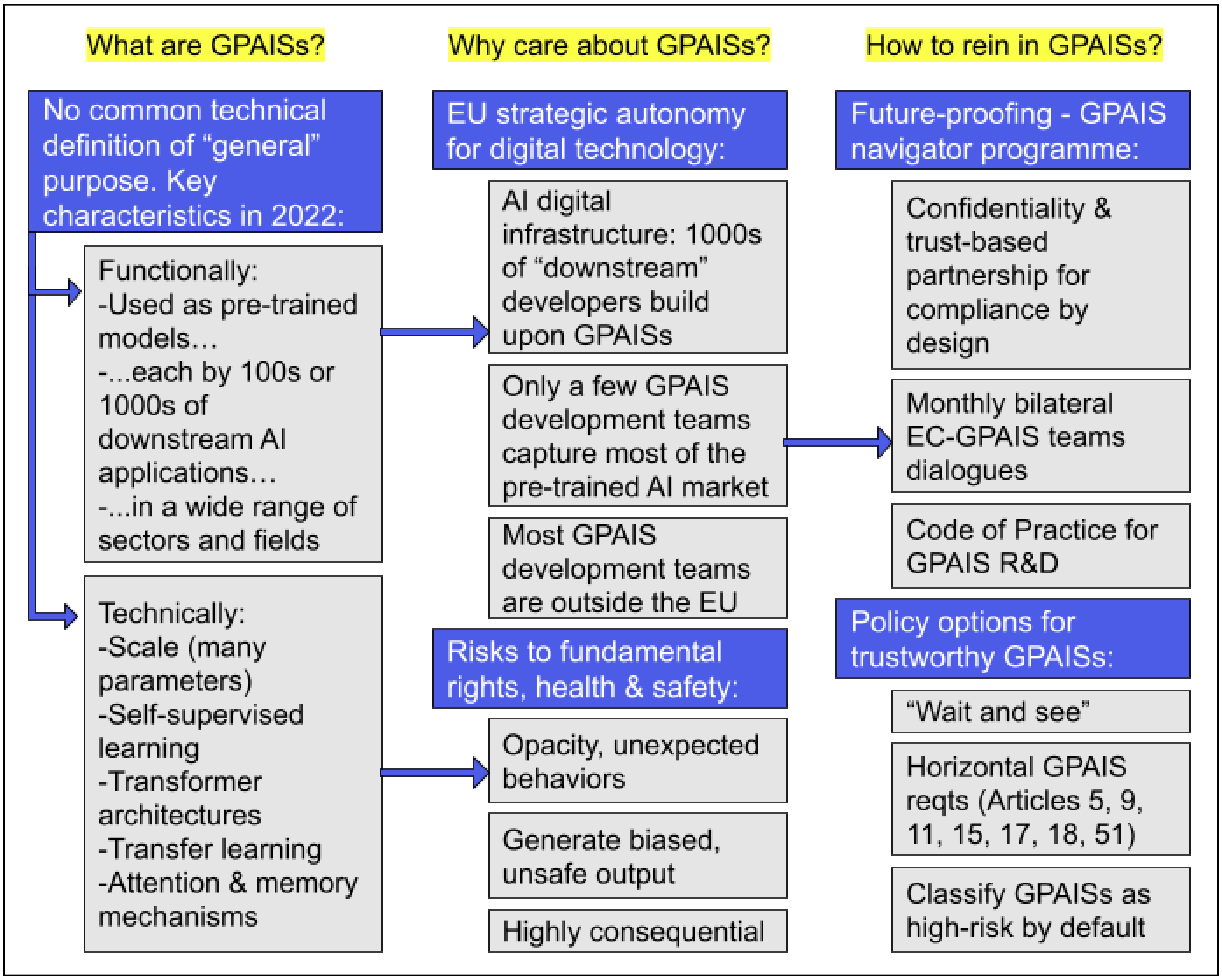
THE
FUTURE
SOCIETY

Fantastic Beasts and How to Tame Them: General Purpose AI fit for the EU

Memo by The Future Society,[1] January 2022

Recent discussions of the EU AI Act have introduced the concept of General Purpose AI Systems (henceforth GPAISs).[2] In this memo, we briefly explain what GPAISs are, why we should care about them, and our recommendations to rein them in for the EU.



What are GPAISs?

No common technical definition of "general" purpose. Key characteristics in 2022:

Functionally:
-Used as pre-trained models…
-...each by 100s or 1000s of downstream AI applications…
-...in a wide range of sectors and fields

Technically:
-Scale (many parameters)
-Self-supervised learning
-Transformer architectures
-Transfer learning
-Attention & memory mechanisms

Why care about GPAISs?

