THE FUTURE SOCIETY

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

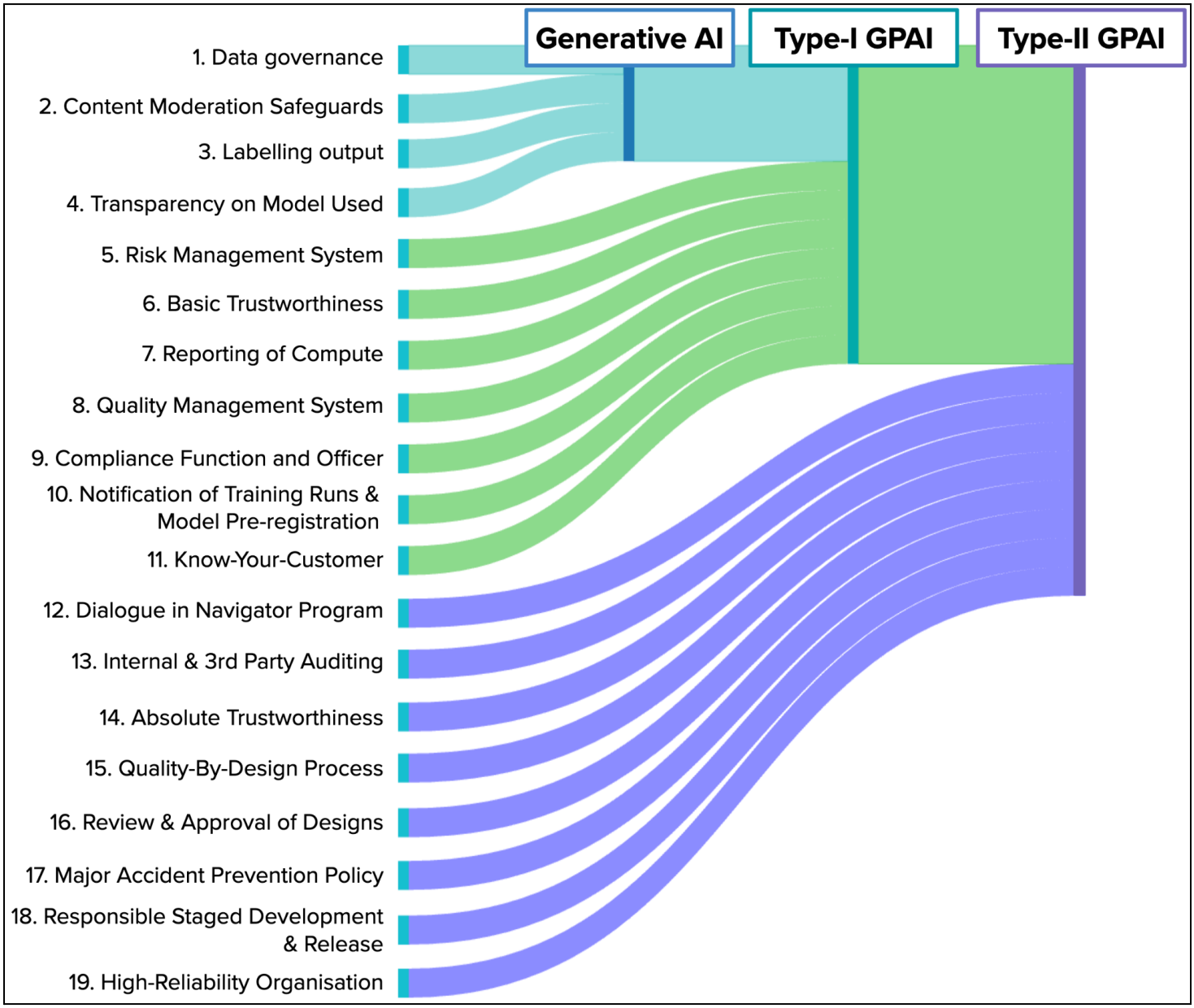
This report provides a **Blueprint for establishing a risk-based tiered system for the governance of General Purpose AI (GPAI) models.** Drawing on 2 years of theoretical, advisory and field research on the governance of these models in the AI Act, this guide distils our findings into actionable steps 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



[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.