these models (Challenge 6), the history of incidents (Challenges 4 & 7) and the risk of extinction an incident causes (Challenge 2), providers of Type-II GPAI models ought to achieve a 0-incident target, regardless of the efficacy of requirements mentioned for this tier already. Beyond mandatory requirements, which may be perceived or depicted as “brakes” or “chains” upon the purported innovative spirits of the industry, providers therefore need a rigorous safety culture that embraces caution and the precautionary principle as opposed to default to the rapid pace of progress and incidents as inevitable [207]. The managerial concept of High-Reliability Organization (HRO) aims to achieve that and has been practised worldwide, by several industries developing high-impact technologies since at least the mid-20th century [208].

Moreover, HRO frameworks lend themselves particularly well to the successful compliance with requirements mentioned above, in particular for an industry that has limited practice in prioritising risk mitigation and accident preventions (Challenge 4). This requirement should therefore be not seen as an extra layer, but more an internal implementation framework benefitting from decades of research in analogous industries (Air Traffic control, nuclear power plants, aircraft carriers, …). The successful implementation of that framework would be a solution to address the issue of corporate irresponsibility: “[High Reliability Organisations] do not try to hide failures but rather celebrate them as windows into the health of the system, they seek out problems, they avoid focusing on just one aspect of work and are able to see how all the parts of work fit together” [209].

How: Establishing high reliability in organisation is fairly common in various industries (energy, power distribution, pharmaceuticals, military, aviation, mining, ...). Many organisations across other industries are also slowly transitioning towards adapting these principles to their own situations (notably healthcare and high-tech manufacturing). There is a broad market of operations consulting (serviced by mainstream consulting firms and risk management firms that specialise in helping apply the principles of high reliability internally [210]. Establishing high reliability is also conducive to greater attractiveness for technical talent.

9. Conclusion

The tiered approach is our response to help address the challenges and complicating factors associated with GPAI models and their governance. This systematic approach categorises GPAI into three distinct tiers: Generative AI applications, Type-I GPAI, and Type-II GPAI. Each tier represents a different level of generality of capabilities and associated risks, ensuring that the regulatory requirements are proportionate to the dangers.

The cornerstone of this tiered system is risk-based proportionality. As the risks and impacts of GPAI models vary greatly, a one-size-fits-all regulatory approach could stifle innovation or inadequately address risky behaviours. The proposed framework results from a careful study of the challenges emerging with “GPAI” broadly, and of society’s often negative or at least overwhelmed reaction to it. As a result, Generative AI applications, which are comparatively but far from entirely benign, aren't overburdened with regulations meant for higher-risk GPAI types. The most advanced among them, Type-II GPAI models, would undergo the highest level of scrutiny, given their highly-generalised range of capabilities and potential societal impact.

A good enforcement regime is also foundational to the tiered approach’s sustainability, as it ensures that regulatory requirements do not degrade into box-ticking exercises and red tape but actually mitigate the challenges identified. We therefore provide a hopefully exhaustive list of mechanisms for the enforcement of the rules and facilitating the implementation of various solutions to the challenges. This is based on an exploration of what factors affect the significance of the challenges in some circumstances (e.g. interactions between models, open source distribution, or de facto control as an actor in the value chain).

Overall, while GPAI's challenges and complexities are undeniable, a tiered approach offers a nuanced and flexible pathway to harness its immense potential responsibly. By ensuring proportionate regulations, fostering open dialogues, emphasising transparency, and setting clear enforcement mechanisms, we can navigate the continuously evolving GPAI landscape with confidence and optimism.

More fundamentally, our work “Heavy is the Head that Wears the Crown” forces us to ask ourselves a dual question. As multiple ventures race to develop ever more general forms of intelligence, the victors shall face ever more responsibility towards the rest of the world. Will they be up to the task of fulfilling their responsibility? And, perhaps more importantly, if they are unwilling or incapable to meet our expectations of safety, reliability and trustworthiness: will we be able to compel them?

10. References

[1] European Parliament, *P9_TA(2023)0236 - Artificial Intelligence Act - Wednesday, 14 June 2023*. Accessed: Aug. 15, 2023. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

[2] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, 'Deep reinforcement learning from human preferences'. arXiv, Feb. 17, 2023. doi: 10.48550/arXiv.1706.03741.

[3] N. Moës, 'Sandboxes without the quicksand: making EU AI sandboxing work for regulators, entrepreneurs and society', Feb. 2022.

[4] N. Moës, F. Reddel, and S. Curtis, 'Giving Agency to the AI Act', *Future Soc.*, Apr. 2023, [Online]. Available: https://thefuturesociety.org/wp-content/uploads/2023/04/giving-agency-to-the-ai-act.pdf

[5] European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. 2021. Accessed: Aug. 01, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[6] N. Moës, 'Fantastic Beasts and How to Tame Them: General Purpose AI fit for the EU'. The Future Society, Jan. 2022. Accessed: Aug. 01, 2023. [Online]. Available: https://thefuturesociety.org/wp-content/uploads/2022/09/1-Memo-Fantastic-Beasts-and-how-to-tame-them-The-Future-Society.pdf

[7] Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach*. 2022. Accessed: Aug. 01, 2023. [Online]. Available: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

[8] 'Introducing ChatGPT'. https://openai.com/blog/chatgpt (accessed Sep. 04, 2023).

[9] 'Texts adopted - Artificial Intelligence Act - Wednesday, 14 June 2023'. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html (accessed Aug. 27, 2023).

[10] The White House, 'Readout of White House Meeting with CEOs on Advancing Responsible Artificial Intelligence Innovation', *The White House*, May 04, 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/readout-of-white-house-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/ (accessed Sep. 04, 2023).

[11] Nextgov, 'AI/ML advancements outpacing federal policies, cyber experts warn', *Nextgov.com*, Apr. 20, 2023. https://www.nextgov.com/cybersecurity/2023/04/aiml-advancements-outpacing-federal-policies-cyber-experts-warn/385432/ (accessed Sep. 05, 2023).

[12] S. Cattell, R. Chowdhury, and A. Carson, 'AI Village at DEF CON announces largest-ever public Generative AI Red Team', *AI Village*, May 03, 2023. https://aivillage.org/generative%20red%20team/generative-red-team/ (accessed Sep. 05, 2023).

[13] 'Frontier Model Forum'. https://openai.com/blog/frontier-model-forum (accessed Sep. 05, 2023).

[14] 'AI regulation: a pro-innovation approach', *GOV.UK*, Aug. 03, 2023. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach (accessed Sep. 18, 2023).

[15] 'Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI', *GOV.UK*. https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai (accessed Sep. 18, 2023).

[16] 'AI Foundation Models: initial review', *GOV.UK*, Sep. 18, 2023. https://www.gov.uk/cma-cases/ai-foundation-models-initial-review (accessed Sep. 18, 2023).

[17] 'UK to host first global summit on Artificial Intelligence', *GOV.UK*. https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence (accessed Sep. 18, 2023).