EU strategic autonomy for digital technology:

AI digital infrastructure: 1000s of "downstream" developers build upon GPAISs

Only a few GPAIS development teams capture most of the pre-trained AI market

Most GPAIS development teams are outside the EU

Risks to fundamental rights, health & safety:

Opacity, unexpected behaviors

Generate biased, unsafe output

Highly consequential

How to rein in GPAISs?

Future-proofing - GPAIS navigator programme:

Confidentiality & trust-based partnership for compliance by design

Monthly bilateral EC-GPAIS teams dialogues

Code of Practice for GPAIS R&D

Policy options for trustworthy GPAISs:

"Wait and see"

Horizontal GPAIS reqts (Articles 5, 9, 11, 15, 17, 18, 51)

Classify GPAISs as high-risk by default

[1] Contact: Nicolas Moës, Head of Operations (nicolas.moes@thefuturesociety.org, +32 488 541917)
[2] In Presidency Compromise Text (2021/0106(COD); 14278/21) of November 29, 2021; Title IVa, Article 52a entitled "General purpose AI systems".

What are GPAISs? Powerful AI systems applicable to many tasks in many fields.

There is no common definition of GPAIS across the scientific literature.[3] GPAISs have emerged and become popular thanks to growing amounts of compute and data and to innovative methods to tap them. **Technically, current GPAISs are characterized by their scale (measured in number of parameters[4])**, 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 generalization). **Functionally, GPAISs are characterized by their widespread use as pre-trained models for other AI systems**. For example, a single GPAIS for language can be used as the core for several hundreds more applied models simultaneously[5] (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[6] last year found that a few GPAISs are the basis for almost all applied models for language.

Originally used for language processing and computer vision, GPAISs are increasingly the foundations for powerful applications in medicine and healthcare (e.g. Clinical BERT), finance (e.g. FinEAS), life sciences and chemistry (e.g. C5T5), and, interestingly, for programming (Codex) and machine learning (e.g. Decision Transformer). **Theoretically, nothing prevents the application to other fields of the techniques underlying existing GPAISs' success.** Moreover, GPAISs are not limited to a single type of information: with enough training and compute, they can process audio, video, textual, physical and structured data feeds, including any representations humans have developed (natural language; genetic, protein, or chemical data; medical, financial and scientific data; etc.). Given the lack of a common definition, it is difficult to assess which models fall under "GPAIS", but these are currently in use: Aleph Alpha's private beta, BART, BERT, CLIP, Codex, DALL•E, data2vec, Ernie 3.0 'TITAN', GLaM, GLIDE, Gopher, GPT-3, GPT-J 6B, Jurassic-1, Megatron-Turing NLG, MuZero, Perceiver IO, RETRO, Switch Transformer, T5, and Wu Dao 2.0 are the main ones to come to mind.

Why care about GPAISs? Strategic autonomy and risks to fundamental rights & health.

GPAISs' importance for EU strategic autonomy cannot be understated, as they are rapidly enabling remarkable progress across many sectors. If data, compute and programming

[3] The most thorough -though exploratory- discussion of GPAIS can be found in Bommasani, Liang et al. (2021) "On the Opportunities and Risks of Foundation Models", which has been used in multiple places throughout this memo.

[4] Parameters in AI are the elements of the program that are tailored automatically through training. GPAISs have several billion parameters or more. The biggest GPAISs available at time of writing have over 1 trillion parameters (Switch Transformer: 1.6 trillion; Wu Dao 2.0: 1.75 trillion). However, there is significant research into training efficient GPAISs 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 (e.g. RETRO, LaMDA, DeepSpeed-MoE)

[5] For example, GPT-3 is used by "tens of thousands of developers" (GPT-3 Powers the Next Generation of Apps) and in over 300 apps currently placed on the market covering over 110 intended purposes. This figure is growing daily: between January 5th 2022 and January 23rd 2022, it went from 303 to 323 (+6.5%). Similarly, the original version of BERT (the model underlying Google Search in 70 languages) has been "forked" into at least 8,400 models (GitHub - google-research/bert: TensorFlow code and pre-trained models for BERT)

[6] *"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)

languages are the raw materials of the AI revolution, GPAISs will become its infrastructure. Because hundreds of applications already rely on GPAISs and many more GPAIS-based systems will be deployed every year, our increasingly smart digital ecosystems will be built upon them.

