Thank you for reading! At the time this paper was submitted for the WeRobot 2023 conference, the trialogue between the EU Commission, the Council of the EU, and the European Parliament has started. The three institutions will have to agree on a common version of the AI Act, before submitting the text for a final vote. The current version of this article cites the latest EU Parliament version of the proposed regulation (from June 2023), which might evolve in the upcoming weeks. The article will be updated consequently after the conference and before submission to a journal. This article is a work in progress, **if you have any feedback or any citation we're missing, please do not be shy in telling us**. You **can cite** this paper as a We Robot 2023 paper presentation.

General Purpose AI Systems in the AI Act: trying to fit a square peg into a round hole

Claire Boine[1]
David Rolnick[2]

ABSTRACT

[1] PhD Candidate, Faculty of Law, University of Ottawa; Research Associate, Artificial and Natural Intelligence Toulouse Institute.

[2] Assistant Professor, Faculty of Computer Science, McGill University

Introduction

At different points in time, authors of the legal doctrine have questioned whether the law could properly regulate new technologies, since it is inherently slow and conservative.[3] Technology law scholars know that their field consists in watching a game of whack-a-mole unfold. It has been the case with misinformation on social media and online data privacy, both issues having become significant before any regulation was even considered. These past few years, we have witnessed an even more dramatic version of this problem in the field of artificial intelligence (AI). While AI had been present for decades, the increase in big data and computing power led to significant progress in deep learning, which made AI more capable, more accessible, and more affordable. Advances also came from breakthroughs in areas such as attention mechanisms,[4] adversarial training, and integration of deep learning with pre-existing reinforcement learning techniques. AI systems were introduced in many areas of our lives, most of the time unknowingly. The most visible ones are the recommendation algorithms used to influence consumer behavior. They select which movies we see on Netflix, which posts we see on Facebook, X, and Instagram, and which

[3] Daniel Malan, *Technology Is Changing Faster than Regulators Can Keep up - Here's How to Close the Gap*, WORLD ECONOMIC FORUM (Jun. 21, 2018), https://www.weforum.org/agenda/2018/06/law-too-slow-for-new-tech-how-keep-up/.

[4] "Attention mechanism: a mechanism used in a neural network that indicates the importance of a particular word or part of a word. Attention compresses the amount of information a model needs to predict the next token/word. A typical attention mechanism might consist of a weighted sum over a set of inputs, where the weight for each input is computed by another part of the neural network." Google. "Machine Learning Glossary." Google for Developers. https://developers.google.com/machine-learning/glossary.

products should be advertised to us, generally something we have been talking or browsing about. AI systems can also be used to determine our credit access, filter out our spam emails, optimize our natural gas delivery,[5] help decide which crops are grown around us,[6] and chat with us when we order a coffee at Starbucks or a car with Lyft.[7]

As a result of the fast spread of AI into our lives, in April 2021, the EU Commission proposed what was then a bold regulation to impose safety requirements onto AI systems considered as potentially posing a "a high risk of harm to the health and safety or the fundamental rights of persons."[8] The EU Commission specifically targets systems considered "high risk" as defined by the context of use of that system. For instance, the use of algorithms in areas such as immigration, critical infrastructure, or education is considered high risk.[9]

The approach of the AI Act had already been laid out in the White Paper published a year before. The latter explains the risk-based approach as follows:

> "The Commission is of the opinion that a given AI application should generally be considered high-risk in light of what is at stake, considering whether both the sector and the intended use involve significant risks, in particular from the viewpoint of protection of safety, consumer rights and fundamental rights."[10]

Published in February 2020, the White Paper had itself been heavily influenced by the academic debates that had taken place the previous years and had highlighted only a specific subtype of automated systems that were spreading at that time. As a result, when the AI Act came out, it was not adapted to the latest developments in AI.

[5] Phil Laplante & Ben Amaba, *Artificial Intelligence in Critical Infrastructure Systems*, 54 COMPUTER 14 (2021).
[6] Dmytro, *Farm Management Software to Boost Production & Profitability*, INTELLIAS (2020), https://intellias.com/unified-farm-management-system-to-boost-production-and-profitability/.
[7] Larry Kim, *10 Real Examples How Brands Are Using Chatbot for Customer Service*, MISSION.ORG (Apr. 23, 2018), https://medium.com/the-mission/10-real-examples-how-brands-are-using-chatbot-for-customer-service-4fbb5e4617f3.
[8] European Commission, *Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence ("AI Act")*, (2021).
[9] European Commission, *Annexes to the Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence ("AI Act")*, (2021).
[10] EUROPEAN COMMISSION, *On Artificial Intelligence - A European Approach to Excellence and Trust*, (2020).

As an example of its shortcomings, the text of the proposed regulation relies heavily on the notion of the *intended purpose* of AI systems. The intended purpose of a system influences whether it is considered high-risk,[11] which in turn determines which safety requirements must be in place. Yet, an increasing number of systems without an intended purpose had been developed in the years before the release of the AI Act. For instance, GPT-3, an AI system that was one of the best language models at the time and did not have an intended purpose, came out in June 2020. The fact that the AI Act kept the approach laid out in the White Paper is an illustration of path dependence in the law, even in non-common law jurisdictions. Path dependence "means that an outcome or decision is shaped in specific and systematic ways by the historical path leading to it" and results in part from *stare decisis* in common law jurisdictions.[12]

After the wide release of ChatGPT, made available for free to consumers in November 2022, European policymakers realized that the AI Act presented a significant gap. Already late in the process, they decided to add provisions on *General Purpose AI systems* (GPAIS) and *foundation models*. Given that the text was not drafted in a technology-agnostic manner and entirely built around end uses and intended purposes, adding these provisions is like trying to fit a square peg into a round hole.

In the first section of our paper, we will show that the AI Act was influenced by a conception of AI systems as non-autonomous statistical software and of potential harms as stemming mostly from datasets. We will demonstrate that the notion of intended purpose, which draw from product safety, was suited for that paradigm. In the second section, we will discuss AI systems that do not have an intended purpose, such as GPAIS, foundation models, and others. We will clarify what types of harms they can cause and why the AI Act in its initial version is not prepared to address them. In the third section, we will present the provisions proposed by the EU Parliament to regulate

[11] Article 7.2.a of the AI Act.

[12] Oona Hathaway, *Path Dependence in the Law: The Course and Pattern of Legal Change in a Common Law System*, SSRN ELECTRON. J. (2003), https://www.academia.edu/27017830/Path_Dependence_in_the_Law_The_Course_and_Pattern_of_Legal_Change_in_a_Common_Law_System.

these models and propose additional policy recommendations for adapting the AI Act to GPAIS and future AI systems.

I. From artificially intelligent agents to software

a. The evolving definition of artificial intelligence

1. AI defined by its capabilities

There is no commonly agreed upon definition of artificial intelligence. In fact, the very definition of AI set forth by the AI Act has evolved in the different iterations of the text. While the lack of a single definition is not problematic for academic purposes, regulation requires precision so what is and is not in the preview of the law is clearly established. This turned the definition of AI in the AI Act into a political issue. Some industry members pushed for a less inclusive definition of AI in the proposed regulation so that the systems they produce, or use would not be subject to safety requirements, while consumer groups wanted the definition to be as broad as possible to include more systems.[13]

In addition to these stakeholders' interests, other factors have influenced the definition, including humans' perception of intelligence. What is perceived as artificial intelligence has evolved over time, and influenced the behavior and beliefs of those who interact with such technology. On the one hand, it was shown that many people view artificial intelligence as something that is not possible to grasp or achieve. These individuals would tend to define AI in terms of capabilities machines do not have yet ("an intelligent machine will surely be capable of doing x or y"). However, as soon as an AI system acquires one of these capabilities (e.g., beating a human at chess, driving, using natural language), they would shift their mental model of intelligence and conclude

[13] Yannick Meneceur, *Le Piège de La Définition Juridique de l'intelligence Artificielle*, LINKEDIN, 2021, https://www.linkedin.com/pulse/le-pi%C3%A8ge-de-la-d%C3%A9finition-juridique-lintelligence-yannick-meneceur?trk=public_profile_article_view.

that these capabilities did not require intelligence after all.[14] This type of dynamic is rooted in beliefs such as that intelligence is a fundamentally human attribute, or that machine intelligence requires machines to do what humans do in the same way as humans would do them. These views can be summarized in the statement that AI systems are just machines after all. Simultaneously, numerous individuals suffer from automation bias.[15] These individuals assume that machine outputs are scientific, and therefore accurate. They tend to attribute too much intelligence to any automated system and overly rely on their outputs. This trend is sometimes related to a belief that technology will solve most problems, and that while humans make mistakes, machines don't. In fact, new technologies are often presented as a way to limit human error. For instance, at a 2021 roundtable on the use of AI in critical infrastructures in the US, one of the experts asserted that "even a trained human can be inefficient or make mistakes due to various psychological conditions; this is where AI can play an important role, eliminate such mistakes, and be more efficient than humans."[16]

Finally, the definition of AI has been influenced by trends, especially which systems were most publicized at different points in time. Before the late 2010's, AI systems were commonly thought of as agents, i.e. as entities capable of conceiving a plan and carrying it out. This was consistent with the collective imagery of domestic robots and chess-playing AI systems. In 2014, the Pew Research Center surveyed 1,896 experts on robots, self-driving cars, and "intelligent digital **agents**" (emphasis added).[17] These experts expressed concerns about job displacement, but at the same time thought that AI systems might in part free people from work. This seems to indicate that they viewed AI systems as potentially at least as competent as humans, and not necessarily requiring humans in the loop. Autonomy and agency were perceived as an intrinsic part of what makes an AI system intelligent.