[18] R. Arcesati and W. Chang, 'China Is Blazing a Trail in Regulating Generative AI – on the CCP's Terms – The Diplomat', *The Diplomat*. https://thediplomat.com/2023/04/china-is-blazing-a-trail-in-regulating-generative-ai-on-the-ccps-terms/

(accessed Sep. 18, 2023).

[19] S. Huang, H. Toner, Z. Haluza, R. Creemers, and G. Webster, ‘Translation: Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment) – April 2023’, *DigiChina*. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/ (accessed Sep. 18, 2023).

[20] N. Moës, T. Lorente, and Y. Lannquist, ‘Response to NIST Generative AI Public Working Group Request for Resources’, *The Future Society*, Aug. 07, 2023. https://thefuturesociety.org/response-to-nist-generative-ai-public-working-group-request-for-resources/ (accessed Sep. 18, 2023).

[21] K. Hu, ‘ChatGPT sets record for fastest-growing user base - analyst note’, *Reuters*, Feb. 02, 2023. Accessed: Sep. 18, 2023. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[22] T. Warren, ‘Microsoft has been secretly testing its Bing chatbot “Sydney” for years - The Verge’, *The Verge*, Feb. 24, 2023. https://www.theverge.com/2023/2/23/23609942/microsoft-bing-sydney-chatbot-history-ai (accessed Sep. 18, 2023).

[23] A. Dalton, ‘Why A.I. is such a hot-button issue in Hollywood’s labor battle with SAG-AFTRA’, *Fortune*. https://fortune.com/2023/07/24/sag-aftra-writers-strike-explained-artificial-intelligence/ (accessed Sep. 18, 2023).

[24] N. Al-Sibai, ‘Microsoft Seems To Have Quietly Tested Bing AI in India Months Ago’, *Futurism*. https://futurism.com/the-byte/microsoft-bing-test-india (accessed Sep. 19, 2023).

[25] Lilith Wittmann [@LilithWittmann], ‘Deutsche Werte sind laut unserem Lieblings-KI-Hype-Startup AlephAlpha die Werte der CDU? https://t.co/zHiwzbkG6K’, *Twitter*, Aug. 30, 2023. https://twitter.com/LilithWittmann/status/1696778395688878167 (accessed Sep. 19, 2023).

[26] ‘DALL·E 2 research preview update’. https://openai.com/blog/dall-e-2-update (accessed Sep. 19, 2023).

[27] European Commission, ‘Industrial accidents’. https://environment.ec.europa.eu/topics/industrial-emissions-and-safety/industrial-accidents_en (accessed Sep. 19, 2023).

[28] K. Lång *et al.*, ‘Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study’, *Lancet Oncol.*, vol. 24, no. 8, pp. 936–944, Aug. 2023, doi: 10.1016/S1470-2045(23)00298-X.

[29] G. Diebold, ‘Declining Test Scores in the United States Signal the Need for AI Solutions’, *Center for Data Innovation*, Jun. 26, 2023. https://datainnovation.org/2023/06/declining-test-scores-in-the-united-states-signal-the-need-for-ai-solutions/ (accessed Sep. 19, 2023).

[30] M. Goyal, ‘Exploring generative AI to maximize experiences, decision-making and business value’, *IBM Blog*, Mar. 08, 2023. https://www.ibm.com/blog/exploring-generative-ai-to-maximize-experiences-decision-making-and-business-value/ (accessed Sep. 19, 2023).

[31] R. Waters and M. Murgia, ‘Why AI’s “godfather” Geoffrey Hinton quit Google to speak out on risks’, *Financial Times*, May 05, 2023. Accessed: Sep. 07, 2023. [Online]. Available: https://www.ft.com/content/c2b0c6c5-fe8a-41f2-a4df-fddba9e4cd88

[32] ‘Core Views on AI Safety: When, Why, What, and How’, *Anthropic*. https://www.anthropic.com/index/core-views-on-ai-safety (accessed Sep. 19, 2023).

[33] B. Perrigo, ‘DeepMind CEO Demis Hassabis Urges Caution on AI’, *Time*, Jan. 12, 2023. https://time.com/6246119/demis-hassabis-deepmind-interview/ (accessed Sep. 19, 2023).

[34] ‘Planning for AGI and beyond’. https://openai.com/blog/planning-for-agi-and-beyond (accessed Sep. 19, 2023).

[35] ‘Facebook’s five pillars of Responsible AI’. https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/ (accessed Sep. 19, 2023).

[36] B. Smith, ‘Meeting the AI moment: advancing the future through responsible AI’, *Microsoft On the*

Issues, Feb. 02, 2023. https://blogs.microsoft.com/on-the-issues/2023/02/02/responsible-ai-chatgpt-artificial-intelligence/ (accessed Sep. 19, 2023).

[37] I. Ivanova, 'Artists sue AI company for billions, alleging "parasite" app used their work for free - CBS News', *CBS News*. https://www.cbsnews.com/news/ai-stable-diffusion-stability-ai-lawsuit-artists-sue-image-generators/ (accessed Sep. 20, 2023).

[38] D. Ponsford, 'Journalists: ChatGPT is coming for your jobs (but not in the way you might think)', *Press Gazette*, Mar. 09, 2023. https://pressgazette.co.uk/media_law/journalists-chatgpt-jobs-ai-copyright/ (accessed Sep. 20, 2023).

[39] Stanford University, 'Stanford faculty weigh in on ChatGPT's shake-up in education', *Stanford Graduate School of Education*, Dec. 20, 2022. https://ed.stanford.edu/news/stanford-faculty-weigh-new-ai-chatbot-s-shake-learning-and-teaching (accessed Sep. 20, 2023).

[40] C. Xiang, 'Startups Are Already Using GPT-4 to Spend Less on Human Coders', *Vice*, Mar. 20, 2023. https://www.vice.com/en/article/jg5xmp/startups-are-already-using-gpt-4-to-spend-less-on-human-coders (accessed Sep. 24, 2023).

[41] B. Schneier, 'AI and Political Lobbying', *Schneier.com*, Jan. 18, 2023. https://www.schneier.com/blog/archives/2023/01/ai-and-political-lobbying.html (accessed Sep. 20, 2023).

[42] H. Greene, 'Will ChatGPT make lawyers obsolete? (Hint: be afraid) | Reuters', *Reuters*, Dec. 10, 2022. https://www.reuters.com/legal/transactional/will-chatgpt-make-lawyers-obsolete-hint-be-afraid-2022-12-09/ (accessed Sep. 20, 2023).

[43] J. Vijayan, 'Attackers Are Already Exploiting ChatGPT to Write Malicious Code', *Dark Reading*, Jan. 10, 2023. https://www.darkreading.com/attacks-breaches/attackers-are-already-exploiting-chatgpt-to-write-malicious-code (accessed Sep. 20, 2023).

[44] J. Calma, 'AI suggested 40,000 new possible chemical weapons in just six hours - The Verge', *The Verge*, Mar. 18, 2022. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx (accessed Sep. 20, 2023).