The few existing GPAIS providers therefore concentrate significant market and political power: a growing number of European software developers' businesses depend on reliable access to them,[7] and millions of EU citizens' lives are or soon will be affected by GPAISs. The current version of the EU AI Act exacerbates this power by shifting all responsibilities for GPAIS compliance and alignment with European values away from GPAIS providers and onto the consumer-facing developers. This disincentivizes the latter to adapt these cutting-edge technologies out of fear they be untrustworthy, since these developers have little control or credible information over e.g. GPAIS biases, accuracy and robustness or GPAIS providers' quality assurance system.

Meanwhile, GPAISs are opaque and have been shown to pose risks to fundamental rights and safety. They are complex and often exhibit surprising behaviors, unexpected even by their own developers. For example, GPT3 -trained to process human natural language- famously acquired by accident the ability to write rudimentary programs in a programming language, to the surprise and excitement of its developers.[8][9] Like other AI systems, these models pose risks: some GPAIS have empirically been shown to exhibit anti-Muslim and anti-Semitic biases, misogyny, and racism.[10] One GPAIS embedded in a chatbot told a test user that it thought the user should kill himself.[11] These risks are fully acknowledged by their developers,[12] and some GPAIS development teams have set up safeguards such as internal risk assessments or lists of use cases for which their model cannot be applied. These self-governance norms and practices are however not sufficiently mature, structured and widespread.

Unlike other AI systems however, **the repercussions of these risks are wide-ranging: a single design flaw in a GPAIS echoes through hundreds or even thousands of applications built upon it.** Given the centralization of their control and their criticality to many apps, they pose a significant cybersecurity, safety, fairness, accountability, and privacy risk. Without greater investment in their reliability and safety, there is little guarantee that their unexpected behaviors witnessed so far will not provoke accidents or highly consequential downtime for EU governments, businesses, and citizens in the future.

[7] 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.

[10] Rooting Out Anti-Muslim Bias in Popular Language Model GPT-3

[11] Nabla - Doctor GPT-3: hype or reality?

[12] Language modelling at scale: Gopher, ethical considerations, and retrieval | DeepMind

How to rein GPAISs in?

1. Future-proofing: GPAIS navigator programme for compliance-by-design.

Harnessing GPAISs' power while mitigating their risks will be crucial to the EU strategic autonomy and to the future of AI adoption and development in the EU. Given what's at stake, **it is crucial to find a balanced approach that reconciles innovation and trustworthiness.** There is fast-paced innovation in the field: following an intriguing and unexpected behavior by their GPAIS, a development team can iterate on it and redevelop the GPAIS for an entirely different field from one month to the next (e.g. from question-answering to automated programming).[13] This lack of predictability reduces the relevance of static, "blanket" (AI industry-wide) governance mechanisms in favor of adaptive, flexible controls tailored to the model. Fortunately, we observe that, so far, GPAISs have often replaced one another. For example, ELMo, ULMFiT, and the two precursor versions of GPT 3 have all waned in popularity and commercial relevance, replaced by some of the 21 models mentioned in the first section. As their ecosystem exhibits oligopolistic dynamics,[14] **only a few GPAIS development teams are relevant to regulators** currently and, if these dynamics persist as the industry matures, we should expect only a linear rather than exponential growth in the number of GPAIS development teams.

Considering these facts, we prefer an agile but scalable governance solution targeted explicitly at the few most important GPAISs at any given time: **we recommend instituting a navigator programme with teams developing AI systems identified as GPAISs to ensure proportionate compliance-by-design.** Inspired by the monitoring and reporting programme on disinformation, this navigator programme would involve the establishment of a joint Code of Practice for GPAIS R&D and regular conversations between GPAIS developers and authorities. The European Commission's trained staff would be in direct bilateral relation with each GPAIS development team to discuss monthly relevant advances and implications, under a strict confidentiality agreement. These discussions would cover:

- the latest progress and experimentation in the GPAIS team including findings related to unexpected behaviors and upcoming research projects,
- design measures taken to identify and mitigate risks prior to development,
- demos of new versions of the model and of its compliance-by-design features,
- brainstorming and suggesting steps to take to manage the model's implications for fundamental rights, health and safety in line with the AI Act, such as accuracy and robustness,
- measures in place among the developers team for quality assurance and risk management,
- the current usage of the GPAIS by other actors, including the estimated number of end-users affected by the GPAIS output monthly, the sectoral, functional, geographic

[13] OpenAI Codex Live Demo; [2107.03374] Evaluating Large Language Models Trained on Code

[14] GPAISs require significant data, compute, and talent to develop, making it profitable for industry actors to supply only a few at a given time. Moreover, on the demand side, once a given GPAIS has been widely leveraged for fulfilling a given function, the users' willingness to pay for another GPAIS for the same function significantly decreases. As a result, *"[a]lmost all state-of- the-art NLP models are now adapted from one of a few [GPAISs]"* (p. 5 in Bommasani, Liang et al. (2021) "On the Opportunities and Risks of Foundation Models")

and demographic distribution of applications based on the GPAIS, the novel applications, etc.
- the adequacy of the self-regulatory measures and of the help provided by the authorities for compliance with the Code of Practice,
- the adequacy of the Code of Practice for EU innovation and AI adoption and trustworthiness,
- the state of the art of GPAIS R&D and the identification of new competing development teams worldwide that would benefit from joining the navigator programme.

