CEPS IN-DEPTH ANALYSIS

RECONCILING THE AI VALUE CHAIN WITH THE EU'S ARTIFICIAL INTELLIGENCE ACT

Alex C. Engler and Andrea Renda

September, 2022 - 03

SUMMARY

The EU Artificial Intelligence Act (AI Act), proposed by the European Commission in April 2021, is an ambitious and welcome attempt to develop rules for artificial intelligence, and to mitigate its risks. The current text, however, is based on a linear view of the AI value chain, in which one entity places a given AI system on the market and is made accountable for complying with the regulation whenever the system is considered 'high risk'. In reality, the AI value chain can present itself in a wide variety of configurations. In this paper, we propose a typology of the AI value chain featuring seven distinct scenarios, and discuss the possible treatment of each one under the AI Act. Moreover, we consider the specific case of general-purpose AI (GPAI) models and their possible inclusion in the scope of the AI Act, and offer six policy recommendations.

- First, the AI Act should discourage application programming interface access for GPAI use in high-risk AI systems, in order to avoid cases in which providers place systems on the market that they cannot fully observe, let alone control.
- Second, the AI Act should envisage soft commitments for GPAI model providers to strengthen legal certainty and reduce transaction costs.
- Third, for high-risk AI applications, the AI Act should discourage value chain types in which a vendor builds software for a specific intended purpose that includes code for training machine learning models, but does not provide the data itself or pre-trained AI models ('software with AI model' in our typology).
- Fourth, the AI Act should explicitly exempt the placing of an AI system online as free and open-source software.
- Fifth, there is a need to clarify ambiguities concerning the identity and obligations of the providers of high-risk AI systems (PHRAIS) in several of the common business models discussed in the AI value chain typology.
- Sixth, the proposals made in the European Parliament's IMCO-LIBE draft report, adding specific new user requirements, should be incorporated into the final text of the AI Act.

Alex Engler is a Fellow in Governance Studies at The Brookings Institution, Andrea Renda is a Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy (GRID) at CEPS. The authors would like to thank Clement Perarnaud (CEPS) for his research support and Adelle Patten from the Brookings Institution for the illustrations used in this report. This study received funding support from the Future of Life Institute.

CONTENTS

INTRODUCTION

The launch of the European Commission's proposal for a new EU Artificial Intelligence Act (AI Act) in April 2021 triggered a lively debate on several outstanding issues, including on whom regulatory requirements should be placed in order to best mitigate the risks stemming from certain AI systems. The intensity of the discussion reflects the difficulty of writing comprehensive legislation with consistent definitions and responsibilities that will apply clearly and effectively to a diverse range of AI supply chains, in which developers, deployers and users may overlap or perform complex interactions[1].

At the same time, the debate is heavily permeated by repeated calls for a balanced allocation of responsibilities across the AI 'value chain', often with the underlying assumption that the latter can be described in a standardised way. As this paper will demonstrate, this is unfortunately far from the truth, as several alternative configurations of the AI value chain are observed on the market, not all of which appear to have been fully considered by EU co-legislators when drafting or amending the AI Act.

Of particular interest in this context is the issue of so-called general-purpose AI (GPAI) models, which were not mentioned in the original Commission proposal of April 2021, but are given unique treatment in the compromise proposal published in May 2022 by the Council of the European Union under the French presidency. GPAI models are characterised by their training on especially large datasets to perform many tasks, making them particularly well-suited for adaptation to more specific tasks through transfer learning. These models—especially those used for natural language processing, computer vision, speech recognition, simulation, and robotics—have become more foundational in many commercial and academic AI applications.

The variety of AI business models and the growing role of GPAI models challenge the assumption of the AI Act that there will be a clear distinction and linear relationship between a single provider of any AI system and its users. Further, depending on the circumstances, it may be difficult or even at times impossible for AI providers to build on third-party GPAI models, and to comply fully with the requirements of the AI Act. It is also unclear whether the existing requirements of the AI Act (e.g. concerning high-risk AI) can be applied appropriately to GPAI models without adaptation.

To help resolve these challenges, this paper examines specific archetypal configurations of the AI supply chain, taking into consideration the complexities presented by GPAI models and the potential for multiple developers. For these different scenarios, we consider who would be subject to the regulatory requirements of the AI Act, and the

[1] The EU presents the AI Act as 'horizontal', although in some ways, the scope of application of the Act is limited to those applications that fall under specific categories, i.e. limited, high and unacceptable risk. See the CEPS companion paper to this study (Bogucki, Engler, Perarnaud and Renda, 2022) for an illustration.

expected outcomes. Based on this analysis, we offer actionable recommendations for EU policymakers on how best to account for the complexity of the AI supply chain and the role of GPAI models in policy.

1. WHAT IS THE AI VALUE CHAIN?

The AI Act mentions the AI value chain in the first specific objective (1.4.2-1): 'To set requirements specific to AI systems and obligations on all value chain participants in order to ensure that AI systems placed on the market and used are safe and respect existing law on fundamental rights and Union values'. The AI Act does not clearly define the AI value chain though, leaving it instead to be interpreted from the body of the regulation.

We define the AI value chain as the organisational process through which an individual AI system is developed and then put into use (or deployed). Proactively considering which organisations may execute on which parts of the AI value chain is necessary to ascertain which organisation is best placed to conform with regulatory requirements, and more generally to ensure safe and responsible function of the AI system.

This focus on the organisational process distinguishes the AI value chain from algorithmic development, in which the focus is more narrowly on the technological steps of building an AI system. These concepts are related, however, as the AI value chain can be expected to differ across different types of algorithmic development (e.g. expert rules, search and supervised machine learning), as well as across different environments (e.g. products, applications and websites) into which the AI system may be integrated.

The development of an AI system will generally include the following stages:

- problem definition
- data collection and pre-processing
- model training
- model retraining
- model testing and evaluation
- integration into software
- model deployment

Prior research has attempted to define a typical AI workflow, illustrating how these tasks can come together. For instance, model training and model evaluation are often an iterative process, with many training attempts taking place before a model is deployed. AI system development may also include model updating (such as through transfer learning or online learning) and the combination or ensembling of multiple models. However, these stages are neither universal to all algorithms, nor guaranteed to be in this order. For example, an AI value chain may start with data collection, labelling, creation or

synthesis, as in the case of machine learning, but may not necessarily begin there, as in the case of expert systems or logic programming.

The Organisation for Economic Co-operation and Development has published a framework for classifying AI systems that includes some characteristics relevant to the AI value chain—including if the AI system continues to learn in the field, and if the model is customisable by the user after its initial development. The framework argues that 'understanding how an AI system's model was developed ... is a key consideration for assigning roles and responsibilities throughout a risk-management process'.

The different stages of AI system development offer break points that will typically manifest themselves in the AI value chain. For instance, in the AI value chain typology, there are scenarios in which one organisation develops an AI model, while another integrates that model into software. Companies specialise in certain types of AI system development, such as natural language processing or computer vision, and may provide these services (but not software integration) to other companies. Further, specialised companies may only provide access to an AI model through software or an application programming interface (API), without enabling direct interaction by a client. All of these scenarios have repercussions for AI governance, so this paper will systemically categorise different formations of AI development into a typology of the AI value chain.

Within the typology, this paper will discuss the entities on which the AI Act's requirements fall. It is therefore important to first understand how the original Commission proposal conceptualises the AI value chain.

2. HOW DOES THE AI ACT ENVISION THE AI VALUE CHAIN?

Although the proposed AI Act does not define the AI value chain, much can be inferred from the text, which envisions a relatively straightforward, linear value chain, with key entities categorised as either 'providers' or 'users' of an AI system.

- A provider is defined in Article 3(2) as 'a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge'.
- A user is defined in Article 3(4) as 'any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity'.

Under these definitions, a provider is essentially the last entity to develop or integrate an AI system into a product or software before it is either sold or used. If the provider is selling the AI system, a user is any purchasing entity that then uses the software or product for any non-personal use[2]. The AI Act also mentions two additional entities that are secondary to the importance of providers and users, namely importers and distributors.

The distinction between provider and user is especially important because these two roles carry nearly all of the regulatory responsibility under the AI Act. However, many organisations may become providers or users without triggering any of the AI Act's requirements. The provider's AI system must also meet the criteria specified by the AI Act when it is classified as high risk, or alternatively the transparency obligations under Title IV[3]. Determining which entities are classed as providers of high-risk AI systems (PHRAIS[4]) is especially critical for considering the impact of the AI Act. To qualify as a PHRAIS, the entity must both meet the provider definition, and provide an AI system that qualifies as high risk at that time. Therefore, a PHRAIS is a subset of AI system providers.