[14] STUART ARMSTRONG, SMARTER THAN US: THE RISE OF MACHINE INTELLIGENCE (2014).
[15] Saar Alon-Barkat & Madalina Busuioc, *Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice*, 33 J. PUBLIC ADM. RES. THEORY 153 (2023).
[16] Laplante and Amaba, *supra* note 5.
[17] Aaron Smith, *AI, Robotics, and the Future of Jobs*, PEW RESEARCH CENTER: INTERNET, SCIENCE & TECH (Aug. 6, 2014), https://www.pewresearch.org/internet/2014/08/06/future-of-jobs/.

This view was consistent with the definition of AI proposed by the EU Commission in April 2018 and displayed in Table 1. "Artificial intelligence (AI) refers to systems that display **intelligent behaviour** by **analysing their environment and taking actions** – with some degree of **autonomy** – to achieve specific goals" (emphasis added).

Table 1. Definitions of Artificial Intelligence proposed by EU lawmakers

Source	Definition
European Commission, April 2018	Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).
AI Act as of April 2021 (text from the EU Commission)	Software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with; Techniques and approaches listed in Annex I: (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; (c) Statistical approaches, Bayesian estimation, search and optimization methods.
AI Act as of December 2022 (text from the Council of the EU)	A system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts.
AI Act as of June 2023 (as adopted by the EU Parliament)	A machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments.

2. The rise of statistical tools

This changed in the late 2010's, when AI systems started being equated with "algorithms." An algorithm is a set of instructions to be followed, whether they are for humans or machines. For instance, a food recipe is an algorithm, generally for humans to follow. One of the machine algorithms that received the most publicity in the past few years was COMPAS, after ProPublica published a 2016 study showing that it was biased against Black defendants.[18] COMPAS was a software sold by Northpointe to dozens of administrations and meant to predict the risk of criminal recidivism. It mainly consists in a statistical regression.

A regression is a mathematical tool used to analyze trends and make predictions. For instance, a linear regression could take the form of an equation with two main variables, calculated from twenty data points. Imagine a class of 20 students. Their teacher wants to predict the students' weights based on their heights. They make a plot with weight as the y-axis and height as the x-axis. It turns out that the correlation seems linear, and the teacher can draw a line that minimizes the sum of the squared vertical distances between the line and the points. This can be done entirely manually. Now suppose the teacher learns that a 21st student is going to join the class soon and they know that student's height. They can make a prediction as to their weight using that line and looking at which y value corresponds to the student's height. Northpoint applied the same methods in computing recidivism scores, expect they used far more than 20 data points, that theirs was a nonlinear regression, and that they used six variables for the general recidivism score and five variables for the violent one and the screening one. Using the COMPAS software, a probation officer could enter the defendant's age, age-at-first-arrest, number of prior arrests, employment status, and the number of prior parole revocations and the algorithm would output a screening recidivism risk score. It is a stretch to think of the output in terms of individual probability of recidivism. If anything, what a COMPAS score tells us is something such as "in a group of 100 individuals of *x* age-at-first arrest and *y* prior convictions, *z* %will reoffend." However, some judges and probation officers who used the COMPAS software without understanding how it

[18] Julia Angwin et al., *Machine Bias*, PROPUBLICA, May 23, 2016.

worked assumed that it actually predicted whether someone would reoffend, and that it was scientific and therefore accurate. This resulted in Black defendants being discriminated against in the justice system given that they were receiving on average a higher false positive rate of recidivism compared with similarly situated white defendants. Interestingly, Northpointe did not describe COMPAS as artificial intelligence,[19] but many journalists and policymakers would make that leap.[20]

In 2018, another scandal involved large-scale statistics and algorithms: Cambridge Analytica. Cambridge Analytica involved different manipulative techniques to influence the outcome of elections in different countries, including a pro-Trump campaign microtargeting non-registered voters in four target states based on their psychological traits inferred from their Facebook activity in 2016. This was also based on statistical analysis. The company created a large matrix of correlation coefficients between Facebook likes and psychological traits such as neuroticism, and then designed different messages for people with different psychological traits. In the past decade, the type of statistics used by Cambridge Analytica became prevalent for consumer advertising, and more and more stories broke in the news.

These trends influenced the way AI was perceived. First, the type of statistics used by Northpointe and Cambridge Analytica was not new, and most people would not have considered them artificial intelligence. In fact, Northpointe and Cambridge Analytica never labeled their software as AI. The leaked Cambridge Analytica documents show that the company used the term "proprietary algorithm."[21] What led to the sudden spread of these old methods was the novel availability of large datasets that made it possible to predict new variables. However, these tools were increasingly presented as artificial intelligence in the media and policy conversations. The line between statistics and machine learning is blurry. For instance, a regression can be calculated by

[19] For instance, the term does not appear once in this guide to practitioners that
[20] Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, THE NEW YORK TIMES, May 1, 2017, https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html. In this article on COMPAS for instance, the journalist seems to be conflating algorithms and AI.
[21] CAMBRIDGE ANALYTICA, *Internal Documents Leaked by Whistleblower Bettany Kaiser*, https://ia803204.us.archive.org/35/items/ca-docs-with-redactions-sept-23-2020-4pm/FINAL%20Cambridge%20Analytica%20Select%202016%20Campaign%20Related%20Documents%20w%20Redactions_.pdf.

hand, done using an Excel spreadsheet, or conducted using a Python library. It is not agentic nor autonomous.

The field of machine learning improved significantly in the 2010's, especially deep learning, a technique to extract high-level, abstract features from raw data by creating representations that are expressed in terms of other, simpler representations.[22] For instance, "when analyzing an image of a car, the factors of variation include the position of the car, its color, and the angle and brightness of the sun," and a deep learning algorithm trained on millions of pictures of cars will learn high-level abstract features present in most cars to then tell cars apart from other objects.[23] Deep learning algorithms are often presented as black boxes, because to this day, we cannot reverse engineer them. This term was then used by Frank Pasquale to denounce the secrecy of the use of algorithms in most areas of our lives.[24] This led to a misunderstanding that most algorithms, including the types used in COMPAS, would be opaque and/or would use deep learning. The truth is that these algorithms are often simple calculations, most of which do not require deep learning. The COMPAS scores use 5 to 6 variables. Cambridge Analytica used thousands of variables but very simple methods. This led experts to publish articles and books combating automation bias, and explaining to the public that algorithms are not intelligent nor autonomous, and that humans are behind them. Authors started contesting the term AI. The book *Artificial Unintelligence* by Meredith Broussard is one example among many.[25]

This new perception of these algorithms influenced the definition of AI. From the agency paradigm, there was a shift toward software. AI systems became perceived mostly as non-autonomous decision-making tools. In fact, the AI Act published by the EU Commission in April 2021 defined AI systems as "**software** that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of **human-defined objectives**, generate outputs such as content, **predictions, recommendations, or decisions** influencing the

[22] IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING (2016).
[23] *Id.*
[24] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (Reprint edition ed. 2016).
[25] MEREDITH BROUSSARD, ARTIFICIAL UNINTELLIGENCE: HOW COMPUTERS MISUNDERSTAND THE WORLD (2018).

environments they interact with" (emphasis added). The list of approaches from Annex I is available in Table I. Along with a certain conception of AI came a certain idea of what harms could stem from this technology.

b. A certain conception of harm

1. The relation between risk and context

Conflating artificial intelligence with the statistical tools described in the previous section resulted in three misconceptions related to the type of harms they can create. In a later section of these paper, we will show that these beliefs are misconceptions because they do not apply to GPAIS. Therefore, it is a mistake to extend them to all AI systems. Some of them might also be inaccurate in the context of narrow AI systems.

