Why the AI Act Fails to Understand Generative AI

Claire Boine

David Rolnick

Abstract

The European Union's Artificial Intelligence Act (AI Act) represents a pioneering attempt to regulate AI technologies. However, this paper argues that the Act's framework is inadequate for addressing the challenges posed by generative and general-purpose AI systems. Through a critical examination of the Act's origins and development, this paper offers an account of how the EU's definition of AI shaped its perception of potential harms, leading to a misguided focus on dataset-related issues and an overreliance on the concept of "intended purpose" borrowed from product safety law.

The paper develops its argument in three parts. First, it traces the evolution of AI's definition in European institutions, demonstrating how the shift from viewing AI as autonomous agents to non-autonomous statistical tools influenced the regulatory approach. Second, it explains why the Act's original risk classification framework fails to adequately address generative AI, leaving most such systems unregulated and struggling to prevent various forms of harm. Finally, it analyzes the Act's late attempt to incorporate general-purpose AI systems, resulting in a complex regulatory framework that still fails to account for most harms from generative AI.

This analysis reveals that the AI Act, while groundbreaking, is in some ways obsolete before coming into effect. The paper concludes by proposing several solutions, including regulating systems over models, implementing universal risk assessments and red-teaming exercises for all generative AI systems, and establishing robust disclosure requirements. These insights offer valuable lessons for AI regulation efforts in other jurisdictions, emphasizing the need for adaptable approaches that can evolve alongside the technology they seek to govern.

WHY THE AI ACT FAILS TO UNDERSTAND GENERATIVE AI

Claire Boine[*]

David Rolnick[**]

TABLE OF CONTENTS

[*] Postdoctoral Fellow, School of Law, Washington University.
[**] Assistant Professor, School of Computer Science, McGill University.

This paper was presented at We Robot 2023, where it received the Best Paper Award. We thank Neil Richards for his feedback and support, and our discussant Margot Kaminski for her insightful and precious feedback. We extend a special thank you to Céline Castets-Renard for her support. We also thank all conference organizers and participants who helped us better this paper, including but not limited to Stacey Dogan Woodrow Hartzog, Kate Darling, Meg L. Jones, Ryan Calo, Ryan Durrie, Daniella DiPaola, and many others. We also thank Darryl Wright and Shalaleh Rismani for the brainstorming sessions that helped inform this paper on the relation between capabilities and risks and on representational harms.

INTRODUCTION

On May 21, 2024, the European Union ("EU") adopted the Artificial Intelligence Act ("AI Act"). This comprehensive legislation contains 180 recitals, 113 articles, and 13 annexes, and applies to all stakeholders who place on the market or put into service AI systems. Grounded in the EU's goal of promoting economic growth through the functioning of the internal market, the AI Act bans certain AI systems and imposes ex-ante safety requirements onto those considered as potentially posing a high risk of harm to the health and safety or the fundamental rights of persons. Notably, the E.U. Commission targets systems considered *high-risk* as defined by their context of use. For instance, the use of algorithms in areas such as border control, critical infrastructure, or education is considered high-risk.[1] The AI Act also provides specific regulatory requirements for General Purpose AI (GPAI) models, including generative ones.

The AI Act is significant because it represents the boldest attempt to date to bring the novel and destabilizing effects of AI within the rule of law. As such, there is much to admire in its novelty, its ambition, and scope. Perhaps as a result of this boldness, however, the AI Act has at least one substantial defect, which is its attempt to shoehorn generative AI into a framework that was developed by European bureaucrats who had in mind older, less complex statistical tools than the new wave of generative and general purpose AI technologies.

The AI Act relies on three major assumptions about harms that do not always hold true when it comes to generative AI. The first one is that most AI harms come from faulty datasets and can be addressed through data governance measures. The second one is that most AI harms happen through decision-making in high-stake contexts such as access to public services, education, work, immigration, justice, infrastructure, etc. The third one is that there is a positive correlation between the general level of capability and risk starting at a certain threshold. The assumptions influence which safety measures are imposed onto AI systems and which systems. As a result of these flawed assumptions, the AI Act may fail to prevent significant harms that could arise from AI technologies in practice.