Through these discussions, the European Commission's staff and GPAIS providers would build mutual understanding and a common evidence base over the years, as the foundations of a trust-based partnership: **the civil servant becomes a navigator for the developers when it comes to the legal implications of their work for health & safety and fundamental rights, and the developers become navigators for the civil servant when it comes to the latest technological advances.** This will help them further converge over the years on the best approaches to governing these systems and give the European Commission experience and knowledge required to design new smart governance mechanisms to implement that approach.

As mentioned, the civil servant in direct communication with the developers of a given GPAIS would be bound by a strict confidentiality agreement. Moreover, civil servants in the navigator programme team would also be barred from applying what they learn for personal or commercial gains through a 5-year cooldown period. To further prevent accidental leakage of commercially sensitive information, **all correspondence would be treated confidentially and each civil servant would be in charge of a single GPAIS at a time, until that model loses relevance.**

The navigator programme's Code of Practice would be developed and reviewed regularly based on the evolution of the field. A GPAIS provider's formal observance of the Code would be required if its GPAIS or versions of its GPAIS is used in apps put into service or made available on the EU market or to EU citizens. The Code of Practice and its signatories would be public, and the navigator programme's team would enforce it.

2. Policy options: wait-and-see, GPAIS requirements or high-risk classification?

Once the navigator programme is in place, the EU has a broader set of policy options available to protect the fundamental rights, health and safety of its citizens and its societal interests. It should be noted for all policy options that the identification of what systems are in-scope is currently difficult: there is no common technical definition nor an actionable legal definition of GPAIS. Specifically, we currently lack a satisfactory definition of "generality" to define *general* purpose. Until 2021, the notion of "scale" (e.g. number of parameters or amount of compute needed for training) was a good proxy for generality, but the latest research in GPAISs has shown that general performance can be achieved without scale.[15]

[15] [2112.04426] Improving language models by retrieving from trillions of tokens and, less explicitly, LaMDA and DeepSpeed-MoE.

While the navigator programme's membership can be extended or restricted following the evolution of the technology, policy should preferably rely on objective criteria.

Nevertheless, an absence of definition does not mean GPAISs are not having implications for EU citizens' fundamental rights, health and safety. We therefore describe three approaches to governing GPAISs. They each differ from each other by the extent to which they apply the precautionary principle. All can be enforced with the support of the navigator programme proposed above.

2.1 Disregard the precautionary principle: wait and see, ready to (re)act and regulate

With this approach, EU authorities hold on before taking regulatory decisions about GPAISs, with the intention to gather more information on the harms and benefits they generate for society. It relies on the GPAIS navigator programme to be robust enough and to provide sufficient timely information to the European Commission. It also requires speedy regulatory reactions before too much harm is done to consumers or businesses applying GPAISs.

The advantage of this approach is that it does not encumber GPAIS development teams of time-consuming compliance procedures and enables maximal freedom to innovate, including in the technical safeguards they might develop. It also is the simplest to implement, as it simply relies on the Code of Practice for GPAIS R&D. The disadvantage, however, is that it precludes legal certainty: while some GPAIS providers have proactively put safeguards around the use of their GPAISs (ethics board, terms and conditions, etc.), some GPAIS providers have been much more hands-off in their public release of powerful systems.[16] Investors, entrepreneurs and developers who are conscious of the risks to fundamental rights, health and safety are disadvantaged vis-a-vis those who ignore these dangers. This race-to-the-bottom reduces the overall trustworthiness in the market for GPAIS, which could percolate in the EU markets as lower AI adoption.

2.2 Soft precautionary principle: horizontal GPAIS requirements and obligations

With this approach, the EU imposes a limited set of requirements to identified GPAIS development teams. Only those requirements that are reasonably expected to reduce the risks of harm "downstream" (i.e. in the hundreds of applications of these GPAISs) should be considered. These requirements should establish outcome-based targets for accuracy, robustness and control rather than procedural requirements which may quickly become obsolete. The outcomes to achieve should evolve in line with the state-of-the-art in the relevant academic fields. Moreover, the requirements should be technology-neutral to avoid distorting the innovation in potential technical and institutional safeguards that GPAIS providers develop. The navigator programme staff would accompany, assess and enforce the compliance with these requirements.