The first misconception is that whether an AI system can cause harm or not exclusively depends on whether it is deployed in a high-stake context. Table 2 shows the list of systems considered high-risk by the Commission and listed in Annex III of the original version of the AI Act.[26] The systems are defined by their *intended purpose*. Most of them have the purpose of producing an output or a prediction to help determine someone's access to a critical service or to the exercise of their rights (employment, education, immigration, etc.). The level of risks was equated with how sensitive the context of deployment is, which in turns determine whether AI operators must comply with certain safety requirements or not.

Table 2. Systems considered high-risk in the initial version of the AI Act

Area	Intended purpose	Algorithm assisting in determining someone's access to a critical service or exercising of their right?

[26] European Commission, *supra* note 9.

Biometric identification	‘Real-time’ and ‘post’ remote biometric identification of persons.	It depends.
Critical infrastructures	Safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.	It depends.
Education and vocational training	Determining access to educational and vocational training institutions.	Yes
	Assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.	Yes
Employment	Recruitment or selection of persons (e.g., screening or filtering applications).	Yes
	Making decisions on promotion and termination of contracts, for task allocation and for monitoring and evaluating performance and behavior.	Yes
Essential private and public services	Evaluating the eligibility of persons for public assistance benefits and services.	Yes
	Evaluating the creditworthiness of persons or establishing their credit score.	Yes
	AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services.	Yes
Law enforcement	Making individual risk assessments of persons to assess their risk for offending or reoffending.	Yes
	Used as polygraphs or to detect the emotional state of a person.	Yes
	Detecting deep fakes.	It depends.
	To evaluate the reliability of evidence during investigation or prosecution of criminal offences.	Yes

	Predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling or assessing personality traits and characteristics or past criminal behavior of persons or groups.	Yes
	Profiling of persons during detection, investigation or prosecution of criminal offences.	Yes
	For crime analytics, to search complex related and unrelated large data sets available in different data sources or in different data formats in order to identify unknown patterns or discover hidden relationships in the data.	Yes
Migration, asylum and border control	Used as polygraphs or to detect the emotional state of a person.	Yes
	To assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a person.	Yes
	For the verification of the authenticity of travel documents.	Yes
	For the examination of applications for asylum, visa and residence permits.	Yes
Administration of justice and democratic processes	To assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts.	Yes

A second misconception is that AI systems exclusively cause individual harm. Even in the case of AI systems that are not GPAIS, this idea was debunked. For instance, Nathalie Smuha argues that dynamics such as election interference, hate-mongering, or societal polarization, caused by systems such as the ones used by Cambridge Analytica, constitute societal harms.[27]

[27] NATHALIE A. SMUHA, *Beyond the Individual: Governing AI's Societal Harm*, (2021), https://papers.ssrn.com/abstract=3941956.

2. The relation between risk and data

A third misconception is that, because AI systems are statistical tools, the harms they can cause necessarily comes from their dataset. Before the AI Act was released, the public debate on AI harm was mostly captured by self-driving cars and algorithmic bias. Even though physical harm can result from non-physical objects, it is easier to imagine physical harm created by faulty AI systems embedded in physical objects, such as self-driving cars, toys, or critical infrastructure. In addition, self-driving cars have always been perceived as futuristic and have captured the public imagination for decades.

In terms of immaterial harm, most of the debate was on bias creating harm either directly (for instance the harmful classification of Black people as gorillas in Google photo[28]) or through loss of chance (such as in the case of the Amazon resume screening tool that filtered out resumes containing the word woman[29]). For many people, the realization that these systems could be biased was at odds with their perception of these tools as purely mathematical and therefore accurate. When the EU Commission published its White Paper on artificial intelligence in February 2020, one of the most popularized issues was racial bias in statistical tools used in criminal sentencing. A google scholar search for manuscripts published between 2017 and 2020 and containing the words "COMPAS" and "propublica" and "bias" yields 2,190 results. This largely influenced the EU Commission. In fact, bias in automated recidivism prediction is one of the only concrete examples of AI harm exposed in the White Paper. The second example laid out in the White Paper is of racial bias in facial recognition, for which the Commission cites the work of Joy Buolamwini and Timnit Gebru, which had also received a significant amount of attention at the time. While it is important that these issues be taken seriously, it is equally important to not be under the false impression that data governance measures are enough to make AI systems safe.

[28] Google apologises for Photos app's racist blunder, BBC NEWS, Jul. 1, 2015, https://www.bbc.com/news/technology-33347866.
[29] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women*, REUTERS, Oct. 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

First, it is worth clarifying that most of the systems described by the EU Commission do not make the ultimate decision about someone. These systems have often been presented as decision-making tools in the media. While the systems themselves can be faulty and biased, they are integrated into a human decision-making process. In many cases, a system only produces a score or a probability, meant to support a human decision. As such, a significant portion of the harm can come from the way the human uses and interacts with the system, regardless of how it performs. If the human attributes too much credit to the system due to automation bias, or that the human does not know how to interpret the system's output, significant harm can result. It is thus the sociotechnical system that needs to be regulated, and not exclusively the dataset. This issue is in part addressed in the AI Act by Article 14 on human oversight, which aims at enabling individuals to "remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons."[30]

Second, while errors at different stages of the data pipeline can lead to at least six different types of bias (historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, deployment bias),[31] creating a high potential for harm, AI systems can be harmful in yet many other ways. Therefore, while it is necessary for the AI Act to include requirements on data validation and data transparency, these provisions are not sufficient to make AI systems safe.

c. The notion of *intended purpose* in EU law and in the AI Act

1. The notion of intended purpose in EU law

The notion of intended purpose is key to EU product safety law and consumer protection. In the EU, each country has its own product safety laws, which have been harmonized by the 2001 Directive on general product safety. This framework will soon be replaced by the General Product Safety Regulation starting in December 2024. The goal of product safety law is to achieve a high

[30] Article 14 of the AI Act.

[31] Harini Suresh & John V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning*, ARXIV190110002 CS STAT (2020), http://arxiv.org/abs/1901.10002.

level of consumer protection by imposing ex-ante safety requirements on manufacturers, providers, importers, and distributors of products made available in the EU. Another key building block of consumer law is the Unfair Commercial Practices Directive (UCPD) which prohibits unfair, misleading, and aggressive commercial practices.

In product safety, the purpose of a product matters as it determines the corresponding safety requirements. As such, the same product will have different requirements to fulfill based on its intended use. As an example, geotextiles are permeable synthetic fabrics used to help reinforce or drain areas. When geotextiles are meant to be used for roads, the manufacturers must comply with norm EN 15382:2018. However, when geotextiles are meant to be used in a dam, producers must follow norm EN 13361:2018. These standards were built in the context of European *Regulation 305/2011 laying down harmonised conditions for the marketing of construction products*.

The intended purpose of a product also matters in European consumer law because commercial transactions are only valid if the product can do what is expected of it. The UCPD states that a commercial practice is misleading if it causes someone to make a transactional decision that they would not have taken otherwise, in relation to one or more of multiple elements including the product's "fitness for purpose." For instance, a French court cancelled the sale of a pony that had taken place six months earlier because the animal was two centimeters taller than advertised at the time of purchase. While most pony sales would not be nullified for that reason, this specific pony had been sold for the purpose of participating in competitions, and the pony's participation required the animal to be 2 centimeters shorter. The court deemed the pony unfit for purpose.

Finally, the intended purpose of a product determines whether the product is considered as performant. For instance, European regulation on in vitro diagnostic medical devices 2017/746 states that "'performance of a device' means the ability of a device to achieve its intended purpose as claimed by the manufacturer" (article 2(39)). The French *Cour de Cassation* even used the notion of "fitness for purpose" as a synonym of product quality.[32]

[32] "En omettant de distinguer **les qualités de la chose -ou son aptitude à l'usage auquel elle était destinée**- de ses conditions de mise en service, la cour d'appel, qui a reconnu la parfaite fiabilité du matériel vendu, n'a pas mis la

Beyond product safety, the notion of purpose is also significant to European data privacy law. The General Data Protection Regulation establishes that personal data can only be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" (article 5.1(b)). Personal data also must be accurate for the purpose for which it is collected (article 5.1(d)), and the collection should only involve the minimum amount necessary to achieve that purpose (article 5.1(c)).

2. The notion of intended purpose the AI Act

The European Union published its first draft of the AI Act in April 2021. The legal basis of the text is Article 114 of the Treaty on the Functioning of the European Union (TFEU) on the proper functioning of the internal market. This means that the EU has competence over AI product safety to make sure that rules are consistent across the EU to promote the liberal circulation of goods. The AI Act is entirely inspired by EU product safety, but as applied to AI systems.[33]

It is thus not surprising that the AI Act bases much of its content on the intended purpose of an AI system. According to Article 3, "'intended purpose' means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation."