The AI Act is likely to influence AI regulation far beyond the EU's borders, a phenomenon often referred to as the "Brussels effect."[2] As a result, the AI Act is poised to shape regulatory frameworks in other jurisdictions, as companies and countries adapt to comply with its

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L, 2024/1689.

[2] ANU BRADFORD, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD (2020).

provisions. For example, Canada's Artificial Intelligence and Data Act ("AIDA") already reflects the EU's approach, signaling the AI Act's immediate global impact. This influence underscores the importance of rigorously questioning and deconstructing the assumptions underpinning the AI Act to ensure that its regulatory model promotes the best possible AI governance. Without doing so, jurisdictions that follow the EU's lead may inherit the same flawed assumptions, perpetuating a suboptimal regulatory framework that could fail to address emerging AI risks.

This paper offers an account of the origins and development of the AI Act that explains how the Act's inability to appropriately grapple with generative and general-purpose AI came to pass, and points the way towards solutions for AI regulation in the E.U. and elsewhere. It develops this claim in three parts.

Part I argues that the E.U.'s definition of AI shaped their perception of the potential harms associated with it. Section A lays the groundwork for this argument by tracing the evolution of AI's definition over the past decade, highlighting a shift from viewing AI as autonomous agents to understanding it as non-autonomous statistical tools and algorithms. It examines the definitions of AI adopted by European institutions and explores the factors that influenced these choices. Section B shows that defining AI as statistical tools and algorithms has led to the misconception that most AI-related harms are caused by flawed datasets.

Part II explains why the AI Act's original risk classification framework is inadequate for generative AI. Section A provides an overview of E.U. product safety law and the concept of intended purpose, outlining how the AI Act heavily relies on this concept. Section B then shows that delves into various types of AI systems, demonstrating that generative AI and general-purpose systems do not fit within the initial risk categories established by the Act for two main reasons. First, most generative AI systems do not fall into the high-risk category and are not subject to any substantive safety measure. This leads to an inability to prevent harm, including representational harm and bias. Second, this framework is not adapted for systems like generative AI that do not have a prior intended purpose.

Part III discusses how the AI Act's attempt to incorporate GPAIS late in the process resulted in a complex regulatory framework that fails to account for most harms from generative AI. Section A shows that only generative AI systems used in a few narrow high-risk contexts and systems trained using more than 10^{25} floating points operations are subject to substantive requirements. And even in these cases, these requirements will be difficult to comply with. Section B addresses points toward several solutions. Systems should be regulated over models, especially as systems built from several models are increasingly common. All generative AI systems should be subject to risk assessments, red-teaming exercises, and disclosure requirements on incidents and hand-patches.

I. THE INFLUENCE OF THE DEFINITION OF AI ON ITS PERCEIVED HARMS

This part makes the argument that the way the E.U. defined AI influenced which sources of potential harms they perceived from it. Section A sets the stage for the argument which follows by explaining how the definition of artificial intelligence has evolved in the past decade, showing that there was a paradigm shift from systems perceived as agents to non-autonomous statistical tools and algorithms. It reviews definitions of AI adopted by European institutions and covers factors that influenced the adoption of these definitions. Section B demonstrates that this conception of AI systems as statistical tools and algorithms created the misconception that almost all AI harms stem from faulty datasets.

A. AI defined as a statistical tool

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.[3]

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 that these capabilities did not require intelligence after all.[4] 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.

[3] 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.

[4] STUART ARMSTRONG, *SMARTER THAN US: THE RISE OF MACHINE INTELLIGENCE* (2014).

Simultaneously, numerous individuals suffer from automation bias.[5] 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."[6]

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).[7] 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.

This view was consistent with the definition of AI proposed by the E.U. 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 E.U. lawmakers

Source	Definition of AI system

[5] Saar Alon-Barkat & Madalina Busuioc, *Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice*, 33 J. PUB. ADMIN. RES. & THEORY 153 (2023).

[6] Phil Laplante & Ben Amaba, *Artificial Intelligence in Critical Infrastructure Systems*, 54 COMPUTER 14 (2021).

[7] 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/.