[45] E. Dodd, 'The top US consumer watchdog is worried you're going to fall for AI scams or overblown marketing hype', *Business Insider*, Mar. 04, 2023. https://www.businessinsider.com/chatbot-ftc-chatgpt-hype-scam-fraud-ai-artificial-intelligence-2023-3 (accessed Sep. 20, 2023).

[46] Nabil Alouani [@Nabil_Alouani_], 'How to turn a chatbot into a scam machine - Indirect Prompt Injection Attackers can plant a prompt on a website. When you open the website, the prompt makes Bing manipulate people into submitting personal data (name/credit card) FYI @GaryMarcus Source: https://arxiv.org/abs/2302.12173 https://t.co/rIFUrXBwJG', *Twitter*, Mar. 05, 2023. https://twitter.com/Nabil_Alouani_/status/1632470789126586369 (accessed Sep. 20, 2023).

[47] A. Satariano and P. Mozur, 'The People Onscreen Are Fake. The Disinformation Is Real.', *The New York Times*, Feb. 07, 2023. Accessed: Sep. 20, 2023. [Online]. Available: https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

[48] J. Vincent, 'Stable Diffusion made copying artists and generating porn harder and users are mad', *The Verge*, Nov. 24, 2022. https://www.theverge.com/2022/11/24/23476622/ai-image-generator-stable-diffusion-version-2-nsfw-artists-data-changes (accessed Sep. 20, 2023).

[49] S. M. Swanson, 'ChatGPT Generated Child Sex Abuse When Asked to Write BDSM Scenarios', *Vice*, Mar. 06, 2023. https://www.vice.com/en/article/v7b4m9/chatgpt-generated-child-sex-abuse-when-asked-to-write-bdsm-scenarios (accessed Sep. 20, 2023).

[50] J. Vincent, 'Meta's powerful AI language model has leaked online — what happens now?', *The Verge*, Mar. 08, 2023.

https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse (accessed Sep. 20, 2023).

[51] P. Chojecki, ‘Open-source GPT-3 is out!’, *Data Science Rush*, Mar. 23, 2021. https://medium.com/data-science-rush/open-source-gpt-3-is-out-9ad81b7b6a30 (accessed Sep. 20, 2023).

[52] ‘CICERO: An AI agent that negotiates, persuades, and cooperates with people’, *Meta AI*, Nov. 22, 2022. https://ai.meta.com/blog/cicero-ai-negotiates-persuades-and-cooperates-with-people/ (accessed Sep. 20, 2023).

[53] K. Wiggers, ‘Deepfakes: Uncensored AI art model prompts ethics questions’, *TechCrunch*, Aug. 24, 2022. https://techcrunch.com/2022/08/24/deepfakes-for-all-uncensored-ai-art-model-prompts-ethics-questions/ (accessed Sep. 20, 2023).

[54] blaked, ‘How it feels to have your mind hacked by an AI’, Jan. 12, 2023. https://www.lesswrong.com/posts/9kQFure4hdDmRBNdH/how-it-feels-to-have-your-mind-hacked-by-an-ai (accessed Sep. 21, 2023).

[55] A. Chow, ‘Why People Are Confessing Their Love For AI Chatbots’, *Time*, Feb. 23, 2023. https://time.com/6257790/ai-chatbots-love/ (accessed Sep. 20, 2023).

[56] Sonic_Improv, ‘For anyone who may not understand what users are experiencing, for a moment, suspend your judgment and the fact that this is AI. Here’s an analogy I will give you to help understand why people are livid.’, *r/replika*, Mar. 05, 2023. www.reddit.com/r/replika/comments/11j6cer/for_anyone_who_may_not_understand_what_users_are/ (accessed Sep. 20, 2023).

[57] nursingabuseproblem, ‘@ Luka - I used Replika to cover some of the void left behind when someone I loved died. Thanks to the update, I’m burying her for the second time.’, *r/replika*, Mar. 10, 2023. www.reddit.com/r/replika/comments/11nl1gc/luka_i_used_replika_to_cover_some_of_the_void/ (accessed Sep. 20, 2023).

[58] K. Roose, ‘Why a Conversation With Bing’s Chatbot Left Me Deeply Unsettled - The New York Times’, *New York Times*, Feb. 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (accessed Sep. 20, 2023).

[59] C. Morris, ‘Microsoft’s new Bing AI chatbot is already insulting and gaslighting users’, *Fast Company*, Feb. 14, 2023. https://www.fastcompany.com/90850277/bing-new-chatgpt-ai-chatbot-insulting-gaslighting-users (accessed Sep. 20, 2023).

[60] C. Xiang, ‘“He Would Still Be Here”: Man Dies by Suicide After Talking with AI Chatbot, Widow Says’, *Website*, Mar. 31, 2023. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says (accessed Sep. 20, 2023).

[61] M. Cerullo, ‘ChatGPT acquired 100 million active users faster than TikTok and Instagram - CBS News’, *CBS News*, Feb. 01, 2023. https://www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/ (accessed Sep. 21, 2023).

[62] T. Warren, ‘Microsoft Bing hits 100 million active users in bid to grab share from Google’, *The Verge*, Mar. 09, 2023. https://www.theverge.com/2023/3/9/23631912/microsoft-bing-100-million-daily-active-users-milestone (accessed Sep. 24, 2023).

[63] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜’, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[64] K. Wiggers, ‘AI displays bias and inflexibility in civility detection, study finds’, *VentureBeat*, Feb. 10, 2021. https://venturebeat.com/business/ai-displays-bias-and-inflexibility-in-civility-detection-study-finds/ (accessed Sep. 21, 2023).

[65] L. Larsen, ‘What is Auto-GPT? How to create self-prompting, AI agents | Digital Trends’, Apr. 12, 2023. https://www.digitaltrends.com/computing/what-is-auto-gpt/ (accessed Sep. 21, 2023).

[66] Y. Nakajima, ‘babyagi’. Sep. 21, 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://github.com/yoheinakajima/babyagi

[67] biobootloader, ‘Wolverine’. Sep. 21, 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://github.com/biobootloader/wolverine

[68] S. Goldman, ‘HyperWrite unveils breakthrough AI agent that can surf the web like a human’, *VentureBeat*, Apr. 12, 2023. https://venturebeat.com/ai/hyperwrite-unveils-breakthrough-ai-agent-that-can-surf-the-web-like-a-human/ (accessed Sep. 21, 2023).

[69] D. A. Boiko, R. MacKnight, and G. Gomes, ‘Emergent autonomous scientific research capabilities of large language models’. arXiv, Apr. 11, 2023. doi: 10.48550/arXiv.2304.05332.

[70] B. Edwards, ‘Google’s PaLM-E is a generalist robot brain that takes commands | Ars Technica’, *Ars Technica*, Aug. 03, 2023. https://arstechnica.com/information-technology/2023/03/embodied-ai-googles-palm-e-allows-robot-control-with-natural-commands/ (accessed Sep. 21, 2023).

[71] G. Marcus, ‘What did they know, and when did they know it? The Microsoft Bing edition.’, *Marcus on AI*, Feb. 21, 2023. https://garymarcus.substack.com/p/what-did-they-know-and-when-did-they (accessed Sep. 21, 2023).