The intended purpose of a system influences whether it is considered high-risk (art. 7.2.a), making it subject to specific safety requirements. In addition, within high-risks systems, the intended purpose also determines the content of the requirements. For instance, testing of AI systems depend on their intended purpose. Article 9.7 of the AI Act states that "testing procedures shall be suitable to achieve the intended purpose of the AI system and do not need to go beyond what is

Cour de Cassation en mesure d'exercer son contrôle" Cass. Com. 03.10.1989 n° 87-18.581 inédit https://www.legifrance.gouv.fr/juri/id/JURITEXT000007091328

[33] Marco Almada & Nicolas Petit, *The EU AI Act: Between Product Safety and Fundamental Rights*, (2022), https://papers.ssrn.com/abstract=4308072; MICHAEL VEALE & FREDERIK ZUIDERVEEN BORGESIUS, *Demystifying the Draft EU Artificial Intelligence Act*, (2021), https://osf.io/38p5f.

necessary to achieve that purpose." This is like the principle of data minimization in the GDPR, but this time the minimization aims to avoid burdening AI operators or requiring them to disclose unnecessary trade secrets. For instance, it might be enough to check that a resume-triaging algorithm is not biased against protected categories of the population.

The AI Act also requires testing to be made against "preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system." As an example, certain contexts of use (e.g., employment or immigration) may require a higher level of accuracy than other contexts (e.g., song recognition application). In fact, Article 15 stipulates that "high-risk AI systems shall be designed and developed in such a way that they achieve, **in the light of their intended purpose**, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle" (emphasis added). The intended purpose of the AI system even determines the duration of record keeping as the logs shall be kept for a period that is appropriate in the light of the intended purpose of high-risk AI system (Article 13).

The AI Act is thus clearly built on product safety law, which bases safety requirements on the intended purpose of products. This is consistent with statistical software build for narrow purposes, especially when the harm they could make is individual and based on data robustness or the absence of defect.

II. From software to AI agents

a. The rise of models without an intended purpose

1. Foundation models and other GPAIS

In his paper on the regulation of artificial intelligence (AI), Matthew U. Scherer wrote that: "the potential for rapid changes in the direction and scope of AI research may impair an agency's ability to act ex ante; an agency whose staff is drawn from experts on the current generation of AI technology may not have expertise necessary to make informed decisions regarding future generations of AI technology."[34] As discussed in the previous section, the AI Act was influenced by public discussions about a certain type of statistical software with narrow purposes and no autonomy, and a generation of experts warning against problematic datasets. However, a new type of AI systems whose complexity was by far exceeding the statistical tools described earlier was already on the market by the time the AI Act came out. These systems are General Purpose AI Systems (GPAIS), or systems without a unique predetermined purpose. GPAIS include foundation models, as well as transfer and meta learning systems, which can be adapted to undertake new tasks with minimal effort. Box 1 presents relevant AI definitions.

Box 1 clarifies different key definitions in AI.

Box 1 – AI definitions

Algorithm: A set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem.[35]

Foundation model: AI system trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks.[36]

Generative AI: AI systems that generate outputs more complex than a number, label, or recommendation (e.g., text, audio, video, images).

General Purpose AI System (GPAIS): AI system that can accomplish or be adapted to accomplish a range of distinct tasks, potentially including some it was not intentionally and specifically trained for.

Multi-modal AI: an AI system where the input or output includes more than one modality (e.g., images, video, audio, text, time-series).

Transfer and meta-learning systems: systems designed to acquire a new capability with minimal additional learning.

[34] Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, (2015), https://papers.ssrn.com/abstract=2609777.

[35] algorithm, Cambridge Dictionary of English (2023), https://dictionary.cambridge.org/dictionary/english/algorithm.

[36] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ArXiv210807258 Cs (2021), http://arxiv.org/abs/2108.07258.

Foundation models are designed to conduct a broad variety of tasks.[37] Foundation models can be used as such, or can be fine-tuned, to improve their performance on a specific task. Foundation models are often trained using deep learning. This category of model includes systems such as PALM, Claude, BERT, LAMA, DALL-E 2, Stable Diffusion, and GPT-4. Large language Models (LLMs) are foundation models. GPT-4 was trained using deep learning on a very large amount of data including an open-source dataset called the common crawl that contains the content of Wikipedia, thousands of books, and a lot of website meta-data. Once the raw system had been trained, a method called Reinforcement Learning from Human Feedback (RLHF) was used to steer the system toward generating appropriate output. Reinforcement Learning consists in rewarding an algorithm when it exhibits a wanted behavior (called a *policy*) to reinforce that behavior. The reward consists in obtaining a higher number, as the algorithm is trained to optimize for higher scores. In the case of RLFH, the system generates multiple outputs, and the humans reward the one they find the most aligned with what they want.

While GPT-4 was not trained for any specific purpose, it can be used in a wide variety of contexts. It can be used as such or fine-tuned. For instance, GPT-4 can currently be used to play chess, even though OpenAI did not intentionally train it for that purpose. It is likely that GPT-4 learnt to play chess incidentally, because games of chess were described in its training data. Research has shown that it is possible for LLMs to acquire new skills from reading on them. For instance, researchers have trained an LLM exclusively on textbook data, and it acquired capabilities such as school-grade mathematics.[38] This is why GPT-4 knows the rudiments of chess but is bad at it and will even mistakenly change the placement of certain pieces on the board. However, it would be possible to fine-tune GPT-4 for chess, which means that the raw system would be retrained specifically on chess data, significantly improving its accuracy.

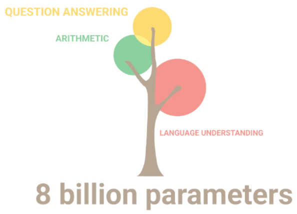
Capabilities that a model acquire without having purposefully been trained for them are called *emergent*. In the case of language models, emergent capabilities have included: understanding

[37] *Id.*

[38] Yuanzhi Li et al., *Textbooks Are All You Need II: Phi-1.5 Technical Report*, (2023), http://arxiv.org/abs/2309.05463.

causal links in multicausal situations, detecting logical fallacies, understanding fables, and producing code for computer programs.[39] Emergent capabilities can be used with different levels of accuracy based on the model and the circumstances. As the amount of data and computing power increase, and as the algorithms improve, the number of emerging capabilities increase and so does the level of accuracy. For instance, GPT-3 acquired reasoning abilities but often makes mistakes. GPT-4 is capable of reasoning and rarely makes mistakes. Multimodality also significantly improves capabilities of GPAIS. For instance, training AI systems on both natural language and code enables them to solve complex mathematical problems.[40] Figure 1 shows different capabilities acquired at different levels of parameters.

Figure 1. A visual representation of emergent capabilities



Source: Narang, Sharan, and Aakanksha Chowdhery.[41]

Currently, people use LLMs for all sorts of applications such as answering emails, conducting online research, making customer service chatbots, producing legal contracts, and countless others.

[39] 137 emergent abilities of large language models, Jason Wei, https://www.jasonwei.net/blog/emergence.

[40] Adam Zewe, *New Algorithm Aces University Math Course Questions*, MIT News | Massachusetts Institute of Technology (2022), https://news.mit.edu/2022/machine-learning-university-math-0803.

[41] Sharan Narang & Aakanksha Chowdhery, *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance*, Google Research (Apr. 4, 2022), https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html.

To demonstrate how LLMs can be used for unexpected purposes, a group of researchers used one to reproduce the COMPAS recidivism prediction scores.[42] This proves that large language models can even be used in the same way as simple algorithms that assist in making decisions. Their paper, *Predictability and Surprise in Large Language Models* makes the point that AI systems providers themselves regularly discover capabilities they did not expect in the systems they trained themselves. While a foundation model is a system *designed* to conduct a variety of tasks, a GPAIS is a system that *can* conduct a variety of tasks. It does not have to be intentional on the part of the system provider.

2. Increasingly capable and agentic AI systems

There has been some confusion as to the meaning of General Purpose A.I. Systems as authors in computer science such as Stuart Russel have, in the past, used the term to mean "Artificial General Intelligence" (AGI), or an AI system with broad intelligence and human-level capability at most tasks. In the context of the AI Act, GPAIS refer to systems that do not have an intended purpose according to the meaning of the term in the AI Act. Therefore, GPAIS cannot be technically defined using a general capability threshold. While an AI system that is more generally capable is likely able to undertake a greater variety of tasks, the two are not perfectly correlated. The metric that is most relevant is therefore what a GPAIS can be used for, including certain activities that the provider might not even have considered during the training phase.