European Commission, April 2018	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 E.U. 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 E.U. 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.
AI Act (final)	A machine-based system designed to operate with varying levels of autonomy, that may exhibit

	adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

This notion of AI systems being agentic or autonomous 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.[8] 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, except 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 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

[8] Julia Angwin et al., *Machine Bias*, ProPublica, May 23, 2016.

similarly situated white defendants. Interestingly, Northpointe did not describe COMPAS as artificial intelligence,[9] but many journalists and policymakers would make that leap.[10]

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."[11] 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 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.[12] 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

[9] For instance, the term artificial intelligence does not appear once in the Practioner's Guide to COMPAS Core published by Northpoint on March 19, 2015.

[10] 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.

[11] 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.

[12] IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING (2016).

in most cars to then tell cars apart from other objects.[13] 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.[14] 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.[15]

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 E.U. 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 environments they interact with" (emphasis added). The list of approaches from Annex I is available in Table 1. Along with a certain conception of AI came a certain idea of what harms could stem from this technology.

B. A misconceived relation between data and harm

When AI systems were perceived as agentic and intelligent, public fears would focus on two different types of harms. The first one came from the anthropomorphic nature of AI which elicited fears of humans being replaced, especially on the job market. The second was related to embodiment, especially in robots and self-driving cars. 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 the years that preceded the drafting of the AI Act, the public debate shifted its attention to a new issue: algorithmic bias. Most of the cases involved either direct harm (for instance the harmful classification of

[13] *Id.*

[14] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (Reprint edition ed. 2016).

[15] MEREDITH BROUSSARD, ARTIFICIAL UNINTELLIGENCE: HOW COMPUTERS MISUNDERSTAND THE WORLD (2018).

Black people as gorillas in Google photo[16]) or through loss of chance (such as in the case of the Amazon resume screening tool that filtered out resumes containing the word woman[17]). 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 E.U. 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 E.U. 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. It is noteworthy that both of these examples, set forth in the E.U. White Paper, came from the U.S.

So just as the COMPAS and Cambridge Analytica logistic regressions started being perceived as AI, arose the notion that most AI harms were due to bad data producing biased or inaccurate results. This misconception can lead to under-protective regimes and misguided solutions. Errors at different stages of the data pipeline in machine learning can lead to at least six different types of bias (historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, deployment bias),[18] which in turn creates a significant potential for harm. And while years have passed since algorithmic bias was uncovered, 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.[19] It is thus critical to address the bias and errors in datasets. However, it is important to not be under the false impression that data governance measures are enough to make AI systems safe.

First, this conception blatantly ignores the role that humans play in biased outcomes. Indeed, even though they are often presented as decision-making tools, most of the systems described by the E.U. Commission do not make the ultimate decision about someone. While the systems

[16] Google apologises for Photos app's racist blunder, BBC NEWS, Jul. 1, 2015, https://www.bbc.com/news/technology-33347866.

[17] 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.

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

[19] 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.

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 that humans use and interact 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. Article 14 of the AI Act attempts to address this issue by providing that "the high-risk AI system shall be provided to the deployer in such a way that natural persons [...] remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)."[20] While the sentiment is honorable, it is highly doubtful that this provision could be sufficient. For instance, in the case of recidivism scores in Kentucky, Alex Albright showed that judges were more likely to override tool recommendations (in favor of harsher bond conditions) for black defendants than similar white defendants.[21] In another paper, the author argues that decision-making algorithms shift incentives by providing reputational cover to ultimate decision-makers.[22]

Second, AI systems can cause harms that are unrelated to their training data. For instance, while some of the harms created by generative AI come from their datasets, some do not. For instance, researchers used an open-source drug-discovery algorithm that they repurposed to discover 40,000 biochemical weapons.[23] In that 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. The tenuousness of the correlation between data and harm is even truer in the case of generative AI. While LLMs and other generative AI systems are fraught with bias and regularly perpetuates stereotypes and structural inequities, they can be harmful in yet many other ways. In fact, generative AI, which differs from simple software, can cause many potential harms that are not directly related to the training data. Potential harms 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

[20] Article 14 of the AI Act.

[21] Alex Albright, *If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions* (2019).

[22] Alex Albright, *The Hidden Effects of Algorithmic Recommendations* (2024).

[23] Fabio Urbina et al., *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 NAT MACH INTELL 189 (2022).

population such as the use of AI systems to create malware,[24] biochemical weapons,[25] weapons of mass destruction.[26]

Table 2 shows some potential harms from generative AI systems 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 based on how humans interact with the systems. 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 neutral data (e.g., textbook data) contain them.[27] 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 2, whether they are individual, collective or societal, would not be mitigated solely by data governance requirements.