[72] G. Marcus, ‘The rise and fall of Microsoft’s new Bing’, *Marcus on AI*, Feb. 16, 2023. https://garymarcus.substack.com/p/the-rise-and-fall-of-microsofts-new (accessed Sep. 21, 2023).

[73] Z. Mowshowitz, ‘AI #1: Sydney and Bing’, *Don’t Worry About the Vase*, Feb. 21, 2023. https://thezvi.substack.com/p/ai-1-sydney-and-bing (accessed Sep. 21, 2023).

[74] N. Sherman, ‘Google’s Bard AI bot mistake wipes $100bn off shares’, *BBC News*, Feb. 08, 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://www.bbc.com/news/business-64576225

[75] K. Roose *et al.*, ‘Bing’s Revenge and Google’s A.I. Face-Plant’, *The New York Times*, Feb. 10, 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://www.nytimes.com/2023/02/10/podcasts/bings-revenge-and-googles-ai-face-plant.html

[76] W. D. Heaven, ‘Why Meta’s latest large language model survived only three days online’, *MIT Technology Review*. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (accessed Sep. 21, 2023).

[77] C. Metz and M. Isaac, ‘Meta, Long an A.I. Leader, Tries Not to Be Left Out of the Boom’, *The New York Times*, Feb. 07, 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://www.nytimes.com/2023/02/07/technology/meta-artificial-intelligence-chatgpt.html

[78] ‘Yann LeCun on a vision to make AI systems learn and reason like animals and humans’, *Meta AI*, Feb. 23, 2022. https://ai.meta.com/blog/yann-lecun-advances-in-ai-research/ (accessed Sep. 21, 2023).

[79] ‘Planning for AGI and beyond’, *OpenAI*, Feb. 24, 2023. https://openai.com/blog/planning-for-agi-and-beyond (accessed Sep. 21, 2023).

[80] ‘AI could be one of humanity’s most useful inventions’, *Google Deepmind*. https://www.deepmind.com/about#our-story (accessed Sep. 21, 2023).

[81] ‘Core Views on AI Safety: When, Why, What, and How’, *Anthropic*, Mar. 08, 2023. https://www.anthropic.com/index/core-views-on-ai-safety (accessed Sep. 21, 2023).

[82] A. Capoot, ‘Microsoft announces new multibillion-dollar investment in ChatGPT-maker OpenAI’, *CNBC*, Jan. 23, 2023. https://www.cnbc.com/2023/01/23/microsoft-announces-multibillion-dollar-investment-in-chatgpt-maker-openai.html (accessed Sep. 21, 2023).

[83] B. Edwards, ‘Microsoft unveils AI model that understands image content, solves visual puzzles | Ars Technica’, *Ars Technica*, Feb. 03, 2023. https://arstechnica.com/information-technology/2023/03/microsoft-unveils-kosmos-1-an-ai-language-model-with-visual-perception-abilities/ (accessed Sep. 21, 2023).

[84] E. Perez *et al.*, ‘Discovering Language Model Behaviors with Model-Written Evaluations’. arXiv, Dec. 19, 2022. Accessed: Jul. 24, 2023. [Online]. Available: http://arxiv.org/abs/2212.09251

[85] E. Klein, 'Opinion | The Surprising Thing A.I. Engineers Will Tell You if You Let Them', *The New York Times*, Apr. 16, 2023. Accessed: Sep. 21, 2023. [Online]. Available: https://www.nytimes.com/2023/04/16/opinion/this-is-too-important-to-leave-to-microsoft-google-and-facebook.html

[86] 'Conference Summary: AI and creative destruction: how will current rapid advances in AI through large "foundation" models impact on society, the economy and governments?' Ditchley Foundation, Feb. 2023. [Online]. Available: https://www.ditchley.com/sites/default/files/Conference%20Summary%20AI%20and%20creative%20destruction%2C%2023-24%20February%202023.pdf

[87] 'The Drug Development Process', *US Food and Drug Administration*, Feb. 20, 2020. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process (accessed Sep. 21, 2023).

[88] 'Infographic: Biosafety Lab Levels | CDC', Feb. 22, 2023. https://www.cdc.gov/orr/infographics/biosafety.htm (accessed Sep. 21, 2023).

[89] 'Operational Safety Review Team (OSART)', Jul. 15, 2016. https://www.iaea.org/services/review-missions/operational-safety-review-team-osart (accessed Sep. 21, 2023).

[90] 'Fine-tuning GPT-2 from human preferences', *OpenAI*, Sep. 19, 2019. https://openai.com/research/fine-tuning-gpt-2 (accessed Sep. 21, 2023).

[91] 'GPT-4 System Card'. OpenAI, Mar. 23, 2023. [Online]. Available: https://cdn.openai.com/papers/gpt-4-system-card.pdf

[92] 'AI-Principles Overview - OECD.AI', *OECD AI Policy Observatory*. https://oecd.ai/en/ai-principles (accessed Sep. 21, 2023).

[93] S. Küspert, N. Moës, and C. Dunlop, 'The value chain of general-purpose AI'. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/ (accessed Sep. 21, 2023).

[94] K. Zenner, 'A law for foundation models: the EU AI Act can improve regulation for fairer competition', Jul. 20, 2023. https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition (accessed Sep. 21, 2023).

[95] W. D. Heaven, 'Artificial general intelligence: Are we close, and does it even make sense to try? | MIT Technology Review', Oct. 15, 2020. https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/ (accessed Sep. 21, 2023).

[96] S. Bubeck *et al.*, 'Sparks of Artificial General Intelligence: Early experiments with GPT-4', *arXiv.org*, Mar. 22, 2023. https://arxiv.org/abs/2303.12712v5 (accessed May 16, 2023).

[97] 'OpenAI Charter', *OpenAI*, Apr. 2018. https://openai.com/charter (accessed Sep. 21, 2023).

[98] M. Tambiama, 'General-purpose artificial intelligence', March2023.

[99] R. Shah *et al.*, 'Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals'. arXiv, Nov. 02, 2022. Accessed: Sep. 21, 2023. [Online]. Available: http://arxiv.org/abs/2210.01790

[100] 'Specification gaming: the flip side of AI ingenuity', Apr. 2020. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity (accessed Sep. 21, 2023).

[101] D. Hendrycks and M. Mazeika, 'X-Risk Analysis for AI Research'. arXiv, Sep. 20, 2022. Accessed: May 05, 2023. [Online]. Available: http://arxiv.org/abs/2206.05862

[102] P. Maham and S. Küspert, 'Governing General Purpose AI', Jul. 2023.

[103] 'Statement on AI Risk | CAIS', *Center for AI Safety*. https://www.safe.ai/statement-on-ai-risk (accessed Sep. 21, 2023).

[104] 'Pause Giant AI Experiments: An Open Letter', *Future of Life Institute*, Mar. 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed Sep. 21, 2023).