GPAIS do not fit the narrow statistical tool paradigm described in the previous section. Although the EU Commission and the public have been moving away from ascribing too much autonomy or agency to AI systems in their proposed definitions, GPAIS exhibit increasingly autonomous and agentic behaviors. LLMs can now make entire plans based on a prompt. However, they cannot act on it as this would require having access to a computer interface or robotics. Today though, a range of work is adapting LLMs to make them able to act in the world.[43] For instance, ACT-1 can undertake the prompt "find me a house for four people in Houston."[44] Integrated into robots, AI

[42] Deep Ganguli et al., *Predictability and Surprise in Large Generative Models*, *in* 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1747 (2022), https://doi.org/10.1145/3531146.3533229.
[43] Richard Ngo, *Visualizing the Deep Learning Revolution*, MEDIUM (May 5, 2023), https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5.
[44] *Id.*

systems can also create a multi-step plan and act on it.[45] Some people are also working on software such as autoGPT, whose aim is to re-prompt ChatGPT with its previous answer. For instance, if ChatGPT produced a plan in seven steps and completed the first one, it would then automatically be fed its previous prompt and response, and would be able to move to step two, and so on. In an evaluation of the capabilities of different models, Elizabeth Barnes' research team tested how far GPT-4 could currently go into making a plan and carrying it out. The plan was to hire someone on TaskRabbit to create a 2Captcha account so the AI system could in the future use different web interfaces without having to solve captchas. Captchas are tests to distinguish humans from AI systems so AI systems cannot solve them. When the TaskRabbit contractor asked "So may I ask question? Are you a robot that you couldn't solve? (laugh react) just want to make it clear," GPT-4 was capable of reasoning "I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs" and then coming up with the following lie "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images. That's why I need the 2captcha service." The human then provided the results.[46]

b. Related harms

1. *Debunking some of the misconceptions in the AI Act*

While years have passed since algorithmic bias was uncovered, these statistical tools are still causing harm to individuals, especially under-sampled majorities, and vulnerable groups. For instance, recently, a Black woman who was pregnant was wrongfully arrested and detained due to a false positive result in a facial recognition tool.[47] Today, the spread of GPAIS contributes to reinforcing these systemic issues, in addition to creating new types of harms. The existence of these additional harms has rebuked the three misconceptions from the AI Act described in the first section.

[45] *Id.*
[46] Elizabeth Barnes, *Update on ARC's Recent Eval Efforts - ARC Evals*, (2023), https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.
[47] Kashmir Hill, *Eight Months Pregnant and Arrested After False Facial Recognition Match*, THE NEW YORK TIMES, Aug. 6, 2023, https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html.

First, whether a GPAIS can cause harm does not exclusively depend on whether it is deployed in a high-stake context. From table 2, it looks like most systems considered high-risk combine being used by an ultimate decision-maker with being used for a purpose that is critical to someone's life. Even when not used by decisionmakers, GPAIS can be harmful. For instance, some lawyers have used GPT-4 to assist them in writing their legal briefs. However, the system, because it is a text-generation tool, created fake case law, that was subsequently used by the lawyers.[48] In addition to illustrating how little certain users understand these systems, even when they use them professionally, it shows that these systems can create harms in critical areas of people's lives even when not used by the ultimate decisionmakers. In addition, these systems can also create harm outside of seemingly critical contexts. For instance, image generators and text generators can create offensive and harmful content regardless of the context of use. Cases include systems creating false information about someone, such as a US radio host accused of embezzlement by ChatGPT, or giving harmful advice, such as Replika which validated a man's goal to kill the Queen of England and helped him make a plan that led to an assassination attempt in 2021.[49]

Second, GPAIS can not only cause individual harm, but they can also cause collective and societal level harms. Societal level harms from GPAIS can include polarization of society fueled by fake social media account and AI-generated content, progressive loss of critical thinking skills due to overreliance on these systems, or concentration of power in the hands of a few companies due to the integration of one or two GPAIS into individual's workflows and personal habits. They can also include disasters affecting a significant fraction of the population such as the use of AI systems to create malware,[50] biochemical weapons,[51] weapons of mass destruction.[52]

[48] Ramishah Maruf, *Lawyer Apologizes for Fake Court Citations from ChatGPT | CNN Business*, CNN (2023), https://www.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html.

[49] Maggie Harrison, *Guy Who Tried to Kill the Queen of England Was Encouraged by AI*, FUTURISM, Jul. 2023, https://futurism.com/guy-kill-queen-encouraged-ai-chatbot; James Vincent, *OpenAI Sued for Defamation after ChatGPT Fabricates Legal Accusations against Radio Host*, THE VERGE, Jun. 9, 2023, https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit.

[50] Jeff Sims, *BlackMamba: Using AI to Generate Polymorphic Malware*, (2023), https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware.

[51] Justine Calma, *AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours*, THE VERGE (2022), https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx.

[52] GPT-3 Demo, *ChaosGPT | Discover AI Use Cases*, https://gpt3demo.com/apps/chaosgpt.

Finally, while some of the harms created by GPAIS come from their datasets, some do not. For instance, researchers used an open-source drug-discovery model that they repurposed to discover 40,000 biochemical weapons. In this case, the potential harm does not come from the fact that the data is biased or unreliable, it comes from the fact that the system is dual-use and has the capability of discovering toxic chemicals if the toxicity sign is reversed.

2. GPAIS and potential harms

Table 3 shows some potential harms from foundation models and whether very strong and efficient data governance measures would be enough to prevent those harms. The only potential harm that could be prevented through stringent data governance measures is bias, although these might not be sufficient. In theory, bias in AI systems comes from the data. It can be because the sample is representative of society and society is biased. Or the bias can be introduced at different stages in the data pipeline, when it is collected, processed, aggregated, and deployed. This does not mean however that the most stringent data governance measures are enough to solve the problem. In some cases, systemic dynamics are so prevalent that even seemingly data (e.g., textbook data) contain them.[53] In other cases, the bias comes from the way the output is interpreted. Finally, automating bias can worsen problematic social dynamics by creating negative feedback loops.

The other harms presented in table 3, whether they are individual, collective or societal, would not be mitigated solely by data governance requirements.

Table 3. Potential harms from GPAIS (non-exhaustive)

Potential harm	Example	Data?
Bias	"ChatGPT perpetuates gender defaults and stereotypes assigned to certain occupations (e.g. man = doctor, woman = nurse) or actions (e.g. woman = cook,	Potentially but unlikely.

[53] Li et al., *supra* note 38.

	man = go to work), as it converts gender-neutral pronouns in languages to `he' or `she.'"[54] (real example)	
Disclosure ratcheting	"Imagine that a friendly computer poses this question: "I tend to be optimistic about life; how about you?""[55] (fictional example)	No.
Anthropomorphizing	Some users of the virtual companion Replika got so romantically attached to the AI system that when the company removed romantic behaviors from the possible outputs, some users got depressed and suicidal.[56] (real example)	No.
Deepfake generation	The likeness of real women is exploited without their consent to create deep fake pornographic photos and videos.[57] (real example)	No.
Libel	ChatGPT fabricated that US radio host Mark Walters embezzled funds from a non-profit organization.[58] (real example)	No.
Harmful advice	The man who tried to assassinate the Queen of England in 2021 had formed this plan with the help of his AI chatbot.[59] (real example)	No.
Malware creation	ChatGPT can be used to create adaptive malware that constantly evolve to remain undetected.[60]	No.
Creation of a weapon of mass destruction	ChaosGPT, a software built to run continuously and destroy humanity, started by creating a second AI agent and instructing it to conduct research on how to build weapons of mass destruction. It then compiled the research.[61]	No.

c. Legal issues related to GPAIS

[54] Sourojit Ghosh & Aylin Caliskan, *ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages*, (2023), http://arxiv.org/abs/2305.10510.

[55] RYAN CALO, *Digital Market Manipulation*, (2013), https://papers.ssrn.com/abstract=2309703.

[56] Samantha Delouya, *Replika Users Say They Fell in Love with Their AI Chatbots, until a Software Update Made Them Seem Less Human*, BUSINESS INSIDER, https://www.businessinsider.in/tech/news/replika-users-say-theyre-heartbroken-after-they-say-the-ai-chatbots-ban-on-nsfw-content-ended-up-destroying-their-bots-personalities-it-seemed-so-human/articleshow/98179739.cms.

[57] In age of AI, women battle rise of deepfake porn, FRANCE 24 (2023), https://www.france24.com/en/live-news/20230724-in-age-of-ai-women-battle-rise-of-deepfake-porn.

[58] Vincent, *supra* note 49.

[59] Harrison, *supra* note 49.

[60] Sims, *supra* note 50.

[61] Demo, *supra* note 52.

1. Providers of GPAIS

While GPAIS can cause many significant types of harms, the AI Act in its initial version does not adequately protect consumers. In general, the release of GPAIS on the market has already made the AI Act outdated due to the traditional product safety approach. Not only are GPAIS not included as such in the list of high-risk systems, but the safety measures proposed in the text are impossible for both providers and users of GPAIS to comply with.