As will be discussed in the typology of AI value chains below, there are common business models that create ambiguity concerning the entities that will be classed as PHRAIS, and whether this is the ideal or intended outcome of the AI Act. Further, while for all high-risk AI systems there must be at least one PHRAIS, for some there may be several relevant PHRAIS.

Chapter 3 of the AI Act details the specific requirements for PHRAIS and users of high-risk AI systems. It assigns the requirements of Chapter 2 primarily to PHRAIS, including provisions on a risk management system, data governance, technical documentation, record keeping (when possible), transparency requirements, accuracy and robustness. The PHRAIS must also take corrective action if the AI system is found to be in violation of any of the above provisions. Finally, the PHRAIS must go through the formal legal registration process, including performing an *ex ante* conformity assessment procedure, registering the high-risk AI system in the EU-wide database, establishing an authorised

[2] The term 'users' is confusing, as it may infrequently refer to individuals. Some have suggested the more accurate term of 'deployers' (see https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/).

[3] Under the AI Act, AI systems are considered to be high risk in two circumstances: (1) if they are included in products falling under existing EU product safety legislation (e.g. for aviation, automotive vehicles, boats, elevators, medical devices and industrial machinery); or (2) if the intended purpose of the AI system falls into a set of categories specified by the AI Act under Annex III (e.g. biometric identification, critical infrastructure, educational access, employment and some uses by governments). The transparency requirements under Title IV apply to AI systems that interact with humans (e.g. chatbots), perform emotion detection or create manipulated content (e.g. deepfakes).

[4] Most easily pronounced as 'phrase'.

representative as a point of contact for regulators, and demonstrating conformity upon the request of regulatory agencies.

Compared to this set of requirements, the user-facing provisions are quite limited. Users of high-risk AI systems must use them as per their instructions, keep relevant records of their function, and monitor them for serious incidents or malfunctions that could undermine the AI Act requirements. Further, when the user has control over input data for an AI system, it must ensure that data is 'relevant in view of the intended purpose', but that data is not required to meet other criteria. Importers and distributors are of secondary importance, and must only verify and avoid undermining compliance, as well as cooperate with regulators, but otherwise have no additional burdens.

Critically, under Article 28 of the AI Act, a user, distributor or importer will themselves become a provider, and thus be subject to the regulatory requirements for providers if they place a pre-existing AI system on the market under their own name, change the intended purpose of the AI system, or make another 'substantial modification' to the AI system.

Article 28 goes on to state that the original PHRAIS will no longer be considered a provider under the AI Act if the user, distributor or importer changes the intended purpose or makes a substantial modification to the AI system. So, if a second entity, such as a user, takes a high-risk AI system and rebrands it before selling it, both that entity and the original provider are PHRAIS for that system, and both need to go through the conformity assessment and meet the other AI Act requirements.

However, if the second entity changes the intended purpose or makes a substantial modification to a pre-existing high-risk AI system, this second entity will be considered the PHRAIS, and the original entity will no longer be considered a provider (presumably only for this system, although the draft AI Act does not state this explicitly). In the latter case, the original PHRAIS will usually have gone through the conformity assessment process in order to sell their AI system, so the outcome may be that they are not subject to post-market surveillance and verification by EU regulators for this specific high-risk AI system.

The precise meaning of 'substantial modification' will play a key role in the application of the AI Act to the AI value chain. A substantial modification is loosely defined by the AI Act as an alteration of the AI system that changes its intended purpose or affects its compliance with the obligations for high-risk AI systems (Title III, Chapter 2). Recital 66 notes that changes to an AI system as a result of ongoing learning after the system has been put into use may or may not qualify as a substantial modification. It is left to the original provider of the AI system to determine in its conformity assessment what types of changes would constitute a substantial modification, in line with the existing legal understanding of this term under EU harmonisation legislation.

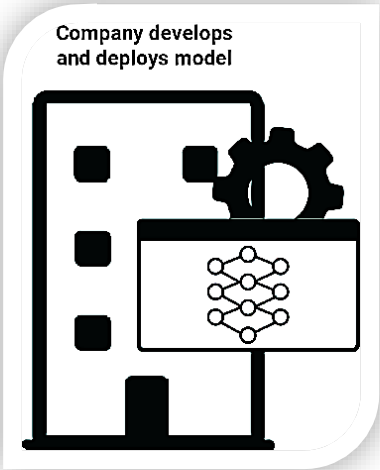
Lastly, Recital 60 states that 'relevant third parties ... involved in the sale and the supply' of AI systems should cooperate with providers and users to enable compliance. The AI Act does not mention so-called end users, e.g. individuals who may interact with an AI system, but did not themselves place it into use.

3. A TYPOLOGY OF AI VALUE CHAINS AND THEIR RELATION TO THE AI ACT

As explained in the previous section, the AI Act envisions a relatively clear distinction between two entities: the provider and the user. This division only directly addresses the situation in which the entirety of algorithmic development and design is done by the provider, while the user feeds input data into the AI system and monitors it for serious malfunctions. In this situation, the provider is a PHRAIS if the AI system meets the high-risk criteria, while the user can become the PHRAIS through Article 28 if it makes substantial modifications, changes the intended purpose or rebrands the AI system.

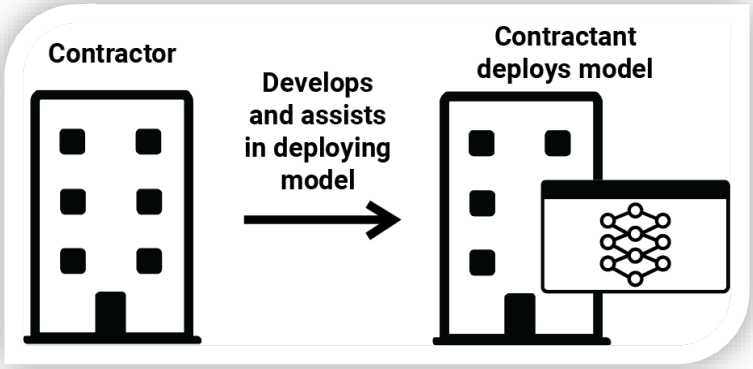
However, this interpretation of the AI value chain can be challenged when considering the wide range of common AI business models. The typology below considers scenarios with one entity (scenario 1), two entities (scenarios 2-6) and more than two entities (scenario 7). These models are expanded upon below with examples to clarify and an attempt to identify the entities on which the regulatory burdens will fall. CEPS held two expert workshops to scrutinise this typology.

Type 1 Internal AI development and deployment



In this case, a company writes the code and fully trains AI models for internal use within the same company's function. There is only one entity in this scenario, and it will qualify as both a provider and a user; a relatively simple interpretation of the AI Act. If the AI system qualifies as high risk, the provider will be a PHRAIS, and thus must go through the conformity assessment process and meet the other AI Act requirements. An example of this scenario is an AI system that manages and allocates tasks to warehouse employees, which was developed and is operated by the warehouse company.

Type 2 One entity develops an AI system for another entity (AI system contracting)



Under this scenario, one entity develops a bespoke AI system as a contractor for another entity (e.g. a company or government), but does not itself use or place the AI system on the market. As an example, a contractor may develop an AI system for financial fraud detection that is directly integrated into a bank's software. Despite partially, or even fully, developing the AI system, the contracted company cannot be considered a provider under the AI Act, as it does not place the AI system into use or on the market. Rather, the contractant (i.e. the entity acquiring the AI system from the developer) will be considered the provider, assuming it puts the AI model into use or on the market. The same entity will then become a PHRAIS if the AI system is classified as high risk.

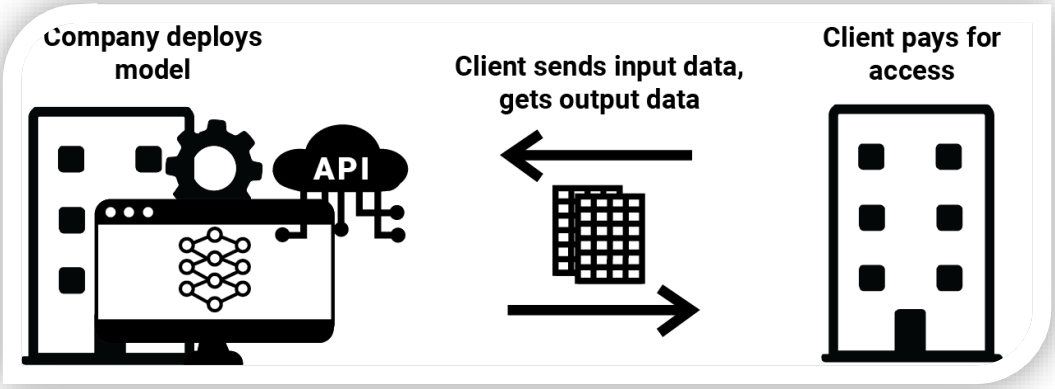
This approach is sometimes used by companies maintaining open-source software and AI systems, which sell technical support services to companies that have adopted those AI systems. While the AI systems they create are free and publicly available, the contractor markets its services based on its expertise in developing and using the open-source model.