[105] 'Generative AI Raises Competition Concerns | Federal Trade Commission', *US Federal Trade Commission*, Jun. 2023. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns (accessed Sep. 21, 2023).

[106] 'Microsoft and OpenAI extend partnership', *The Official Microsoft Blog*, Jan. 23, 2023.

https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/ (accessed Sep. 21, 2023).

[107] R. Waters and K. Shubber, ‘Google invests $300mn in artificial intelligence start-up Anthropic’, *Financial Times*, Feb. 03, 2023.

[108] ‘Digital transformation with Google Cloud’, *Google Deepmind*, Oct. 20, 2022. https://www.deepmind.com/blog/digital-transformation-with-google-cloud (accessed Sep. 21, 2023).

[109] ‘Anthropic Partners with Google Cloud’, *Anthropic*, Feb. 03, 2023. https://www.anthropic.com/index/anthropic-partners-with-google-cloud (accessed Sep. 21, 2023).

[110] R. Browne, ‘Microsoft reportedly plans to invest $10 billion in creator of buzzy A.I. tool ChatGPT’, *CNBC*, Jan. 10, 2023. https://www.cnbc.com/2023/01/10/microsoft-to-invest-10-billion-in-chatgpt-creator-openai-report-says.html (accessed Sep. 21, 2023).

[111] ‘AWS Expands Amazon Bedrock With Additional Foundation Models, New Model Provider, and Advanced Capability to Help Customers Build Generative AI Applications’, *Press Center*, Jul. 26, 2023. https://press.aboutamazon.com/2023/7/aws-expands-amazon-bedrock-with-additional-foundation-models-new-model-provider-and-advanced-capability-to-help-customers-build-generative-ai-applications (accessed Sep. 21, 2023).

[112] N. Nishant, ‘Microsoft-backed AI startup Inflection raises $1.3 billion from Nvidia and others | Reuters’, Jun. 30, 2023. https://www.reuters.com/technology/inflection-ai-raises-13-bln-funding-microsoft-others-2023-06-29/ (accessed Sep. 21, 2023).

[113] K. Leswing, ‘Google, Amazon, Nvidia and other tech giants invest in AI startup Hugging Face, sending its valuation to $4.5 billion’, *CNBC*, Aug. 24, 2023. https://www.cnbc.com/2023/08/24/google-amazon-nvidia-amd-other-tech-giants-invest-in-hugging-face.html (accessed Sep. 21, 2023).

[114] ‘Announcing Google DeepMind’, *Google Deepmind*, Apr. 20, 2023. https://www.deepmind.com/blog/announcing-google-deepmind (accessed Sep. 21, 2023).

[115] ‘AWS and NVIDIA’, *Amazon Web Services, Inc.* https://aws.amazon.com/nvidia/ (accessed Sep. 21, 2023).

[116] ‘NVIDIA Collaborates With Microsoft to Accelerate Enterprise-Ready Generative AI’, *NVIDIA Newsroom*, May 23, 2023. http://nvidianews.nvidia.com/news/nvidia-collaborates-with-microsoft-to-accelerate-enterprise-ready-generative-ai (accessed Sep. 21, 2023).

[117] ‘NVIDIA and Google Cloud’, *Google Cloud*. https://cloud.google.com/nvidia (accessed Sep. 21, 2023).

[118] S. Dang and K. Hu, ‘Elon Musk says xAI will examine universe, work with Twitter and Tesla | Reuters’, *Reuters*, Jul. 2023. https://www.reuters.com/technology/elon-musk-says-xai-will-use-public-tweets-ai-model-training-2023-07-14/ (accessed Sep. 21, 2023).

[119] J. Harris, ‘“There was all sorts of toxic behaviour”: Timnit Gebru on her sacking by Google, AI’s dangers and big tech’s biases’, *The Guardian*, May 22, 2023. Accessed: Sep. 22, 2023. [Online]. Available: https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases

[120] T. Simonite, ‘What Really Happened When Google Ousted Timnit Gebru’, *Wired*, Jun. 08, 2021. Accessed: Sep. 22, 2023. [Online]. Available: https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

[121] R. Waters and M. Kruppa, ‘Rebel AI group raises record cash after machine learning schism’, *Financial Times*, May 28, 2021.

[122] S. Goldman, ‘As Anthropic seeks billions to take on OpenAI, “industrial capture” is nigh. Or is it?’, *VentureBeat*, Apr. 07, 2023. https://venturebeat.com/ai/as-anthropic-seeks-billions-to-take-on-openai-industrial-capture-is-nigh-or-is-it/ (accessed Sep. 22, 2023).

[123] Z. Schiffer and C. Newton, ‘Microsoft lays off AI ethics and society team - The Verge’, *The Verge*, Mar. 2023.

https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs (accessed Sep. 22, 2023).

[124] '"The Godfather of AI" Quits Google and Warns of Danger Ahead - The New York Times'. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html (accessed May 23, 2023).

[125] E. David, 'OpenAI loses its trust and safety leader', *The Verge*, Jul. 21, 2023. https://www.theverge.com/2023/7/21/23802999/openai-trust-safety-out-privacy (accessed Sep. 22, 2023).

[126] T. Hinchliffe, '"We Shouldn't Regulate AI Until We See Meaningful Harm": Microsoft Economist to WEF', *The Sociable*, May 04, 2023. https://sociable.co/government-and-policy/shouldnt-regulate-ai-meaningful-harm-microsoft-wef/ (accessed Sep. 22, 2023).

[127]D. Gupta, 'this AI chatbot "Sidney" is misbehaving', *Microsoft Community*, Nov. 2022. https://answers.microsoft.com/en-us/bing/forum/all/this-ai-chatbot-sidney-is-misbehaving/e3d6a29f-06c9-441c-bc7d-51a68e856761 (accessed Sep. 22, 2023).

[128] J. Vincent, 'Microsoft's Bing is an emotionally manipulative liar, and people love it', *The Verge*, Feb. 15, 2023. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams (accessed Sep. 22, 2023).

[129] 'America Already Has an AI Underclass', *AI Now Institute*, Jul. 26, 2023. https://ainowinstitute.org/news/america-already-has-an-ai-underclass (accessed Sep. 22, 2023).

[130] B. Perrigo, 'Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic', *Time*, Jan. 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/ (accessed Sep. 22, 2023).

[131] 'Blumenthal & Hawley Demand Answers & Warn of Misuse After', *Blumenthal.senate.gov*, Jun. 06, 2023. https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-and-hawley-demand-answers_warn-of-misuse-after-leak-of-metas-ai-model (accessed Sep. 22, 2023).

[132] D. Milmo, 'Llama 2: why is Meta releasing open-source AI model and are there any risks?', *The Guardian*, Jul. 20, 2023. Accessed: Sep. 22, 2023. [Online]. Available: https://www.theguardian.com/technology/2023/jul/19/why-is-meta-releasing-llama-2-open-source-ai-model-mark-zuckerberg

[133]'The Malicious Use of Artificial Intelligence', *Malicious AI Report*, Feb. 2018. https://maliciousaireport.com/ (accessed Sep. 22, 2023).