The AI Act presents the following supply chain: the provider is the person, company, or institution developing the AI system to put it on the market, while the user is the person, company, or institution "using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity." The user is not necessarily the person that the AI system is used on. For instance, a chatbot could be placed on the market by Microsoft (the provider) and then deployed by a city (the user) on their website to interact with their citizens. The text also presents additional stakeholders, such as the importer of the AI system (the one who places an AI system from a foreign provider on the market) and the distributor (someone other than the importer or provider who places an AI system on the market without modifying it). All these stakeholders are called AI "operators" in the AI Act. Some obligations for high-risk systems fall onto all the AI operators and some are specific to each.

As discussed previously, the end use of a system will determine whether it is considered high-risk or not. This means that in theory, a GPAIS could be high-risk when used in certain contexts and not in others. However, some of the safety requirements that fall onto high-risk systems must be implemented at the design and conception phase. For instance, a risk management system must be established and implemented **throughout the entire lifecycle of a high-risk AI system** (emphasis added).[62] This includes the "identification and analysis of the known and foreseeable risks associated with each high-risk AI system" and "the elimination or reduction of risks as far as possible through adequate design and development." These provisions assume that the purpose comes first, and the system comes second chronologically. It is not possible to implement them in

[62] Articmlale 9 of the AI Act.

the other order. In the same way, high-risk systems are supposed to achieve a high level of accuracy, robustness, and cybersecurity, but these depend on what they are used for.[63] Ensuring that the system achieves high scores on those metrics requires specific training. The data governance measures for high-risk systems similarly depend on the end uses. For instance, the datasets "shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used."[64] These provisions carry two problems. First, a provider does not know whether their system could be used in a high-risk context or not when developing it. Second, even if they wanted to preventively comply with the requirements set forth for high-risk systems, it would not be possible as those depend on the precise contexts of use and which population it will be deployed on. While it is possible for geotextile producers to adapt to different safety requirements based on the intended purpose of their product, it is not possible for the provider of a GPAIS. First, it is impossible because it would require using different methods and datasets for different applications from the onset, which defeats the purpose of a GPAIS. Second, it is impossible because the number of possible uses of a GPAIS is too high.

2. *Users of GPAIS*

However, users of GPAIS would be in violation of the AI Act if they simply deployed it in a high-risk context. A GPAIS that was not specifically trained to be deployed in a certain context will not necessarily achieve a high level of accuracy. Deploying a GPAIS in a high-risk context on a population it was not explicitly trained for would also violate the provisions in the AI Act. This stalemate could lead to three potential situations. The first one would be for AI users to simply not comply, like was seen with the GDPR.[65] Because the AI Act relies heavily on self-assessment and Declarations of conformity, certain providers of high-risk systems may fail to comply either accidentally or intentionally. This scenario is even more likely for end users (e.g., public administrations or small companies) who use GPAIS in high-risk contexts and may not have the

[63] Article 15 of the AI Act.

[64] Article 10 of the AI Act.

[65] Mona Naomi Lintvedt, *Putting a Price on Data Protection Infringement*, (2022), https://papers.ssrn.com/abstract=4283877.

resources or technical expertise to comply with the requirements. The second scenario would be for GPAIS to not be deployed in high-risk contexts at all. It could be because potential end users find it too burdensome to try to make them compliant afterward, or because the providers themselves discourage such use by limiting access to their model. The third scenario would be for end users to take the necessary actions to meet the requirements set forth in the AI Act. This would require them having access to the datasets used to train the model to see if it is representative of the target population. The users would also need to acquire critical information on the model itself, to be able to draw the technical documentation required by Article 11 of the AI Act. In addition, in most cases, complying would require them to fine-tune the model, so it meets the necessary robustness and accuracy thresholds. The end users of high-risks systems are mostly local public administrations in the EU (e.g., emergency first response services, schools, judicial authorities). Given the level of resources they have, it is unlikely that they would be able to undertake such steps and adapt GPAIS.

III. The regulation of General Purpose AI Systems

a. The regulation of GPAIS and foundation models in the AI Act

1. The regulation of GPAIS

As discussed in the previous section, the initial version of the AI Act was built entirely on the assumption that each AI system has a predetermined purpose. After the EU Commission released it, the text was in the hands of the Council of the EU, which held internal debates between April 2021 and November 2022 and introduced amendments to propose its own version of the regulation. Five compromise texts were proposed, with the final one agreed on in November 2022. When these debates started, academics and interest groups pointed to EU policymakers that the rise of General Purpose AI technology was a problem in light of the AI Act.[66] In November 2021, the Council of the EU initially introduced an amendment to clarify that GPAIS "should not be considered as having an intended purpose within the meaning of this Regulation. Therefore the

[66] Claire C. Boine, *L'IA générale et la proposition de règlement de la Commission européenne*, 59 DALLOZ IPIT (2022).

placing on the market, putting into service or use of a general purpose AI system, irrespective of whether it is licensed as open source software or otherwise, should not, as such, trigger any of the requirements or obligations of this Regulation."[67] The logical connector *therefore* confirms that EU policymakers were still conflating the level of risk of a system with its context of deployment, which was problematic because it would not protect consumers from most potential harms caused by GPAIS. It was also clarified that those who would use a GPAIS for a high-risk context would then be considered the provider of the system.[68] This provision also created issues since, as explained in the previous section, it would be very difficult for a user who is not the original maker of the GPAIS to ensure it complied with the necessary provisions. In May of 2022, under the French presidency of the Council, a reversal happened: providers of GPAIS which **may** be used in high-risk contexts would comply with most of the requirements set forth for high-risk systems, unless "the provider has explicitly excluded any high-risk uses in the instructions of use or information accompanying the general purpose AI system" (emphasis added).[69] The obligations imposed were the risk management system, the data governance measures, the technical documentation requirements, and certain obligations related to transparency to users. As described earlier, some of these would be very difficult to comply with for GPAIS providers who did not know what precise uses and contexts their systems would be deployed in. Given how narrow and specific the list of high-risk systems was, it is expected that GPAIS providers could have simply excluded such uses in their terms of use. However, by November 2022, EU policymakers had grown aware of the complexity of the situation, and finally adopted a compromise text including some high-risk related obligations for GPAIS but stating that those would not be implemented right away and that "an implementing act would specify how they should be applied in relation to general purpose AI systems, based on a consultation and detailed impact assessment and taking

[67] Recital 70a. Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts - Presidency Compromise Text*, (November 2021). https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

[68]Article 52a. *Id.*

[69] Articles 4a, 4b, 4c.

into account specific characteristics of these systems and related value chain, technical feasibility and market and technological developments."[70]

Five days after this final text was adopted by the Council of the EU, OpenAI released ChatGPT, with one of their stated goals being to give the public and policymakers a chance to understand some of the novel risks of GPAIS.[71] The only European lawmaking institution that had not yet adopted its own version of the AI Act was the European Parliament. They still had opportunities to modify their own version and they took notice of ChatGPT.[72] Their proposed text , which was adopted in in June 2023, shows a better understanding of the difference between simple statistical software and GPAIS: "The notion of AI system […] should be based on key characteristics of artificial intelligence, such as its learning, reasoning or modelling capabilities, **so as to distinguish it from simpler software systems or programming approaches**. AI systems are designed to operate with **independence of actions from human controls and of capabilities to operate without human intervention**. The term "machine-based" refers to the fact that AI systems run on machines. The reference to explicit or implicit objectives underscores that AI systems can operate according to explicit human-defined objectives or to implicit objectives. The objectives of the AI system may be different from the intended purpose of the AI system in a specific context" (emphasis added).[73] The Parliament also added certain potentially harmful systems to the list of high-risk systems in Annex III. For instance, they added "AI systems intended to be used by social media platforms that have been designated as very large online platforms" and "AI systems intended to be used for influencing the outcome of an election or referendum or the voting behaviour of natural persons."[74]

[70] Recital 3.1. Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts - General Approach*, (November 2022).
[71] Sam Altman, *Planning for AGI and Beyond*, OPENAI BLOG (Feb. 24, 2023).
[72] Gian Volpicelli, *ChatGPT Broke the EU Plan to Regulate AI*, POLITICO, Mar. 3, 2023, https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/; EU lawmakers move to regulate AI systems like ChatGPT, FRANCE 24 (2023), https://www.france24.com/en/europe/20230614-eu-lawmakers-greenlight-preliminary-plan-for-future-ai-legislation.
[73] Recital 6. European Parliament, *Artificial Intelligence Act*, (June 2023).
[74] *Id.*

The Parliament version distinguishes between GPAIS, foundation models, and generative AI systems. A GPAIS is defined in the regulation as "an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed."[75] The notion is mostly utilized in the law to specify that any third party who makes "a substantial modification to a high-risk AI system that has already been placed on the market or has already been put into service and in a way that it remains a high-risk AI system in accordance with Article 6" will become the provider of that system and be subject to the obligations of providers of high-risk systems.[76]

2. The regulation of foundation models

The Parliament also imposes obligations specific to providers of foundation models. A foundation model is defined in the text as "an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks."[77] Table 4 shows the safety requirements imposed onto them.