To avoid a fragmentation of the governance regime in the AI value chain, these requirements should be consistent with the EU AI Act as much as possible. Moreover, as GPAISs have by definition a general-purpose rather than an "intended purpose", only these requirements that

[16] Propaganda-as-a-service may be on the horizon if large language models are abused | VentureBeat

relate to its technical design and development would be relevant. Concretely, they could include any subset of the following (the related EU AI Act articles are indicated in parentheses):

- Registering in the EU database for Standalone high-risk AI systems (Article 51)
- Monitoring and preventing the use of GPAIS in prohibited AI practices (Article 5)
- Developing and maintaining a risk management system (Article 9)
- Writing up a technical documentation (Articles 11 and 18)
- Ensuring accuracy, robustness, control and cybersecurity (Article 15)
- Developing and maintaining a quality management system (Article 17)

Note that, given the EU AI Act focuses on AI systems with an intended purpose (as opposed to GPAISs), it is possible that some efficient regulatory requirements for GPAISs would be inefficient for other AI systems. The navigator programme would help identify these GPAIS-specific requirements over time.

The advantage of this approach is that it encumbers GPAIS development teams minimally. Through the navigator programme, it is also relatively easy to enforce, especially since it does not necessitate compliance with the broader set of “post-deployment” requirements in the context of market surveillance. The disadvantage, however, is that the developers team will face greater scrutiny from the navigator programme, which might prevent the accidental innovation from GPAISs’ unexpected behaviors, in particular by reducing the occurrence of these behaviors. Moreover, GPAIS-specific requirements will have to evolve alongside the technology in order to remain relevant without hindering innovation, which will take significant expertise and regular updates. The effectiveness of this approach will therefore hinge on the quality and efficacy of the navigator programme at generating policy-relevant expertise for developing smart governance mechanisms.

2.3 Hard precautionary principle: consider GPAISs as high-risk AI systems

With this approach, the EU AI Act would classify GPAISs as high-risk AI systems, regardless of what use cases they are applied to, in order to prevent the harms they might generate outside the Act’s identified high-risk areas. Concretely, this would require amending the Article 6 of the EU AI Act -“Classification rules for high-risk AI systems”- with a third paragraph specifying that general purpose AI systems (to be formally defined) will be considered high-risk AI systems if they are applied into AI systems that are placed on the market or put into service in the EU.

Instead of a soft regulatory approach, the navigator programme would then become an enforcement mechanism. It would help monitor compliance with ex ante and ex post requirements and obligations facing high-risk AI systems (Title III of the EU AI Act), in partnership with the mandated market surveillance authorities.

The advantage of this approach is that it shifts the burden for compliance of big, opaque GPAISs away from consumer-facing applications developers and onto the GPAIS providers themselves. GPAIS providers have greater control and information over the design and development of their GPAISs, and therefore would be best able at cost-effectively complying with requirements and obligations that pertain to design and developments. Once the GPAIS in question is compliant, applications developers would then be able to further assess whether their own re-use of the GPAIS is considered high-risk. If so, they'd be able to rely on the GPAIS compliance and conformity assessment as a first step and would thus only have to address obligations and requirements that pertain to the modification, deployment or use of the AI system built upon the GPAIS. Given that one GPAIS can be re-used by hundreds or even thousands of applications, having a single conformity assessment for the GPAIS would save authorities and industry collectively significant time and resources that would otherwise have to be dedicated for conformity assessment of the potentially hundreds of high-risk applications of the GPAIS.

The disadvantage of this approach is that it would involve significant adaptation to market surveillance. The consumer-facing applications developers would be considered like "distributors" of the GPAIS, yet face some conformity assessment obligations themselves. Adapting the market surveillance institutional ecosystem could prove challenging, even with the support of the navigator programme. Moreover, the developers team will face greater scrutiny, which might slow down innovation. It is also unclear to what extent the existing requirements for AI systems envisioned as products with intended purposes are cost-effective for governing GPAISs which are often construed as R&D projects rather than products. Additionally, the perceived compliance burden might incentivize GPAIS development teams to seek loopholes or to operate on a stealth-basis (i.e. away from public and regulators scrutiny through multiyear "private betas" of several thousand people) until they obtain sufficient capital to invest in complying with the EU AI Act. Finally, the GPAIS providers might choose to prohibit applications leveraging their GPAIS to be used by EU developers or for EU citizens, which would significantly slow down AI adoption in the EU.