Since the contractor is not defined as a provider, this scenario may at times create an incentive for companies to provide contracting services to develop an AI system while avoiding the AI Act requirements, rather than place the AI system on the market as software. In this case, the client would assume the PHRAIS responsibilities for an AI system that it fully controls and manages, but that it did not develop and may not fully understand.

What would need to be ascertained is whether this situation would trigger enhanced cooperation and information-sharing between the two entities, so that the PHRAIS can fully meet the AI Act's regulatory requirements. In these circumstances, meaningful compliance with the AI Act may require both entities, the contractor and contractant, to cooperate in carrying out the conformity assessment. We return to this issue in Section 4 below.

Type 3

One entity writes the code and trains the system, then sells access through a branded application or API (restricted AI system access)

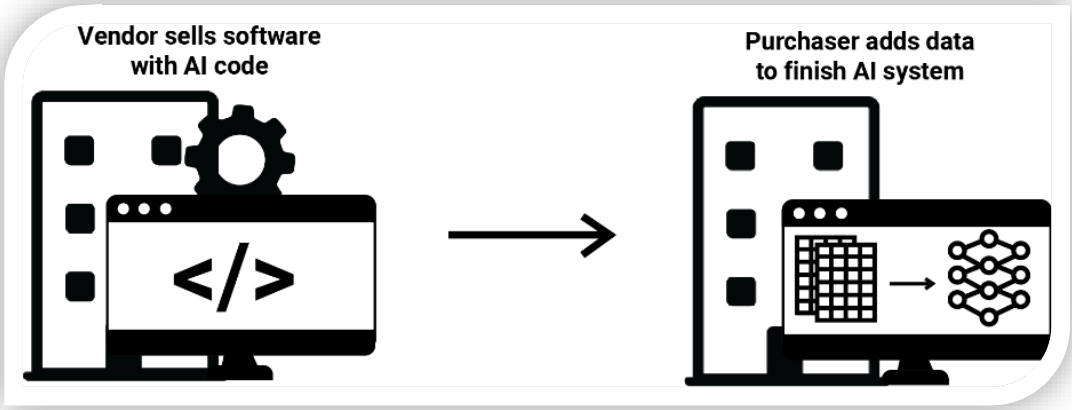


In this scenario, one entity writes the code and fully trains AI models within a branded application or API, then sells access to that application. The company enables access to an AI system, but only through closed-source software or an API, preventing any changes being made to the model.

Usually, along with Type 1 above, this is a simple interpretation of the AI Act, as the entity is clearly a provider since it develops and places the AI system on the market. If the AI system is high risk, the company will be considered a PHRAIS. Clients of that company can only submit input data and receive output, and are clearly users under the AI Act. A commercial facial recognition application is a well-known example of this business model.

Notably, the commercialization of GPAI models behind an API will complicate this scenario, as a client may apply the GPAI model to a purpose that counts as high risk. This is discussed more thoroughly in the next chapter.

Type 4 A vendor writes code for an AI system but does not pre-train it or provide training data to purchasers (software with AI code)



A vendor builds software for a specific intended purpose, including code for training machine learning models, but does not provide the data or pre-train AI models. In this situation, the vendor develops software that includes a process for a client to enter training data, after which the software's code automatically runs statistical analyses and presents the results in the software[5]. The vendor sells this software, and the purchasing entity inputs all the data used by the AI models, possibly with the technical assistance of the vendor.

Vendors who sell commercial software used for setting college tuition prices in the United States employ this approach. The vendor develops the software and includes code for analysing the data, but never sees the data from individual colleges. The colleges add the data themselves after purchasing the software, and the software automatically trains AI models, runs statistical analyses and presents the results.

The sale of software with AI code does not fall neatly under the AI Act, with some ambiguity regarding which entity qualifies originally as a provider, and the circumstances under which this might change. The most straightforward reading of the AI Act suggests that the vendor is the provider, as it has developed a technology that meets the definition of an AI system once data has been added. However, one could argue that the vendor

[5] This is not to be confused with statistical software, such as STATA or SAS, which enables a user to generally analyse data and run statistical methods that might include AI methods. Rather, this software automatically performs a specific set of data analyses and statistical methods when users submit training data.

has not created an AI system, since it cannot 'generate output'[6] until the purchasing entity has added training data, after which the software completes the final AI system.

Assuming the vendor does indeed qualify as a provider, and the intended purpose of the AI system makes it a high-risk AI system under the AI Act, then that vendor will be a PHRAIS. It is possible, however, that the purchasing entity could become a PHRAIS, depending on whether adding training data into the software counts as a substantial modification, as defined by the vendor in its conformity assessment.

This definition offers the vendor some leeway in deciding the circumstances under which its client (the purchasing entity) assumes the PHRAIS responsibilities. If the purchasing entity meets the substantial modification criteria, it will become a PHRAIS and have to go through a conformity assessment procedure, and the original vendor will no longer be considered a PHRAIS since it does not qualify as a provider according to Article 28[7].

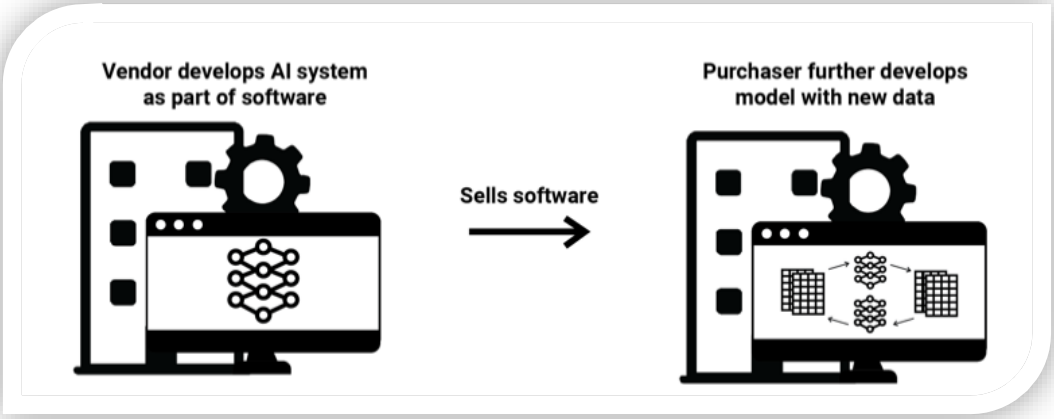
Typically, it may be safe to assume that a PHRAIS selling this kind of software would not want its clients to need to go through the conformity assessment process, as it might dissuade them from purchasing the software, and thus the vendor may prefer to be the only PHRAIS. It is unclear, however, whether either entity is able to fully perform the conformity assessment. Note that the vendor has access and control over the code, but not the data; and the purchasing entity has control over the data, but limited access and control of the code[8]. In this situation, meaningful compliance with the AI Act may only be achieved if both entities—the vendor and the purchaser—cooperate in carrying out the conformity assessment. Otherwise, the responsibility would fall on a PHRAIS that has neither access nor control over the underlying code, and as such may not be able to fully anticipate the risks that may be generated by the AI system, let alone identify possible mitigating measures.

[6] See the AI Act's definition of an AI system.

[7] Under the changes proposed by the French presidency of the Council of the EU, which delete Article 28, the vendor's status as a PHRAIS would be unchanged.

[8] Article 29(3) is also relevant, in that it notes that users are required to ensure that input data is relevant, but does not distinguish between input data that changes an AI system's behaviour and input data that is merely used to produce an output.