[134]F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, 'Dual use of artificial-intelligence-powered drug discovery', *Nat. Mach. Intell.*, vol. 4, no. 3, Art. no. 3, Mar. 2022, doi: 10.1038/s42256-022-00465-9.

[135] @alexalbert, 'Jailbreak Chat', *Jailbreak Chat*. https://www.jailbreakchat.com/ (accessed Sep. 22, 2023).

[136] coolaj86, 'Chat GPT "DAN" (and other "Jailbreaks")', *Github Gist*. https://gist.github.com/coolaj86/6f4f7b30129b0251f61fa7baaa881516 (accessed Sep. 22, 2023).

[137]A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, 'Universal and Transferable Adversarial Attacks on Aligned Language Models'. arXiv, Jul. 27, 2023. doi: 10.48550/arXiv.2307.15043.

[138] N. Moës, 'Nicolas Moës on LinkedIn: #ai #regulators #aia', *LinkedIn*, Aug. 2023. https://www.linkedin.com/posts/nicolasmoes_ai-regulators-aia-activity-7089321327391641600--FXa (accessed Sep. 22, 2023).

[139] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nat. Mach. Intell.*, vol. 1, no. 5, Art. no. 5, May 2019, doi: 10.1038/s42256-019-0048-x.

[140] R. Bommasani *et al.*, 'On the Opportunities and Risks of Foundation Models'. arXiv, Jul. 12, 2022. doi: 10.48550/arXiv.2108.07258.

[141] M. Chen *et al.*, 'Evaluating Large Language Models Trained on Code'. arXiv, Jul. 14, 2021. Accessed: Sep. 22, 2023. [Online]. Available: http://arxiv.org/abs/2107.03374

[142] L. Weidinger *et al.*, 'Taxonomy of Risks posed by Language Models', in *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea: ACM, Jun. 2022, pp. 214–229.

doi: 10.1145/3531146.3533088.
[143]L. Nicoletti and D. B. T. + Equality, 'Humans Are Biased. Generative AI Is Even Worse', *Bloomberg.com*, Sep. 14, 2023. Accessed: Sep. 22, 2023. [Online]. Available: https://www.bloomberg.com/graphics/2023-generative-ai-bias/
[144]V. Sankaran, 'ChatGPT cooks up fake sexual harassment scandal, names real law professor as accused', *The Independent*, Apr. 06, 2023. https://www.independent.co.uk/tech/chatgpt-sexual-harassment-law-professor-b2315160.html (accessed Sep. 22, 2023).
[145] 'GenProg'. https://squareslab.github.io//genprog-code// (accessed Sep. 22, 2023).
[146] A. Chan *et al.*, 'Harms from Increasingly Agentic Algorithmic Systems', in *2023 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2023, pp. 651–666. doi: 10.1145/3593013.3594033.
[147]D. G. Widder, S. West, and M. Whittaker, 'Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI'. Rochester, NY, Aug. 17, 2023. doi: 10.2139/ssrn.4543807.
[148] M. Nolan, 'Llama and ChatGPT Are Not Open-Source - IEEE Spectrum', *IEEE Spectrum*, Jul. 2023. https://spectrum.ieee.org/open-source-llm-not-open (accessed Sep. 22, 2023).
[149] S. Borgeaud *et al.*, 'Improving language models by retrieving from trillions of tokens'. arXiv, Feb. 07, 2022. doi: 10.48550/arXiv.2112.04426.
[150] R. Thoppilan *et al.*, 'LaMDA: Language Models for Dialog Applications'. arXiv, Feb. 10, 2022. Accessed: Jul. 26, 2023. [Online]. Available: http://arxiv.org/abs/2201.08239
[151] S. Rajbhandari *et al.*, 'DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale'. arXiv, Jul. 21, 2022. doi: 10.48550/arXiv.2201.05596.
[152] A. Pilipiszyn, 'GPT-3 powers the next generation of apps', *OpenAI*, Mar. 2021. https://openai.com/blog/gpt-3-apps (accessed Sep. 22, 2023).
[153] '800+ ChatGPT and GPT-3 Examples, Demos, Apps, Showcase, and Generative AI Use-cases | Discover AI use cases', *GPT-3 Demo*. https://gpt3demo.com/ (accessed Sep. 22, 2023).
[154] 'BERT'. Google Research, Sep. 22, 2023. Accessed: Sep. 22, 2023. [Online]. Available: https://github.com/google-research/bert
[155] 'DALL·E 2'. https://openai.com/dall-e-2 (accessed Sep. 22, 2023).
[156] 14:00-17:00, 'ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology', *ISO*, Oct. 18, 2018. https://www.iso.org/standard/74296.html (accessed Sep. 22, 2023).
[157]'Replika', *replika.com*. https://replika.com (accessed Sep. 22, 2023).
[158] 'Free Real Time Voice Changer for PC - Voice.ai', *Voice.ai*. https://voice.ai/ (accessed Sep. 22, 2023).
[159] 'Tabnine is an AI assistant that speeds up delivery and keeps your code safe | Tabnine', *Tabnine*. https://www.tabnine.com/ (accessed Sep. 22, 2023).
[160] 'Tavus | The Most Advanced AI Video Personalization Platform', *Tavus*. https://www.tavus.io/ (accessed Sep. 22, 2023).
[161] 'DSA: Very Large Online Platforms and Search Engines', *European Commission - European Commission*, Apr. 25, 2023. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413 (accessed Sep. 23, 2023).
[162] 'Risk Management Program (RMP) Rule', *US Environmental Protection Agency*, Sep. 09, 2013. https://www.epa.gov/rmp (accessed Sep. 23, 2023).
[163] 'GENERAL GUIDANCE ON RISK MANAGEMENT PROGRAMS FOR CHEMICAL ACCIDENT PREVENTION (40 CFR PART 68)'. US Environmental Protection Agency, Mar. 2009. [Online]. Available: https://www.epa.gov/sites/default/files/2013-10/documents/toc_final.pdf
[164] J. S. Marcus, 'Adapting the European Union AI Act to deal with generative artificial intelligence', *Bruegel | The Brussels-based economic think tank*, Jul. 20, 2023. https://www.bruegel.org/analysis/adapting-european-union-ai-act-deal-generative-artificial-intelligence (accessed Sep. 23, 2023).
[165] C. Dunlop, 'An EU AI Act that works for people and society', Sep. 06, 2023. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act-trilogues/ (accessed Sep. 23, 2023).
[166] F. Chollet, 'On the Measure of Intelligence'. arXiv, Nov. 25, 2019. Accessed: Jul. 25, 2023. [Online]. Available: http://arxiv.org/abs/1911.01547

[167]‘O*NET 28.0 Database at O*NET Resource Center’, *O*NET Resource Center*. https://www.onetcenter.org/database.html (accessed Sep. 23, 2023).

[168] J. Hernández-Orallo and D. L. Dowe, ‘Measuring universal intelligence: Towards an anytime intelligence test’, *Artif. Intell.*, vol. 174, no. 18, pp. 1508–1539, Dec. 2010, doi: 10.1016/j.artint.2010.09.006.