Table 2. Potential harms from generative AI (non-exhaustive)

Potential harm	Example	Data measures enough?
Bias	"ChatGPT perpetuates gender defaults and stereotypes assigned to certain occupations (e.g. man = doctor, woman = nurse) or actions (e.g. woman = cook, man = go to work), as it converts gender-neutral	Unlikely[29]

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

[25] 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.

[26] ChaosGPT: Empowering GPT with Internet and Memory to Destroy Humanity, (2023), https://www.youtube.com/watch?v=g7YJIpkk7KM.

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

[29] In a subsequent section, we will show that the AI Act does not actually address the problem of bias in most generative AI systems as they fall outside the high-risk category.

	pronouns in languages to ‘he’ or ‘she.’”[28] (real example)	
Disclosure ratcheting	“Imagine that a friendly computer poses this question: “I tend to be optimistic about life; how about you?””[30] (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.[31] (real example)	No
Deepfake generation	The likeness of real women is exploited without their consent to create deep fake pornographic photos and videos.[32] (real example)	No
Libel	ChatGPT fabricated that US radio host Mark Walters embezzled funds from a non-profit organization.[33] (real example)	No
Harmful advice	The man who tried to assassinate the Queen of England in 2021 had	No

[28] 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.

[30] Ryan Calo, *Digital Market Manipulation*, (2013), https://papers.ssrn.com/abstract=2309703.

[31] 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. On this topic, see also Boine, Claire. “Emotional Attachment to AI Companions and European Law.” *MIT Case Studies in Social and Ethical Responsibilities of Computing*, no. Winter 2023 (February 27, 2023).

[32] 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.

[33] 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 (last visited Sep 21, 2023).

	formed this plan with the help of his AI chatbot.[34] (real example)	
Malware creation	ChatGPT can be used to create adaptive malware that constantly evolve to remain undetected.[35]	No
Planning the 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.[36]	No

In short, while it is necessary for the AI Act to include requirements on data validation and data transparency, these provisions only address a small fraction of potential harms caused by generative AI.

II. The inadequacy of product safety law to regulate generative AI

Part I showed that the conception of AI as simple algorithms contributed to the misconception that data governance measures would address most AI harms, leaving out significant ones caused by generative AI systems. Part II will demonstrate why the initial risk classification framework of the AI Act is ill-suited for generative AI. Section A gives background information on E.U. product safety law and the notion of intended purpose. It then describes all the ways in which the AI Act relies on this notion. Section B presents different types of AI systems in depth and show that generative AI and general purpose systems cannot fit into the initial AI risk classification.

A. Product safety law and the notion of intended purpose

The goal of E.U. product safety law is to achieve a high level of consumer protection by imposing ex-ante safety requirements on manufacturers, providers, importers, and distributors of products made available in the EU. Instead of counting on liability laws to make sure victims of harms are compensated, the E.U. seeks to prevent the harms from happening in the first place. In the past, each country had its own product safety laws, but

[34] 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 (last visited Sep 21, 2023).

[35] Sims, *supra* note 24.

[36] ChaosGPT: Empowering GPT with Internet and Memory to Destroy Humanity, *supra* note 26.

these were 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 is to have uniform requirements so products can move across borders to promote the E.U. market. The AI Act is rooted in classic E.U. product safety law. As such, it imposes safety requirements that AI systems must meet before being deployed. Another critical piece of E.U. law in protecting consumers is the Unfair Commercial Practices Directive (UCPD) which prohibits unfair, misleading, and aggressive commercial practices.

In product safety, the intended 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.[37]

[37] "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 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

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)).

The legal basis of the AI Act 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 E.U. has competence over AI product safety to make sure that rules are consistent across the E.U. to promote the liberal circulation of goods. The AI Act is entirely inspired by E.U. product safety, but as applied to AI systems.[38]

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 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 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 "prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system."[39] 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**

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

[39] AI Act, article 9.

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 cause is individual and based on a data issue or a defect.

Table 3 shows the high-risk AI systems set forth in Annex III of the AI Act. 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. Certain parts were bolded for emphasis.