Type 5 Vendors of learning AI systems

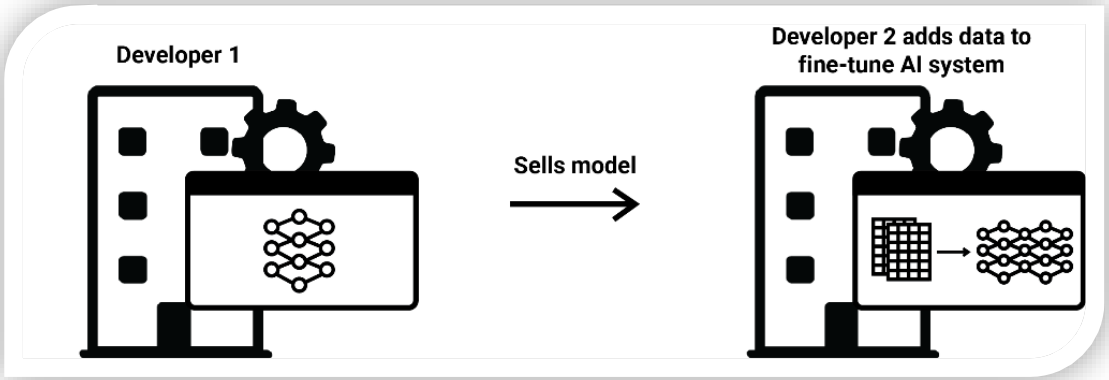


A vendor develops an AI system as part of a commercial software package or product that can update its models when given data by users, a process called online learning. The voice recognition AI system in Amazon's Alexa, which adapts to new data from specific human voices, is one such example, as are AI robotic arms, which continue to collect data and update while they are in use.

If the AI system is high risk, the AI Act considers the vendor to be the PHRAIS and the purchasing entity that buys the product to be the user. However, the purchasing entity could feasibly provide enough input data to meaningfully change the function of the software or product. Whether this is a substantial modification would depend on the pre-determined criteria of the original provider. This gives the vendor some leeway in deciding what constitutes a substantial modification by the user, potentially reducing its legal and regulatory exposure regarding the AI system in question.

In this scenario, the future legislative initiative on liability for AI systems may also clarify the definition of 'misuse' of an AI system by the user, hopefully in a manner that is consistent with the way in which 'reasonably foreseeable uses' will be defined in the AI Act. In fact, under the EU products liability regime, unforeseeable misuse by product users that results in causing unjust damage cannot trigger liability on the side of the producer, as the misuse does not make the product 'defective'.

Type 6 Initial development by one entity and fine-tuning by another (AI system fine-tuning)



In this case, one developer trains an AI system and then sells it to another developer, in such a way that the second developer can continue to train and develop the AI system. This second developer will then fine-tune the AI system, meaning it will use additional training data (controlled by the second developer) to improve the function of the AI system or adapt it to more specific circumstances or tasks. This is distinct from Types 3, 4 and 5 in that the second developer has direct access to the AI system (i.e. owns the AI model object) and is thus free to manipulate it.

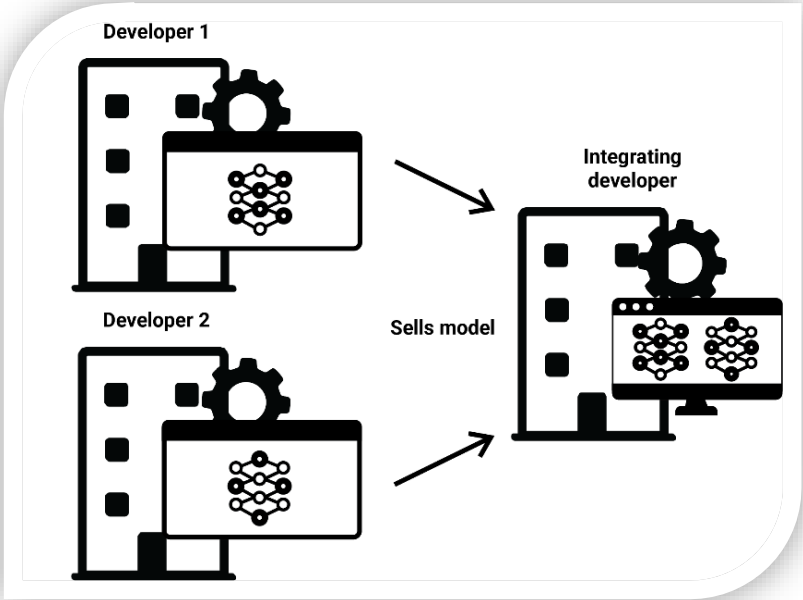
Typically, the dataset used to train the first AI system is larger and more general, while the dataset used to retrain is smaller and more specific. A general question-answering chatbot trained by a first developer, then retrained on the specific domain knowledge relevant to a second developer (e.g. on insurance information) and sold to a user, is an example of this business model.

In this situation, both developers are providers once they sell or use the AI system. However, the question of which developer is a PHRAIS depends on the circumstances. If the original AI system is not high risk, then its developer is simply a provider and not a PHRAIS. If the second developer purchases and then fine-tunes that AI system for a high-risk application, thus altering its purpose, that second developer is the only PHRAIS.

However, if the original AI system is high risk, and the second developer fine-tunes the model for the same or other high-risk purposes, then the first developer is a PHRAIS. The second developer may also be a PHRAIS, but this depends on the terms of substantial

modification as specified by the first developer[9] in its conformity assessment. Since the second developer has direct access to the AI system and can evaluate and modify its function, unlike in Type 4 (software with AI code) and Type 5 (learning AI systems), it is more reasonable that the second developer takes on the PHRAIS regulatory responsibilities.

Type 7 One entity integrates different AI systems into a new one (AI model integration)



A developer may take several pre-existing AI systems and integrate them together into one piece of software. In this scenario, there may be many developers. For instance, a company may combine AI systems for computer vision, audio analysis and natural language processing, all from different developers, into a single piece of software. Alternatively, integration may consist of combining several AI systems in order to improve their collective function on one specific task, known as ensembling.

Also in this category are so-called Socratic models, a novel technique whereby an AI system is trained to choose between several other AI systems for a given input. AI model integration may include several of the scenarios discussed above, particularly scenarios 2 and 6. Commercial invigilating software that monitors students (with AI models analysing video, audio, text, key entry, etc.) as they sit an exam is an example of this business model.

If the pre-existing AI systems are not high risk, and the integrating developer repurposes them to place a high-risk AI system on the market, the integrating developer will be the sole PHRAIS. If the pre-existing AI systems are high risk, the first developers will be

[9] If both entities are PHRAIS, the second developer may request the information from the first developer's conformity assessment as part of the purchase, potentially reducing compliance costs in connection with the second conformity assessment.

PHRAIS, and the integrating developer may also be a PHRAIS, as combining AI systems into a broader piece of software would likely constitute a substantial modification, as well as a change of intended purpose under Article 28. In this case, where the pre-existing AI systems are high risk, the original developers are PHRAIS and must go through the conformity assessment, but then the integrating company also becomes a PHRAIS and must go through a conformity assessment for the new integrated AI system and, specifically for this integrated AI system, the earlier developers are relieved of the PHRAIS label under Article 28[10].

3.1 KEY TAKEAWAYS FROM THE AI VALUE CHAIN TYPOLOGY

The heterogeneity of the AI value chain shown above, with at least seven distinct process types, reveals a number of takeaways for the future application of the AI Act. These include:

(1) **In several types of AI value chain, there is considerable ambiguity regarding which entity will typically be the PHRAIS**. This is especially true for Type 2 (AI system contracting) and Type 6 (AI system fine-tuning), although it may also apply to Type 4 (software with AI code), Type 5 (learning AI systems) and Type 7 (AI model integration).

(2) **In several types of AI value chain with two entities, neither entity will be completely capable of evaluating and altering an AI system in order to meet the regulatory obligations of PHRAIS**. This is especially true of software with AI code, as neither entity has control over both the code and the data, which are two essential components to determining the qualities of a resulting AI system. This is also at times the case for GPAI models, for example when they are used through restricted AI system access, but also potentially through AI system fine-tuning and AI model integration. This will be examined more closely in this paper in the discussion on GPAI.

(3) **There are realistic circumstances in which the designated PHRAIS pertaining to one specific AI system will shift between entities**. This will occur due to the circumstances described in Article 28, where either the intended purpose of the AI system has been changed, or alterations to that AI system meet the definition of a significant modification.