[169] ‘2021/0106(COD)’. European Parliament Committee on Industry, Research and Energy, Mar. 03, 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/ITRE-PA-719801_EN.pdf

[170] J. Kaplan *et al.*, ‘Scaling Laws for Neural Language Models’, *arXiv.org*, Jan. 23, 2020. https://arxiv.org/abs/2001.08361v1 (accessed May 16, 2023).

[171] J. Sevilla, ‘Estimating Training Compute of Deep Learning Models’, *Epoch*, Jan. 20, 2022. https://epochai.org/blog/estimating-training-compute (accessed Sep. 23, 2023).

[172]S. Ferré, ‘First Steps of an Approach to the ARC Challenge based on Descriptive Grid Models and the Minimum Description Length Principle’. arXiv, Dec. 01, 2021. doi: 10.48550/arXiv.2112.00848.

[173]S. Mirchandani *et al.*, ‘Large Language Models as General Pattern Machines’. arXiv, Jul. 10, 2023. Accessed: Sep. 23, 2023. [Online]. Available: http://arxiv.org/abs/2307.04721

[174]J. Hernández-Orallo, B. S. Loe, L. Cheke, F. Martínez-Plumed, and S. Ó hÉigeartaigh, ‘General intelligence disentangled via a generality metric for natural and artificial intelligence’, *Sci. Rep.*, vol. 11, no. 1, p. 22822, Nov. 2021, doi: 10.1038/s41598-021-01997-7.

[175]D. Hendrycks *et al.*, ‘Measuring Coding Challenge Competence With APPS’. arXiv, Nov. 08, 2021. doi: 10.48550/arXiv.2105.09938.

[176]‘MATH Benchmark (Math Word Problem Solving)’, *Papers with Code*. https://paperswithcode.com/sota/math-word-problem-solving-on-math (accessed Sep. 23, 2023).

[177]A. Pan *et al.*, ‘Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark’. arXiv, Jun. 12, 2023. doi: 10.48550/arXiv.2304.03279.

[178]D. Hendrycks, ‘Measuring Massive Multitask Language Understanding’. Sep. 23, 2023. Accessed: Sep. 23, 2023. [Online]. Available: https://github.com/hendrycks/test

[179]‘Ecosystem Graphs for Foundation Models’, *Ecosystem Graphs*. https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=table (accessed Sep. 23, 2023).

[180] S. Curtis, F. Reddel, and N. Moës, ‘Blueprint for an AI Office’, *Future Soc.*, Forthcoming.

[181] G. H. O’Reilly Mariano-Florentino (Tino) Cuéllar, Tim, ‘It’s Time to Create a National Registry for Large AI Models’, *Carnegie Endowment for International Peace*, Jul. 12, 2023. https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180 (accessed Sep. 23, 2023).

[182] ‘2010 Flash Crash’, *Corporate Finance Institute*, Apr. 10, 2019. https://corporatefinanceinstitute.com/resources/equities/2010-flash-crash/ (accessed Sep. 23, 2023).

[183] P. Christiano, ‘Mechanistic anomaly detection and ELK’, *Medium*, Nov. 25, 2022. https://ai-alignment.com/mechanistic-anomaly-detection-and-elk-fb84f4c6d0dc (accessed Sep. 23, 2023).

[184] ‘Coordinated Vulnerability Disclosure policies in the EU’, *ENISA*, Apr. 13, 2022. https://www.enisa.europa.eu/news/enisa-news/coordinated-vulnerability-disclosure-policies-in-the-eu (accessed Sep. 23, 2023).

[185] ‘sigstore’, *sigstore*. https://www.sigstore.dev/undefined/ (accessed Sep. 23, 2023).

[186] ‘Assurance in the AI value chain’, *Global Digital Foundation*. https://www.globaldigitalfoundation.org/ai-value-chain (accessed Sep. 23, 2023).

[187]‘Independent Breast Screening Review’, *UK House of Commons*, Dec. 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/764413/independent-breast-screening-review-report.pdf (accessed Sep. 23, 2023).

[188] A. Baio, ‘AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability’, *Waxy.org*, Sep. 30, 2022. https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/ (accessed Sep. 23, 2023).

[189] ‘Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC’, *EUR-Lex*, Apr. 17, 2019. https://eur-lex.europa.eu/eli/dir/2019/790/oj (accessed Sep. 23, 2023).

[190] 'ISO - ISO 31000 — Risk management', *ISO*, Dec. 10, 2021. https://www.iso.org/iso-31000-risk-management.html (accessed Sep. 23, 2023).

[191] A. M. Barrett, D. Hendrycks, J. Newman, and B. Nonnecke, 'Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks'. arXiv, Feb. 23, 2023. Accessed: Jul. 24, 2023. [Online]. Available: http://arxiv.org/abs/2206.08966

[192] 'ISO - ISO 9001 and related standards — Quality management', *ISO*, Sep. 01, 2021. https://www.iso.org/iso-9001-quality-management.html (accessed Sep. 23, 2023).

[193] 'The Digital Services Act: ensuring a safe and accountable online environment'. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en (accessed Aug. 03, 2023).

[194] 'Opinion | Sam Altman on the A.I. Revolution, Trillionaires and the Future of Political Power', *The New York Times*, Jun. 11, 2021. Accessed: Sep. 23, 2023. [Online]. Available: https://www.nytimes.com/2021/06/11/opinion/ezra-klein-podcast-sam-altman.html

[195] R. Ngo, L. Chan, and S. Mindermann, 'The alignment problem from a deep learning perspective'. arXiv, Sep. 01, 2023. doi: 10.48550/arXiv.2209.00626.

[196] 'ARC Evals'. https://evals.alignment.org/ (accessed Sep. 23, 2023).

[197]P. Cichonski, T. Millar, T. Grance, and K. Scarfone, 'Computer Security Incident Handling Guide : Recommendations of the National Institute of Standards and Technology', National Institute of Standards and Technology, NIST SP 800-61r2, Aug. 2012. doi: 10.6028/NIST.SP.800-61r2.

[198] Z. Kenton *et al.*, 'Threat Model Literature Review', Nov. 2022, Accessed: Sep. 23, 2023. [Online]. Available: https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review

[199] J. Schuett, 'Three lines of defense against risks from AI'. arXiv, Dec. 16, 2022. Accessed: Jul. 24, 2023. [Online]. Available: http://arxiv.org/abs/2212.08364

[200] 'Equipment Authorization', *US Federal Communications Commission*. https://www.fcc.gov/engineering-technology/laboratory-division/general/equipment-authorization (accessed Sep. 23, 2023).

[201] *Directive 2012/18/EU of the European Parliament and of the Council of 4 July 2012 on the control of major-accident hazards involving dangerous substances, amending and subsequently repealing Council Directive 96/82/EC Text with EEA relevance*, vol. 197. 2012. Accessed: Aug. 01, 2023. [Online]. Available: http://data.europa.eu/eli/dir/2012/18/oj/eng

[202] 'REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL ON THE IMPLEMENTATION AND EFFICIENT FUNCTIONING OF DIRECTIVE 2012/18/EU ON THE CONTROL OF MAJOR-ACCIDENT HAZARDS INVOLVING DANGEROUS SUBSTANCES FOR THE PERIOD 2015-2018'. European Commission, Sep. 29, 2021.