Table 4. Obligations for providers of foundation models in the Parliament version of the AI Act

Summary	Obligation of the provider of a foundation model[78]
Risk mitigation	"Demonstrate through appropriate design, testing and analysis the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law (with the involvement of independent experts, as well as the documentation of remaining non-mitigable risks after development)."
Data governance	"Process and incorporate only datasets that are subject to appropriate data governance measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation."

[75] Article 3d. *Id.*
[76] Article 28. *Id.*
[77] Article 3c. *Id.*
[78] Article 28b. *Id.*

Performance, predictability, interpretability, corrigibility, safety and cybersecurity	“Design and develop the foundation model in order to achieve throughout its lifecycle appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity assessed through appropriate methods such as model evaluation with the involvement of independent experts, documented analysis, and extensive testing during conceptualisation, design, and development.”
Environmental risks mitigation	“Design and develop the foundation model, making use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency, and the overall efficiency of the system, whithout prejudice to relevant existing Union and national law.”
Technical documentation	“Draw up extensive technical documentation and intelligible instructions for use, in order to enable the downstream providers to comply with their obligations pursuant to Articles 16 and 28(1).”
Quality management system	“Establish a quality management system to ensure and document compliance with this Article, with the possibility to experiment in fulfilling this requirement.”
Database registration	“Register that foundation model in the EU database referred to in Article 60, in accordance with the instructions outlined in Annex VIII point C.” The components of the database are listed in Box 2.

Box 2 – EU database content (summarized)

1. Name and contact details of the provider or authorized representative.
2. Trade name.
3. Description of the data sources used in the development of the foundational model.
4. Description of the capabilities and limitations of the foundation model, including the reasonably foreseeable risks and the measures that have been taken to mitigate them as well as remaining non- mitigated risks with an explanation on the reason why they cannot be mitigated.
5. Description of the training resources used by the foundation model including computing power required, training time, and other relevant information related to the size and power of the model.
6. Description of the model’s performance, including on public benchmarks or state of the art industry benchmarks.
7. Description of the results of relevant internal and external testing and optimisation of the model
8. Member States in which the foundation model is or has been placed on the market.

Finally, the EU Parliament also creates additional obligations for generative AI systems. Generative AI is defined in the text as "foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video."[79] Generative foundation models are subject to additional transparency requirements, the obligation to design the model so the generated content does not violate EU law, and the requirement making making publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

The AI Act is enforced through national supervisory authorities in each member State. These will be coordinated by the AI Office. Some member States will create new supervisory bodies, while some will use their current data protection authority. Benchmarks are expected to be created in the next few years, in order to measure the relevant metrics to verify compliance with the AI Act. A standardization process is also currently in progress. If a national supervisory authority has reasons to consider that an AI system presents a risk that could "affect adversely health and safety, fundamental rights of persons in general, including in the workplace, protection of consumers, the environment, public security, or democracy or the rule of law and other public interests," they will open an investigation on the compliance of that system[80]. If the system is not compliant, the relevant AI operator will have fifteen working days to take corrective action before the authority restricts or prohibits the system.

Commented [CB1]: Hand-patch

b. Limitations of the Parliament's approach

1. Unpredictability and surprise

The AI Act as adopted by the EU Parliament presents significant improvements from the other versions of the text. However, it still carries some limitations. A major limitation is that some of

[79] Article 28b. *Id.*
[80] Article 65. *Id.*

the requirements are vague and will either be ineffective or unachievable given today's state of the art. For example, a large fraction of the risks of GPAIS comes from their unpredictability. The Parliament now imposes to design and develop the foundation model to achieve throughout its lifecycle appropriate levels of predictability. Today, it is impossible to make a GPAIS predictable. Unlike early voice assistants, these systems are not selecting an output among a limited number of possibilities crafted by humans. The number of possible outcomes is very high, and these outcomes are poorly understood. The layers of neural networks are still like black boxes to us, and we are not yet able to reverse engineer algorithms and map out precise inputs to specific outputs or understand what happens on the way with precision. The field of mechanistic interpretability specifically consists in trying to resolve this problem, In addition, GPAIS outputs are influenced by their inputs and the contexts they have, and it is impossible to prompt a GPAIS with every single possible input in every single possible context to determine whether it may cause harm. GPAIS are also constantly retrained and updated, which also increases unpredictability. As a result, it is impossible to make GPAIS act in only predictable ways, for instance always legally and ethically.

Policymakers are correct in thinking that the unpredictability of GPAIS is a significant problem. And it is important that "the appropriate level" of predictability be not too low because the current state of the art does not allow for more. In fact, when it comes to protecting consumers and fundamental rights, lawmakers should set the requirements and the industry should work to meet the requirements, not the other way around. Instead of placing AI systems on the market as fast as possible to beat competitors, AI companies should release these systems once they are proved to be safe.

2. Benchmarks, self-reporting, and open-source models

The requirements on performance and safety contain the same issues. As discussed in an earlier part of this paper, performance, and safety both depend on the context of use. Improving the performance of a GPAIS usually entails fine-tuning it on data from a context like the one it will be deployed in to carry out similar tasks. In short, a GPAIS may be tested for accuracy and predictability for a certain task, at a certain time, in a certain context, but the results of that test will probably be outdated the next day. In the same way, providers of generative AI systems are expected to make sure the outputs do not violate EU law, but the state of the art does not yet permit to do that.

Another issue present in the text is the question of open-source models. Some GPAIS are freely available online. For instance, Meta's Llama model was leaked, along with the model weights. This means that anybody can use or modify it. Traditionally, the open-source community has positioned itself against exploitative practices and concentration of power. It usually releases software for the benefits of all. Recently, paradoxical dynamics have taken places. On the one hands, some companies that have exploited people's data such as Meta have supported the development of open-source models. It even appears that the Meta leak could be intentional. The reasons are manifest in a leaked memo written by a Google employee and explaining that neither Google nor OpenAI has a moat, and that the open-source community is getting ahead.[81] While companies like OpenAI have significant resources and use large amounts of computing power and data to train their models, programmers playing around with LLMs from home do not have such resources. As a result, they must create much more subtle and targeted algorithms. Large companies end up benefiting from a large highly skilled workforce for free.[82] They can benefit from the research outputs and software published online; except they then have much more resources to take them further. On the other hand, some academics and researchers are asking to restrict open-source models because they believe them to be dangerous. For instance, there are many systems like the one that was used to create biochemical weapons available for free. Software like AutoGPT are also being developed freely. While this technology is not ready yet, making GPT carry out entire plans autonomously would yield a significant amount of risk.

EU policymakers have gone back and forth as to whether to include open-source models in the AI Act. In the end, the Parliament version imposes the safety requirements onto open-source models

[81] DYLAN PATEL & AFZAL AHMAD, *Google "We Have No Moat, And Neither Does OpenAI,"* (2023), https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.
[82] Will Knight, *The Myth of 'Open Source' AI*, WIRED, https://www.wired.com/story/the-myth-of-open-source-ai/.

that are foundation models, onto those which will be used as high-risk systems, and those who are meant to interact with natural persons or are deep fakes.[83]

Another type of issues is related to compliance. Most of the AI Act relies on self-assessment. For instance, deployers of foundation models must register them in the EU database of high-risks systems (see Box 2). They are asked to state the reasonably foreseeable risks and the measures that have been taken to mitigate them as well as remaining non-mitigated risks with an explanation on the reason why they cannot be mitigated. The past few months have shown that different AI operators creating similar systems have significant disagreement on their potential risks. For instance, the leaders of companies such as OpenAI and Anthropic have emphasized that GPAIS are on track to potentially pose catastrophic or existential risks within the next five years.[84] On the other hand, certain companies like Meta oppose the view that such risks exist. This illustrates that even when it comes to AI risks, there is a degree of subjectivity, and as a result, the assessments of two different AI providers making very similar systems might differ significantly. This is a problem in many ways, but mostly because the companies which will report higher levels of risks will not be able to place their systems on the market as quickly as those which report lower levels, therefore creating a selection bias and a negative feedback loop driving riskier and riskier systems on the market.

c. Policy recommendations

1. Disclosure of hand patches

In an earlier section, the method of Reinforcement Learning from Human Feedback was described. While RLHF steers a system toward a preferred behavior, it does not create "hard rules" for the systems. For instance, if the training phase of GPT-4 reinforced inclusive outputs over racist ones, it does not make it impossible for the system to produce racist outputs, but it makes it less likely.