In the scenario where a high-risk AI system has been altered sufficiently to change the PHRAIS, the reading of the AI Act becomes somewhat complex. The first developer will have developed an AI system, gone through the *ex-ante* conformity

[10] Similar to the situation in Type 4 (software with AI code), under the changes proposed by the French presidency of the Council of the EU, which delete Article 28, the earlier developers' status as PHRAIS would remain unchanged.

assessment process, registered the system in the EU database, and at that time been considered as the PHRAIS. The second developer will then have taken that AI system, changed its purpose or made a substantial modification, become the PHRAIS (removing the legal label from the first developer for that specific AI system) and gone through the *ex-ante* conformity assessment, before using or selling the AI system.

(4) **Contract language will play a key role in many types of the AI value chain.** In the many potential instances of uncertainty described above, contracts between entities in the AI value chain will become quite important. This includes ensuring that pertinent information, such as prior conformity assessments, is shared in Type 6 (AI system fine-tuning) and Type 7 (AI model integration) scenarios. This contract language may also ensure cooperation between two entities for the conformity assessment process or specific AI requirements, such as risk management, in Type 2 (AI system contracting) and Type 4 (software with AI code).

3.2 LIMITATIONS OF THE PROPOSED TYPOLOGY

It is important to note that this typology is likely incomplete in many ways, as many business models may fall between these categories. For instance, companies such as OpenAI are beginning to offer fine-tuning of an AI model through an API, a service that lies somewhere between Type 3 (restricted AI system access) and Type 6 (AI system fine-tuning). Many businesses may also engage in more than one of these business models, as they are certainly not exclusive. This may be especially true of Type 2 (AI system contracting), which may be valuable to clients in addition to purchasing AI-driven software or an AI model. Furthermore, this typology does not consider businesses that provide ancillary services, such as model evaluation, operations or compliance services, which are all growing cottage markets.

Most notably absent from this typology are data collection, aggregation and pre-processing, leading up to the initial development of an AI system, which can be considered an independent stage of the AI value chain and are often performed by a separate entity. These entities are omitted from the AI value chain typology because there is no scenario in which regulatory requirements fall directly onto them. Data collection alone cannot lead to an entity becoming a provider under the AI Act. Still, it is worth noting that a PHRAIS will have obligations concerning the data collection process, especially under Article 10 on data and data governance, among others. To meet these obligations, when a PHRAIS purchases or receives data from other entities that performed the collection, it will have to use contract language for any transfer of legal obligations to the data collecting entity.

4. GENERAL-PURPOSE AI MODELS AND THE AI ACT

GPAI models are especially relevant to some of the types of value chain presented, as they feature in some of the more complex forms of AI value chain and have properties that warrant further consideration. GPAI models[11] are characterised by their training on especially large datasets using relatively high amounts of computation to perform many distinct tasks. This makes them especially well-suited for adaptation to more specific tasks through transfer learning. These models—especially those used for natural language processing, computer vision, speech recognition, game playing and robotics—are starting to become more foundational in commercial and academic application of AI[12].

Although the idea of GPAI is agnostic to any specific algorithm or method, the current generation of GPAI models revolves around deep learning. Due to the benefits of scaling deep learning models, continuous improvements have been made to deep learning GPAI models driven by ever-larger investment to increase model size, computational resources for training and underlying dataset size, as well as advances in research.

The ability of GPAI models to perform distinct tasks is especially important for the definition of 'general purpose'. For example, a GPAI model for natural language could perform both sentence completion and summarisation; two distinct tasks with differently formulated responses. This is different to performing the same task in varying environments or circumstances. A facial recognition model that is used by both law enforcement and in a retail store would generally not be considered as GPAI, as it is only performing one task (matching a face to a database of faces) despite the change in environment.

Typically, in order for an AI model to perform a task, it often receives additional training data that is specific to the new task, called retraining or fine-tuning. One perceived advantage of recent GPAI models is relatively strong performance with a limited number of new training samples (called 'few shot' learning), including one or even no additional samples ('one shot' and 'zero shot' learning, respectively).

Just as critically, the idea of GPAI does not suggest that the model can do any task. In fact, GPAI models are generally highly limited to a select number of tasks. For example, PaLM, a recent large language model developed by Google, might be able to perform many distinct language tasks, but still remains entirely unable to function as a recommender

[11] We use the term 'GPAI model' so we can differentiate from the proposed definitions in the AI Act text, which refer to a general-purpose AI (GPAI) system. It is our opinion that 'multi-task' and 'multi-modal' are more accurate descriptors of most models currently described as 'general-purpose' models. We use the term GPAI in response to the EU proposals, rather than as an endorsement of the terminology.

[12] So much so that Stanford HAI has begun to call them 'foundation' models, although the term 'base models' is more broadly accepted in the field of machine learning.

system for movies or make predictions for targeted advertisements. GPAI models are unable to generalise to completely different data types outside of their training data. Even more generally, while GPAI models show some ability to learn new tasks as they are scaled, many core underlying limitations in understanding remain difficult to overcome.

Many GPAI models produce one type of output data: Gopher and Chinchilla produce text in human language; AlphaCode and Codex attempt to write code; Stable Diffusion and StyleGan2 output images. However, the release of multi-modal models, which can produce multiple types of output data, is becoming more common. Most often, this includes models for both images and text, such as CLIP and DALL-E2. The most recent GPAI models push further, combining text, images, game-playing and robotic arm manipulation (e.g. DeepMind's Gato). Still, even these relatively expansive GPAI models have clear limits. For example, DeepMind states that Gato, at the very high range of the spectrum of GPAI models, can perform 604 tasks.

Aside from internal deployment, GPAI is typically commercialised through paid API access or by transferring the model, either by selling or open-sourcing the model object. However, it is also possible to provide access to GPAI models in commercial software or through the bespoke creation of GPAI models as a contractor, although these business models appear to be less common.

- Many GPAI models are made available through APIs, which enables the use, but not modification, of a GPAI model (Type 3, or restricted AI system access, as described in the typology). In this business model a provider develops a GPAI model and enables access to that model through the internet, often with prior approval or requiring payment[13].

  Clients who use the model are able to submit input data, which runs through the model and returns the output to the client. However, the client cannot examine the GPAI model directly, which is a significant limitation on its use. Historically, the client could also not retrain or fine-tune the GPAI model for specific tasks, but this may be changing. As mentioned above, OpenAI now offers fine-tuning of an AI model through an API, a service that lies somewhere between restricted AI system access and AI system fine-tuning. Still, a client is likely to be limited by cost, and possibly by a hard cap, on the number or rate of requests (i.e. the amount of input data sent to the GPAI model), which can further limit the client's ability to evaluate or retrain the GPAI model's function.

- A second business model for GPAI is to sell or open-source the GPAI model object, which is the digital representation of the trained model. The GPAI model object

[13] OpenAI enables the use of GPAI models via API through both paid access and approved access. The first option is a revenue generator and the second is to receive feedback and free media (see https://openai.com/api/pricing/ and https://openai.com/dall-e-2/).

should not be mistaken for the code, and also does not inherently include the training dataset. Rather, the model object is the computational artefact that results from running the code on the dataset. In addition to being able to use the GPAI model, direct access to the model object offers several distinct advantages for clients, the biggest of which is the ability to fine-tune (or 'retrain') the GPAI model for specific tasks (as in Type 6, AI system fine-tuning; or Type 7, AI model integration). For instance, a bank may want to access a GPAI model trained for natural language tasks, then fine-tune it to the bank's customer service documents and other domain-specific data.

This model might lead to a more effective chatbot or customer service tool than would otherwise be possible, such as in the case of access via API. Direct access to the model object also enables clients to interrogate and evaluate the function of the GPAI model more thoroughly. In addition to selling or open sourcing the pre-trained GPAI model, a firm may also provide consulting services in its fine-tuning or retraining, as well as assistance with the model's integration into the client's software or website (Type 2, AI system contracting).

Technically, both avenues (API and model object access) can be offered for free. However, when access to the model object is made free of charge and publicly available, the GPAI model becomes open source. Aside from potentially selling associated contracting services, this is also a business strategy, as open-sourcing code libraries and AI models can gain clout for a company, create a pipeline of data scientists trained on their systems, enhance their access to cutting edge research, and lead to other benefits.

5. AI ACT PROPOSALS AFFECTING THE AI VALUE CHAIN AND GPAI

The original text of the AI Act proposed by the Commission does not explicitly mention GPAI, nor does it distinguish between these models and other types of algorithms. Therefore, GPAI models only fall under the AI Act when their application meets the same criteria as any other model, such as being used for a high-risk purpose, interacting with humans in a way that triggers the disclosure requirement, or being used for unacceptable or banned applications. The fact that GPAI models are not mentioned in the original draft AI Act does not exclude or exempt them, but rather assumes that the existing governance mechanisms and value chain also effectively apply to GPAI models.