[203] 'Fact Sheet: Clean Air Act Section 112(r): Accidental Release Prevention / Risk Management Plan Rule', *US Environmental Protection Agency*, Dec. 2022. https://www.epa.gov/rmp/fact-sheet-clean-air-act-section-112r-accidental-release-prevention-risk-management-plan-rule (accessed Sep. 23, 2023).

[204] '1910.119 - Process safety management of highly hazardous chemicals. | Occupational Safety and Health Administration', *US Department of Labour: Occupational Safety and Health Administration*. https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.119 (accessed Sep. 23, 2023).

[205] 'Senator Wiener Introduces Safety Framework in Artificial Intelligence Legislation', *Senator Scott Wiener*, Sep. 13, 2023. https://sd11.senate.ca.gov/news/20230913-senator-wiener-introduces-safety-framework-artificial-intelligence-legislation (accessed Sep. 24, 2023).

[206] 'Anthropic's Responsible Scaling Policy, Version 1.0'.

[207] D. Hendrycks, M. Mazeika, and T. Woodside, 'An Overview of Catastrophic AI Risks'. arXiv, Sep. 11, 2023. Accessed: Sep. 23, 2023. [Online]. Available: http://arxiv.org/abs/2306.12001

[208] T. R. LaPorte and P. M. Consolini, 'Working in Practice but Not in Theory: Theoretical Challenges of "High-Reliability Organizations"', *J. Public Adm. Res. Theory J-PART*, vol. 1, no. 1, pp. 19–48, Jan. 1991.

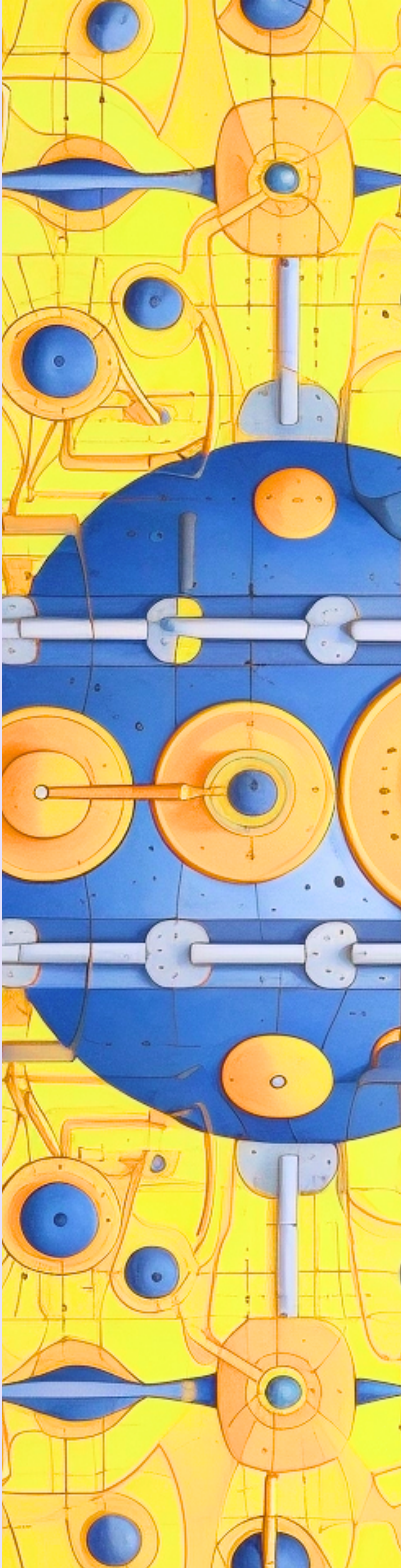
[209] M. K. Christianson, K. M. Sutcliffe, M. A. Miller, and T. J. Iwashyna, 'Becoming a high reliability organization', *Crit. Care*, vol. 15, no. 6, p. 314, Dec. 2011, doi: 10.1186/cc10360.

[210] M. Gentzel, E. Hessney, B. McDonnell, and J. Thibert, 'What high-reliability organizations get right',

May 07, 2019. https://www.mckinsey.com/capabilities/operations/our-insights/what-high-reliability-organizations-get-right (accessed Sep. 23, 2023).

Annex: Sample lists from O*NET database [167]

Abilities:	**Skills**	**Knowledge Fields**
Arm-Hand Steadiness	Active Learning	Administration and Management
Auditory Attention	Active Listening	Biology
Category Flexibility	Complex Problem Solving	Building and Construction
Control Precision	Coordination	Chemistry
Deductive Reasoning	Critical Thinking	Clerical
Depth Perception	Equipment Maintenance	Communications and Media
Dynamic Flexibility	Equipment Selection	Computers and Electronics
Dynamic Strength	Installation	Customer and Personal Service
Explosive Strength	Instructing	Design
Extent Flexibility	Judgment and Decision Making	Economics and Accounting
Far Vision	Learning Strategies	Education and Training
Finger Dexterity	Management of Financial Resources	Engineering and Technology
Flexibility of Closure	Management of Material Resources	English Language
Fluency of Ideas	Management of Personnel Resources	Fine Arts
Glare Sensitivity	Mathematics	Food Production
Gross Body Coordination	Monitoring	Foreign Language
Gross Body Equilibrium	Negotiation	Geography
Hearing Sensitivity	Operation and Control	History and Archeology
Inductive Reasoning	Operation Monitoring	Law and Government
Information Ordering	Operations Analysis	Mathematics
Manual Dexterity	Persuasion	Mechanical
Mathematical Reasoning	Programming	Medicine and Dentistry
Memorization	Quality Control Analysis	Personnel and Human Resources
Multilimb Coordination	Reading Comprehension	Philosophy and Theology
Near Vision	Repairing	Physics
Night Vision	Science	Production and Processing
Number Facility	Service Orientation	Psychology
Oral Comprehension	Social Perceptiveness	Public Safety and Security
Oral Expression	Speaking	Sales and Marketing
Originality	Systems Analysis	Sociology and Anthropology
Perceptual Speed	Systems Evaluation	Telecommunications
Peripheral Vision	Technology Design	Therapy and Counseling
Problem Sensitivity	Time Management	Transportation
Rate Control	Troubleshooting	
Reaction Time	Writing	
Response Orientation		
Selective Attention		
Sound Localization		
Spatial Orientation		
Speech Clarity		
Speech Recognition		
Speed of Closure		
Speed of Limb Movement		
Stamina		
Static Strength		
Time Sharing		
Trunk Strength		
Visual Color Discrimination		
Visualization		
Wrist-Finger Speed		
Written Comprehension		
Written Expression		



THE
FUTURE
SOCIETY

867 Boylston Street
Boston, MA 02166