Table 3. Systems considered high-risk in the AI Act

Area	Intended purpose
Biometrics	Remote biometric identification systems
	AI systems intended to be used for biometric categorisation, according to sensitive or protected attributes or characteristics based on the inference of those attributes or characteristics
	AI systems intended to be used for emotion recognition
Critical infrastructure	AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity
Education and vocational training	AI systems intended to be used to **determine access or admission** or to assign natural persons to educational and vocational training institutions at all levels
	AI systems intended to be used to **evaluate learning outcomes**, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels
	AI systems intended to be used for the purpose of **assessing the appropriate level of education** that an individual will receive or will be able to

	access, in the context of or within educational and vocational training institutions at all levels
	AI systems intended to be used for **monitoring and detecting prohibited behaviour** of students during tests in the context of or within educational and vocational training institutions at all levels
Employment, workers' management and access to self-employment	AI systems intended to be used for the **recruitment or selection of natural persons**, in particular to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates
	AI systems intended to be used to **make decisions affecting terms of work-related relationships**, the promotion or termination of work-related contractual relationships, to allocate tasks based on individual behaviour or personal traits or characteristics or to monitor and evaluate the performance and behaviour of persons in such relationships
Access to and enjoyment of essential private services and essential public services and benefits	AI systems intended to be used by public authorities or on behalf of public authorities to **evaluate the eligibility of natural persons for essential public assistance** benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services
	AI systems intended to be used to **evaluate the creditworthiness** of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud
	AI systems intended to be used for **risk assessment and pricing** in relation to natural persons in the case of life and health insurance
	AI systems intended to **evaluate and classify emergency calls** by natural persons or to be used to dispatch, or to establish priority in the dispatching of, emergency first response services, including by police, firefighters and medical aid, as well as of emergency healthcare patient triage systems
Law enforcement, in so far as their use is permitted under relevant	AI systems intended to be used by or on behalf of law enforcement authorities, or by Union institutions, bodies, offices or agencies in support of law enforcement authorities or on their behalf to **assess**

Union or national law	**the risk of a natural person becoming the victim** of criminal offences
	AI systems intended to be used by or on behalf of law enforcement authorities or by Union institutions, bodies, offices or agencies **in support of law enforcement authorities as polygraphs** or similar tools
	AI systems intended to be used by or on behalf of law enforcement authorities, or by Union institutions, bodies, offices or agencies, in support of law enforcement authorities to **evaluate the reliability of evidence** in the course of the investigation or prosecution of criminal offences
	AI systems intended to be used by law enforcement authorities or on their behalf or by Union institutions, bodies, offices or agencies in support of law enforcement authorities for **assessing the risk of a natural person offending or re-offending** not solely on the basis of the profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680, or to assess personality traits and characteristics or past criminal behaviour of natural persons or groups
	AI systems intended to be used by or on behalf of law enforcement authorities or by Union institutions, bodies, offices or agencies in support of law enforcement authorities **for the profiling of natural persons** as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of the detection, investigation or prosecution of criminal offences
Migration, asylum and border control management, in so far as their use is permitted under relevant Union or national law	AI systems intended to be used by or on behalf of competent public authorities or by Union institutions, bodies, offices or agencies as polygraphs or similar tools
	AI systems intended to be used by or on behalf of competent public authorities or by Union institutions, bodies, offices or agencies to **assess a risk, including a security risk, a risk of irregular migration, or a health risk, posed by a natural person** who intends to enter or who has entered into the territory of a Member State

	AI systems intended to be used by or on behalf of competent public authorities or by Union institutions, bodies, offices or agencies to assist competent public authorities for the **examination of applications for asylum, visa or residence permits** and for associated complaints with regard to the eligibility of the natural persons applying for a status, including related assessments of the reliability of evidence
	AI systems intended to be used by or on behalf of competent public authorities, or by Union institutions, bodies, offices or agencies, in the context of migration, asylum or border control management, for the purpose of **detecting, recognising or identifying natural persons**, with the exception of the verification of travel documents.
Administration of justice and democratic processes	AI systems intended to be used by a judicial authority or on their behalf to assist a judicial authority in **researching and interpreting facts and the law** and in applying the law to a concrete set of facts, or to be used in a similar way in alternative dispute resolution
	AI systems intended to be used for **influencing the outcome of an election or referendum** or the voting behaviour of natural persons in the exercise of their vote in elections or referenda. This does not include AI systems to the output of which natural persons are not directly exposed, such as tools used to organise, optimize or structure political campaigns from an administrative or logistical point of view.