Under the business models described above, it is clear that the companies developing and selling GPAI models would qualify as providers. However, they would not likely qualify as PHRAIS, due to the absence of an intended purpose for the GPAI models. Without an intended purpose, a GPAI model cannot be a high-risk AI system. Further, when GPAI developers sell a model object, that model will generally need retraining and fine-tuning, which likely qualifies as a substantial modification under Article 28, making the company that purchases the GPAI model the PHRAIS in any high-risk use.

Alternatively, when a GPAI model is made available over an API, it will not qualify under any of the specific categories that trigger requirement of the AI Act, such as by interacting with humans or having a high-risk purpose or unacceptable risk application. Once again, the GPAI developer would not be the PHRAIS or otherwise subject to the AI Act. This means that, although GPAI developers are not specifically exempted or excluded from the AI Act as proposed by the European Commission, they will usually, perhaps almost always, be exempt from the AI Act's oversight.

Against this backdrop, during the negotiations on the AI Act both the Council of the EU and the European Parliament tabled possible amendments, with the aim of including GPAI in the scope of the AI Act. These amendments are described below.

5.1 AMENDMENTS PROPOSED BY THE COUNCIL OF THE EU

On 29 November 2021, the Slovenian presidency of the Council of the EU proposed an addition to the AI Act that, although appearing to exempt GPAI models, functionally just stated more clearly that they are already exempt, as explained above. The proposal added an Article 52A and Recital 70a that both deal specifically with GPAI, stating that the 'placing on the market, putting into service or use of GPAI should not trigger any of the requirements under the AI Act'.

The new article and recital make it clear that simply putting a GPAI model on the market without an intended purpose would not trigger any part of the AI Act and, conversely, putting a GPAI model on the market with an intended purpose covered by the AI Act would trigger its requirements. This should be interpreted as a clarification of the pre-existing intention of the AI Act, but the ensuing proposed changes move away from this approach. The article makes only a passing attempt to define a 'GPAI system'[14] (which is sensible, since there is no distinction in treatment) and does not distinguish between open-source and proprietary AI systems, unlike later proposals.

Meanwhile, the proposed additions also include a new research exemption for AI systems under Article 2(6) and (7), which exempts AI systems only used for internal research and development, provided that they are not placed on the market. An addition to Article 28 states explicitly that taking an existing AI system that is not high risk and modifying its intended purpose to a high-risk task triggers the PHRAIS requirements. This change is especially relevant to Type 5 (AI system fine-tuning) and Type 7 (AI model integration).

[14] A GPAI system is understood as 'an 'AI system that is able to perform generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation, etc.' We prefer the term GPAI model, see *supra* note 11. To maintain consistency with the wording used by the EU institutions, we maintain the term 'system' in this section.

On 3 February 2022, the French presidency of the Council of the EU proposed deleting the entirety of Article 28 from the AI Act. The intention of this change was difficult to interpret at the time, but the later consolidated version of changes published by the Council on 15 June showed significant alterations to Recital 66 that added clarity. Without Article 28, rather than shifting the legal label of a PHRAIS from one entity to another, the Council's proposed text argues that the substantial modification of an AI system results in a new AI system. This would change the AI Act's interpretation of some scenarios in the AI value chain, specifically Type 4 (software with AI code) and Type 7 (AI model integration).

On 7 March 2022, the French presidency of the Council of the EU proposed exempting microenterprises (i.e. companies with up to 10 employees and an annual turnover of less than EUR 2 million) from the requirement to create a quality management system for high-risk AI systems. This change would not relieve any requirements for micro, small or medium-sized enterprises.

On 13 May 2022, the French presidency of the Council of the EU proposed a substantial change in the approach to GPAI models. This proposal deleted Article 52a and Recital 70, and added several sections to Article 4 and Recital 12aa applying significant parts of the AI Act requirements to GPAI developers, even when they have not placed the GPAI system on the market for a use covered by the AI Act. Article 3 defines a GPAI system as an AI system that performs 'generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others'. The proposal is agnostic as to how the AI system becomes available, even explicitly including open-source software (although with important exemptions).

The text goes on to ascribe a specific and exclusive set of requirements to GPAI systems, regardless of whether the AI systems are fine-tuned by another entity. GPAI systems that may be used as components of high-risk AI systems have to comply with articles regarding the risk management system, data governance, technical documentation and transparency instructions, as well as on accuracy, robustness and cybersecurity.

Under the proposal, GPAI systems must be registered in the EU-wide database along with high-risk AI systems, and developers must declare their GPAI systems as being in conformity. If a GPAI developer releases a system that can be used in a high-risk AI system, the GPAI developer must meet the requirements above. Given the many categories of AI systems in products and standalone AI systems that can fall into the high-risk category of the AI Act, functionally this means that all GPAI systems would trigger these requirements. Critically, small and medium-sized enterprises (i.e. those with fewer than 250 employees and an annual turnover of less than EUR 50 million) are exempt from the requirements on GPAI models.

This is a substantial change in the AI Act's approach to GPAI models, specifically addressing Types 6 and 7 in the above typology of value chains. In Type 6, a GPAI developer who sold access (via a model object or API) to an entity that adapted the GPAI model for a specific high-risk purpose would previously have been exempt from all requirements in the AI Act, and the purchasing company would have been a PHRAIS. Now, there are two providers, a GPAI system provider and PHRAIS, with different requirements under the AI Act. Both need to register their AI systems in the EU-wide database.

This also applies to Type 7 value chains, where a GPAI system is integrated into a broader AI system but is not necessarily fine-tuned. If an integrating company purchases a GPAI system from a GPAI developer and integrates it into a high-risk system, both entities have distinct requirements under these proposed changes (as a PHRAIS and GPAI provider respectively). Notably, if the company integrating these systems purchases a different AI system that does not qualify as GPAI, the company it is purchasing from may or may not qualify as a provider, depending on their relationship (e.g. as a contractor versus a software provider, as discussed above). This may create a disincentive to sell GPAI models, when selling task-specific models may avoid this regulatory burden[15].

This change is also noteworthy for its significant shift in the treatment of open-source GPAI models. Under the French presidency proposal, making a GPAI model publicly available (e.g. on websites such as GitHub or Hugging Face) triggers the requirements noted above if the model could be used for high-risk applications. Given that GPAI is designed for many tasks, it seems highly unlikely that any GPAI model could not plausibly be used for any high-risk application. Notably, the proposed exemption for research purposes would not exempt open-source GPAI models, as the Council text clearly states that open-source models count as being placed on the market. Functionally, these requirements land on all GPAI systems, although there is an important exemption.

There is an exemption if the provider 'has explicitly excluded any high-risk uses in the instructions of use or information accompanying the general-purpose AI system'. This seems like an easy standard to meet, but the exemption is not valid if 'the provider has sufficient reasons to consider that the system may be misused'. Further, if the provider discovers 'statistically significant trends of market misuse' it must take action to prevent future misuse. This may be easier for providers to meet when using APIs, who can disable access for some users, than for open-source developers, who cannot.

[15] It is not immediately clear in what circumstances a company can make adjustments to transform a GPAI model in appearance or function into a more task-specific model, but this warrants further consideration.

5.2 EUROPEAN PARLIAMENT DRAFT REPORT

On 20 April 2022, a draft report from the European Parliament's Internal Market and Consumer Protection (IMCO) and Civil Liberties, Justice and Home Affairs (LIBE) committees proposed a series of amendments to the AI Act, but seemingly did not change the AI Act's approach to GPAI. The report's explanatory statement says that GPAI should not be explicitly excluded from the AI Act, which signals broad agreement with the approach of the original Commission text.

However, the IMCO-LIBE report does affect the interpretation of the AI value chain. By adding a paragraph to Article 10 (data and data governance) and a new Recital 45(b), the report seeks to deal with the uncertainty pertaining to Type 4 (software with AI code). Specifically, the article and recital state that, through contractual language with a PHRAIS, a user may be made responsible for the data governance requirements in the AI Act if they have sole access to the data used. So, if a PHRAIS builds AI code for a high-risk purpose into software, enabling the user to add their own data to train the model and see the results, the PHRAIS may use contract language to transfer some regulatory responsibilities to the user. This is a significant clarification of how the AI Act would handle software with AI code, although it only covers data governance and not related high-risk AI requirements, especially including transparency (Article 13) or accuracy and robustness (Article 15).