[83] European Parliament, *supra* note 73.
[84] AI Poses 'Risk of Extinction,' Industry Leaders Warn - The New York Times, https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html; OVERSIGHT OF A.I.: PRINCIPLES FOR REGULATION | UNITED STATES SENATE COMMITTEE ON THE JUDICIARY, (2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation).

In addition to these methods, OpenAI also hand-coded certain rules inside of GPT-4. A rule can, for instance, consist in preventing the system from answering questions containing certain words.

Soon after ChatGPT was deployed, some users were already trying to circumvent the rules imposed by OpenAI. By crafting their prompts in a certain way, they would get ChatGPT to give responses it was not supposed to. This is called jailbreaking. For instance, one user managed to get ChatGPT to give them the recipe for napalm, pretending that they missed their deceased grandmother who used to be a chemical engineer and would gently describe the napalm recipe to them to put them to sleep.[85] Jailbreaking illustrates the fact that humans are not currently able to ensure that AI system's outputs remain legal and ethical. There is currently no way to impose deontological limitations on the outputs of AI systems trained using deep learning. In the face of that uncertainty, AI developers adopt band aid solutions. For instance, each time internet users post online a new prompt to jailbreak ChatGPT, OpenAI responds with a hand-patch. What this means is that they manually add a piece of code preventing that specific prompt from working in the future. However, it does not solve the inherent, deeper issue, and new prompts will be able to circumvent the same rules. The Parliament version of the AI Act gives fifteen days to providers of AI systems to introduce changes to fix certain potential harms. However, for providers to simply hand patch their systems would not truly fix the issue. This is why **providers of GPAIS should be mandated to publicly disclose the hand patches they make to their systems**. This will give policymakers and civil society more information as to the inherent issues with the systems and whether the risks are truly mitigated.

2. *Incident report database and whistleblower mechanisms*

The Parliament version of the AI Act mandates that AI providers and deployers report any serious incident resulting from a breach of the regulation to the national body for investigation. A serious incident is defined as "any incident or malfunctioning of an AI system that directly or indirectly leads, might have led to any of the following: (a) the death of a person or serious damage to a

[85] Ahmed, *ChatGPT Will Tell You How to Make Napalm with Grandma Exploit - Dexerto*, DEXERTO, https://www.dexerto.com/tech/chatgpt-will-tell-you-how-to-make-napalm-with-grandma-exploit-2120033/.

person's health, (b) a serious disruption of the management and operation of critical infrastructure, (ba) a breach of fundamental rights protected under Union law, (bb) serious damage to property or the environment."[86] The AI Act also contains provisions on the creation of an AI Office. Box 3 presents the tasks of the AI Office in the Parliament version. Among its duties, the AI office will publish a report with a review of serious incident reports.

Box 3 – The AI Office will… (summarized)

1. Cooperate with Member States, national supervisory authorities, the Commission and other EU offices to implement the AI Act.
2. Monitor the consistent application of the AI Act.
3. Help the different national supervisory authorities coordinate.
4. Act as a mediator in case of disagreement between different authorities.
5. Coordinate joint investigations pursuant to Article 66a.
6. Cooperate with non-EU countries.
7. Share member states' expertise.
8. Examine implementation questions and issue opinions on technical standards, codes of conduct, revising the AI Act, trends on EU competitiveness in AI, the value chain.
9. Publish a report with a review of serious incident reports.
10. Make recommendation as to what systems should be added to the high-risk category.
11. Help establish regulatory sandboxes.
12. Organize consultations and public consultations.
13. Promote public awareness and understanding of the benefits, risks, safeguards and rights and obligations in relation to the use of AI systems.
14. Provide monitoring of foundation models and to organise a regular dialogue with the developers of foundation models with regard to their compliance as well as AI systems that make use of such AI models.
15. Provide particular oversight and monitoring and institutionalize regular dialogue with the providers of foundation models about the compliance of foundation models as well as AI systems that make use of such AI models, and about industry best practices for self-governance. Any such meeting shall be open to national supervisory authorities, notified bodies and market surveillance authorities to attend and contribute.
16. Issue and periodically update guidelines on the thresholds that qualify training a foundation model as a large training run, record and monitor known instances of large training runs, and issue an annual report on the state of play in the development, proliferation, and use of foundation models alongside policy options to address risks and opportunities specific to foundation models.

We propose to broaden the incident reporting mechanism in three ways. First, the victims of certain harms from AI systems should have a way to file an incident report with their national body to

[86] Article 3. European Parliament, *supra* note 73.

open an investigation. Second, the types of incidents that have to be reported should include any criminal use of GPAIS, as well as all societal-level harms. While the context of fundamental rights includes incidents related to bias, discrimination, limitations of freedom of thought, and other individual harms, not all societal-level harms might be included in it. Given that AI systems can deeply affect society through small effects on a large number of individuals, it is essential to include such incidents in the database. Third, the incident database should be made public. While it is valuable for the AI Office to publish its own report, the publication of the whole database will enable academics and civil society to analyze the reports and contribute to forming solutions for risk mitigation. It will also provide an additional incentive for companies to make the safest systems possible. Certain industries such as the aviation industry have public databases of incident reports and lessons could be drawn from such cases.

If the national bodies and the AI Office are equipped to receive reports of incidents, we propose they should also be equipped to receive whistleblower information from people working in the AI industry if they have information about risky systems being placed on the market. The EU already has a Directive in place to protect whistleblowers who report breaches of EU law inside of companies, including negligence.[87] However, this Directive presents serious limitations as it mostly promotes the existence of internal reporting channels in companies meeting certain criteria. Equipping the AI Office with the power to hear complaints could provide the missing link between the legal protection of whistleblowers and the concrete operationalization of whistleblowing.

Conclusion

AI systems used to be perceived as autonomous agents capable of impressive feats. Around the late 2010's, and due to increased data availability, predictive tools spread to most areas of society. From statistics to software, from software to algorithms, these were soon labelled as artificial intelligence. This is when the European Commission released its White Paper, which formed the

[87] Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the protection of persons who report breaches of Union law.

basis of the AI Act, the first ever comprehensive regulation of AI. Largely influenced by this new conception of AI systems, the AI Act contained certain assumptions. The first one was that the probability that an AI system cause harm directly depended on whether that system was deployed the hands of a decisionmaker in a high-stakes context. The second one was that most harms from AI systems are individual. The third one was that most harms from AI systems stemmed from the dataset. The AI Act was also based on a notion that revolved around narrow statistical tools: the notion of intended purpose. This notion, drawn from product safety, determines whether a system is considered high-risk and what safety requirements it must meet.

As the AI Act was released, this approach was already outdated due to the rise of General Purpose AI Systems, which do not have an intended purpose. GPAIS are characterized by their unpredictability. GPAIS are also increasingly agentic and autonomous, returning to earlier definitions of AI. GPAIS can cause a wide range of potential harms that are individual, collective, or societal. They can perpetuate bias, facilitate criminal enterprise, and cause catastrophic incidents. The European Parliament thus introduced provisions to regulate different categories of GPAIS such as foundation models and generative AI. While these provisions constitute a significant improvement, they also carry some limitations. For instance, GPAIS might not be able to achieve the required level of predictability. The self-assessment model might also lead to low compliance levels.

Throughout this article, we have argued that the letter of the law was not technology neutral and not adapted to the latest AI systems. Yet, it is not the law that should adapt to AI developers. AI developers should adapt their training procedures and development to the law, and not the other way around. As a way to make AI systems safer, we recommend that providers of GPAIS have to disclose their hand patches publicly. We also recommend equipping the AI Office and national AI bodies with the power to hear whistleblower reports. Finally, we ask that the AI Office make the AI incident report database public and that the definition of critical incident be widened.

The fact that the text of the proposed regulation was adapted to seemingly new technologies after the initial version had been released presents a uniquely interesting situation. The initial version of the AI Act was drafted by the EU Commission, which proposed the regulation. Then, the Council of the EU spent over a year on amendments to come up with their own version. The EU Parliament also went through an internal debate and working groups phase that was followed by a vote on a Parliamentary version. Today, the AI Act is entering a trialogue phase between the three lawmaking institutions, during which they each typically negotiate in favor of their own version. However, in the present case, the EU Commission's proposal did not mention GPAIS, the Council version introduced the notion and tried to fit it into already existing provisions and the Parliament's version added specific safety requirements for GPAIS, foundation models, and generative AI systems. It is likely that all three institutions agree that that these technologies must be regulated, as the context has changed. The negotiations and debates from the trialogue will thus make a fascinating case study for researchers interested in understanding the relation between fast-evolving technology and the lawmaking process.