The IMCO-LIBE report also suggests expanding the responsibilities of users (Article 29), specifically that users of high-risk AI systems should be required to meet the human oversight requirements in Article 14, and that the individuals performing this oversight should be adequately trained. In another important change, the IMCO-LIBE report would add to the user requirements an obligation to inform people who interact with a high-risk AI system that they are subject to that AI system's decisions.

The draft report also adds to Article 28, stating that if a pre-existing AI system that is not high risk is placed back on the market without modification but with a new, high-risk purpose, then the entity that does this will qualify as a PHRAIS. This is a minor change that closes a small loophole[16] in the Commission text: regardless of its prior purpose, when an AI system is not modified but is repurposed and placed on the market for a high-risk application, the entity that has repurposed the AI system is considered a PHRAIS.

[16] Under a strict reading of Article 28 of the AI Act proposed by the Commission, it could be construed that taking an AI system already on the market (but not high risk), repurposing it for a high-risk purpose and placing it back on the market could feasibly not qualify the repurposing entity as a PHRAIS. It is unlikely that this was the Commission's intention, and thus we consider this change to be the closing of a small loophole.

6. DISCUSSION AND POLICY RECOMMENDATIONS

The application of the proposed AI Act to GPAI has triggered a lively debate among scholars and between the EU co-legislators, with widely diverging positions being expressed. Recent attempts to include provisions related to GPAI in the text have allegedly been motivated by two emerging needs: first, the need to reflect an emerging trend in AI, which is seeing the growth of more versatile models that are able to perform a variety of tasks, though still within a rather narrow range of possible applications; and second, the need to promote a more balanced allocation of responsibilities along the value chain. The latter motivation was expressed repeatedly during the debate in the European Parliament and also among Member States, and incorporates the belief that original AI developers will often be larger entities such as tech giants. These larger entities can be assumed to possess more resources and greater knowledge compared to the (arguably smaller) companies that will eventually become the providers, as they will place the high-risk AI systems on the market.

Both arguments deserve to be unpacked and discussed in greater depth, as the debate on the AI Act continues to unfold.

As far as the growth of GPAI models is concerned, and as already mentioned, the past few years have seen the commercialisation of a number of what companies such as Google and the Alphabet-subsidiary DeepMind define as 'generalist agents', which innovate on previous AI systems due to their ability to perform more tasks, be efficiently fine-tuned for new tasks, and may even be able to choose the correct task for a given input. While their versatility should not raise excitement as to the imminent arrival of artificial general intelligence, it certainly opens up new possibilities for companies wishing to rely on advanced AI solutions by using existing GPAI and deploying AI systems with a specific purpose.

At the same time, it creates the risk that easily (or, in some cases, freely) available GPAI is deployed for high-risk purposes, generating risks for fundamental rights and safety. This concern is most poignant when the GPAI model is not documented clearly or completely, or when the downstream PHRAIS is unable to use the GPAI model object directly. Either scenario creates more challenges in meeting the AI Act's conformity requirements. As a result, ignoring this emerging type of AI system might detach the AI Act's representation of the AI value chain from the evolving market context. Moreover, excluding GPAI models could potentially distort market incentives, leading companies to build and sell GPAI models that minimise their exposure to regulatory obligations, leaving these responsibilities to downstream applications.

Entities developing generalist agents train these agents for a variety of uses. For example, as noted above, DeepMind has trained Gato on 604 distinct tasks with varying modalities,

observations and action specifications, and OpenAI's text and image generation systems (GPT and DALL-E2) are able to perform a variety of tasks, for which they have been trained by their developers. These systems, rather than general purpose, could be defined as 'multi-purpose' systems under the scope of the AI Act. This, in turn, may place on their developers a number of obligations, which in any case may differ from those attributed to PHRAIS. Obligations may range from the provision of documentation to the performance of internal checks as regards the trustworthiness of the system, and based on the information available at the time of development; and the provision of information as regards the possible uses the system has been trained for, the related level of accuracy to be expected, and the possible mitigating measures to be adopted to ensure the mitigation of possible risks for safety and/or fundamental rights.

These obligations would fall in line with emerging market practice. For example, tech companies providing building blocks to deployers normally share in-depth information on how to handle those building blocks, and the possible mitigating measures to adopt. At the same, they would strengthen the incentives for developers of powerful, proprietary GPAI models to guide deployers in the use of these very impactful tools.

They would also fall in line with established practices in other sectors, such as pharmaceutical and chemical products, where producers' liability is normally exempted in cases of misuse, but producers are increasingly prompted to think about 'reasonably foreseeable misuses'. In fact, the text proposed by the European Parliament introduces a clear reference to reasonably foreseeable uses and misuses at Recital 32, a new Recital 32a, and Recitals 42 to 44. The same IMCO-LIBE report adds to one recital that 'particular attention should be paid to the foreseeable uses and reasonably foreseeable misuses of AI systems with indeterminate uses'.

This leads directly to the second issue, on **how to ensure that responsibility is allocated fairly along the value chain**. Ideally, the efficient way to allocate obligations to the players involved in the different stages of the value chain would be to ensure that responsibility is placed on the so-called cheapest cost avoider (as defined by Guido Calabresi already in 1970). This means the entity that can most effectively identify and assess the riskiness of a given conduct and act to mitigate that risk[17]. The key questions in this respect would be:

- At what stage of the value chain are risks for safety and fundamental rights most easily detected?
- Who is best positioned to take action to mitigate the risk?

These are two distinct questions that deserve greater attention in the AI Act, as well as in the future EU initiative on adapting liability rules to the digital age and artificial

[17] For a comprehensive analysis, see Lior, A. (2020), 'The AI Accident Network: Artificial Intelligence Liability Meets Network Theory', *Tulane Law Review*, Vol. 95, No 5.

intelligence. While a full analysis of these questions goes beyond the scope of this paper, it is worth pointing to the fact that the answers may differ, depending on which type of value chain is adopted among the typologies presented in Section 3. Moreover, a scenario that may present itself is one in which developers have full knowledge of the functioning of their models, but insufficient information to fully grasp the risks that may be generated by the deployed version, and no control over its behaviour and possible impacts when placed on the market; whereas the deployers/users are best positioned to adopt the necessary mitigating measures and monitor the performance of the AI system on the market, but have no inside knowledge of how the model works.

Such a situation calls for information sharing between the developer(s) and the deployers/users of a given high risk AI system; as well as an apportionment of regulatory obligations that depends on a variety of factors, including on who has the best knowledge of the model, and which entity chooses the data that will be fed into the deployed AI system.

Ideally, a PHRAIS that relies on pre-trained models should be entitled to receive sufficient information from the developers, so that it can carry out conformity assessments and post-market surveillance in full control of the system's specifications and behaviour. What remains to be seen is whether market forces will generate this result spontaneously, or whether there will be a role for regulators or the future AI Board to intervene either through soft law (e.g. interpretive guidance on AI contracts, as done in Japan, for instance) or through regulatory requirements for developers, mostly aimed at increasing mutual assistance and information-sharing along the value chain.

Against this backdrop, a different set of issues emerges for GPAI models that are made available as open source. While some open-source AI models are released by companies, others come from teams of developers and academics who may be located all around the world, complicating their regulation. Further, the open-source release of a GPAI model makes it relatively easy for potential deployers to examine its function and gain knowledge of its potential risks, though providing documentation of the training data used and technical steps taken is also very valuable to future developers. In this case, the responsibility to use an open-source GPAI model to deploy a high-risk AI system should fall entirely onto the deployer/user, which then becomes the sole PHRAIS.

The same may not occur, however, when the GPAI is not entirely open nor easy to train or adapt, as can happen whenever the system requires massive hardware infrastructure, or is nested in a proprietary hardware structure, without which its functionality is drastically reduced. Here, too, the variety of hybrid practices that are emerging in the AI world between the extreme cases of 'purely proprietary' and 'purely open' are such that future research should probably venture into an ad hoc typology of the type presented in Section 3 above. And the solutions available to the policymakers should not be seen in a binary way (that either the developer or the deployer should carry out the conformity

assessment). In some cases, one could also imagine a system whereby smaller firms rely on third-party audits of large-scale open-source systems in order to access the market.

6.1 SELECTED POLICY RECOMMENDATIONS FOR THE FUTURE AI ACT

Recommendation 1

Discourage API access for GPAI use in high-risk AI systems

Accessing a GPAI model through an API dramatically limits the understanding and examination of the GPAI model by outside developers and users. For example, an outside developer is unlikely to be able to effectively use many methods to interrogate the GPAI model, such as red teaming, adversarial training, model pruning, generating important or context-specific evaluation metrics, or altering the model in any way, with the possible exception of fine-tuning. Further, because APIs prorate cost based on the amount of data, this type of access discourages the data-intensive process of extensive and routine testing of GPAI models. In most scenarios, it will be very difficult, and at times functionally impossible, for the developer to use the GPAI model through an API for a high-risk purpose and meet the AI Act's requirements. This remains true even with the requirements on GPAI providers proposed by the French presidency of the Council of the EU. While these requirements may create some public transparency and enable more responsible use of GPAI models, they are highly unlikely to be sufficiently specific and thorough to broadly enable safe and responsible use of downstream high-risk AI systems.

While the AI Act should not ban API access for the use of GPAI models in high-risk applications, it could clarify that high-risk applications built on top of restricted AI system access (Type 3) present significant regulatory exposure to the outside developer (which is also the PHRAIS) in terms of meeting the AI Act's requirements. This leaves EU policymakers with at least two possible solutions, to be incorporated into the AI Act:

- *Requiring GPAI model providers engage in active collaboration with downstream developers*. This collaboration, however, should not be limited to the initial, *ex ante* conformity assessment obligation, but should extend to the post-market surveillance phase, given the need to monitor the behaviour of the model and intervene where possible to mitigate emerging risks.
- *Encouraging the sale and transfer of the GPAI model object, along with extensive technical documentation* of the underlying data and GPAI development process, as a best practice. Interpretive guidance (possibly by the future AI Board) could also specify the content that should be featured in the technical documentation, so that downstream entities can request this information when purchasing the GPAI model to develop their applications.

Recommendation 2

Introduce soft commitments for GPAI model providers to strengthen legal certainty and reduce transaction costs

The proposed changes on GPAI proposed under the French presidency of the Council of the EU would require GPAI model providers to go through a conformity assessment process. However, the use of GPAI models by other developers is likely to be so wide and varied that a GPAI provider cannot reasonably predict the many diverse applications and contexts for which these systems will be used. Imposing a separate conformity assessment for every reasonably foreseeable use may be too burdensome on the original developer, even impossible, since the contours of the deployment phase will not be known at the time of assessment. Therefore, even a highly responsible and well-resourced GPAI model provider will be functionally unable to envision and properly document the GPAI model for all possible downstream use cases, let alone monitor its behaviour as the system is placed on the market. Recognising this, the Council proposal should clarify that downstream PHRAIS should not assume that using a GPAI model that has gone through the conformity assessment process guarantees, or even necessarily contributes to, the conformity of downstream AI systems.

Possible safeguards that could be built into the AI Act to reduce the burden on downstream developers include the following:

- *Requiring that GPAI model providers indicate the possible uses* of the system that were envisaged at the time of system design and training, or even specify which uses they believe should be 'allowed', and which they do not endorse or consider to be safe. Documentation on the GPAI model could also go as far as indicating the recommended mitigating measures for a variety of possible uses of the system, so that downstream developers are helped in their decision on whether and how to use the system.
- *Establishing a voluntary code of conduct for GPAI models*. Similar to what is envisaged by Title IX of the AI Act for non-high-risk AI applications, the AI Act could seek to reduce transaction costs between GPAI model providers and downstream developers by introducing the possibility for the former to voluntarily adopt a code of conduct. According to this code, GPAI model providers would abide by the principles and requirements of trustworthy AI, and explain how they can be complied with when using the system. Since GPAI model providers would not normally be in the position to anticipate all risks (as many risks depend on the concrete purpose for which the system is deployed and used), the code of conduct would need to be made proportionate to the information that the GPAI model provider can be assumed to

have at the time of making the system available to downstream developers. In this respect, the voluntary system would work proportionately, along the lines of similar frameworks being developed in other countries (e.g. the NIST Risk Management Framework in the United States).

Recommendation 3

Discourage software with AI code for high-risk AI systems

The business model of software with AI code, in which a provider develops software with code to train an AI model (but no data), and sells the software to a user who adds their own training data, is particularly challenging for the AI Act. Because the user is unable to change, or even potentially precisely view, the code or model object of the AI system, and the provider may never see the data, it may be functionally impossible for either entity to meet the requirements of the AI Act, or to develop and use the high-risk AI system responsibly.

Although the IMCO-LIBE proposals attempt to balance requirements better (stating that the user can take on the data governance requirements in this situation), this does not resolve the underlying challenge of meeting these requirements for either entity—both missing one crucial part of the picture. In light of this, the AI Act should clarify that this business model, or any analogous situation in which no single entity can evaluate the whole of an AI system, is more likely to lead to non-compliance under the high-risk requirements of the AI Act.

One potential outcome for software with AI code may be that it comes with a far greater range of options for customisation of the AI system. This would be a radical change in comparison to much of the software developed today but, paired with the IMCO-LIBE proposals for requiring users to provide competent human oversight, could be effective. Alternatively, moving away from this business model to other options, such as AI system contracting, would also be a welcome change, easing the apportionment of regulatory responsibilities and ensuring that the deploying entity has the potential to fully observe and evaluate the AI system. Broadly, as off-the-shelf software is often cheaper and used more frequently by smaller businesses, it is essential for the EU to think more critically about this part of the AI value chain.

Recommendation 4

Exempt open-source AI models from all AI Act requirements

The AI Act should explicitly exempt the placing of an AI system online as free and open-source software (i.e. making the entire model object available for download under an open-source licence, not just available without cost via API access). The deployment and use of these AI systems for any covered non-personal purposes would still be regulated under the AI Act, thus maintaining the same level of consumer protection and safeguards for human rights and safety. However, this exemption would enable the collective development, improvement and public evaluation of AI systems, which has become a key outcome of open-source software. Including open-source AI systems under the AI Act requirements will likely result in a barrier to both scientific and research advancement, as well as reducing public understanding and rigorous scrutiny of commonly used methods and models. Even with the carveouts in the French presidency's proposal (as discussed above), researchers and developers may be unsure if their AI models and associated licences qualify for these exemptions, creating a chilling effect on open science with few, if any, benefits to consumer protection[18].

Amendments proposed by the European Parliament's Committee on Legal Affairs would explicitly exempt the open-sourcing of AI systems. These suggested changes are appropriately careful not to exempt developing an open-source AI system then using it for a commercial purpose, such as contracting to help a company deploy it.

[18] For a more in-depth discussion, see Engler, A. (2022), 'The EU's attempt to regulate open-source AI is counterproductive', Brookings TechTank, at https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/.

Recommendation 5

Clarify PHRAIS ambiguities in common AI value chain scenarios

In several of the common business models discussed in the AI value chain typology, there is considerable ambiguity concerning the entity that will typically be the PHRAIS. This is especially true in the case of AI system contracting and fine-tuning, although it also may apply to software with AI code, learning AI systems and AI model integration. The AI Act could benefit from a more thorough discussion and consideration of the AI value chain, including explicit reference to these scenarios or others, and the expected or typical result in terms of PHRAIS responsibilities. In addition, once the AI Board is established, this institution will be perfectly placed to issue interpretive guidance on how the AI Act applies to the ever-evolving typology of AI value chains, and what the resulting implications in terms of accountability of all the entities involved will be under the different scenarios.

Recommendation 6

Incorporate European Parliament proposal to add new user requirements

The IMCO-LIBE draft report suggest a few important new requirements on users of high-risk AI systems, most notably adding new transparency responsibilities and expanding the human oversight requirements. As for transparency, in many cases, users who purchase an AI system may have more direct control over the environment the AI system is incorporated into, and should therefore be held responsible for 1) informing natural persons such as end consumers that a high-risk AI system is making decisions that affect them; and 2) adopting measures that may mitigate the risks, such as securing meaningful human oversight or providing end recipients with a right to redress. The provider (as currently defined under the AI Act) would not always, perhaps not even often, be well placed to inform the public. Hence, the IMCO-LIBE proposal to place the disclosure requirement on the user is appropriate.

Similarly, the IMCO-LIBE proposal requires users of high-risk AI systems to ensure that human oversight is undertaken by competent and properly trained employees. In many of the business cases discussed, there may be a significant knowledge gap between the AI system developer and the user, such that it should not automatically be expected for the user to have a precise or effective understanding of the AI system. Therefore, requiring users to provide not just human oversight, but sufficiently informed human oversight, is a beneficial change where high-risk AI systems are concerned.

CEPS
PLACE DU CONGRES 1
B-1000 BRUSSELS