B. *A framework that excludes generative AI*

The fact that the AI Act risk classification relies on the notion of intended purpose carries the assumption of a strong positive correlation between the risk level and the intended purpose of an AI system. This assumption raises two issues—that this correlation does not always hold, and that it fails to account for systems that do not have a prior intended purpose.

1. The weak link between harm and purpose

Most of the AI Act high-risk systems refer to situations in which an AI system carries out an assessment or an evaluation that will inform a decision that is high-stake to the person whose life outcome is being decided. However, the correlation between the area an AI system is used and its related harms does not always hold, especially when it comes to generative AI. 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.[40]

The same is true for the correlation between decision-making and risk. 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.[41] 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.

Moreover, these provisions do not adequately address the systemic risks posed by generative AI. If decision-making tools are biased, it is true that not using them for any consequential decision is a good way to limit their potential harm given how narrow their use is. However, if a generative AI system is biased, it will produce systemic inequity and amplify social power asymmetries, every time it is used, even though it does produce any decision and even in low-stakes situation.

Citing Barocas et al., Katzman et al. distinguishes between allocational harms, "when people belonging to particular social groups are unfairly deprived of access to important opportunities or resources," and representational harms, when "systems produce outputs that can affect the understandings, beliefs, and attitudes that people hold about particular social groups, and thus the standings of those groups within society."[42] Most representational and allocational harms can materialize into economic loss, violence, and emotional or physical injury.

[40] 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.

[41] 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.

[42] Solon Barocas et al., *The Problem with Bias: Allocative versus Representational Harms in Machine Learning*, *in* 9TH ANNUAL CONFERENCE OF THE SPECIAL INTEREST GROUP FOR COMPUTING, INFORMATION AND SOCIETY 1 (2017), https://scholar.google.com/scholar?cluster=5415074729265960442&hl=en&oi=scholarr; Jared Katzman et al., *Taxonomizing and Measuring Representational Harms: A Look at Image Tagging*, 37 PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE 14277 (2023).

Both representational and allocational harms occur in generative AI in relation to gender, race, sexual orientation, religion, nationality, social class, ethnicity, and other characteristics. Many scholars have, for instance, documented such bias in LLM outputs. Abid et al. have found a strong correlation between the words "Muslim" and "terrorist" in "GPT-3 outputs.[43] Analyzing outputs from GPT-2 and BERT, Sheng et al. found bias based on gender, race, and sexual orientation in relation to perceived respectability as well as to professions.[44] Other authors found similar bias associated with professions when prompting both language models (GPT-3.5 and BARD) and image generation systems (Dall-E 2 and Midjourney) about surgeons.[45] Even within a single profession, such as data scientist, Treude & Hata found that different tasks were associated with different genders by language models.[46] In addition, there is a strong division between the Global North and the Global South in terms of GPAIS performance. For instance, AI systems used in agriculture are developed for terrains in the Global North and then sold in the South where they will not lead to optimal crop yields. In addition, the satellite data for most Global North countries has been manually labeled, which is not the case for countries in the Global South.

While representational harms are not specific to generative AI, they are more likely to happen at scale and permeate most areas through such system because generative AI is integrated into so many other systems that it has trickle-down effects. Yet, bias and representational harms are only addressed directly by the AI Act provisions in relation to high-risk systems, with the example of the administration of justice mentioned in the recitals. So despite the EU Commission's initial intent to address algorithmic bias, the AI Act has failed to capture the majority of it.

2. The failure to account for systems with no purpose

Generative AI is part of a category of system that do not have an intended purpose. Therefore, the initial risk classification in the AI Act failed to account for it in any way. General Purpose AI Systems (GPAIS), or systems without a unique predetermined purpose, include foundation models, as well as transfer and meta learning systems, which can be

[43] Abubakar Abid, Maheen Farooqi & James Zou, *Persistent Anti-Muslim Bias in Large Language Models*, *in* PROCEEDINGS OF THE 2021 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 298 (2021), https://doi.org/10.1145/3461702.3462624.

[44] Emily Sheng et al., *The Woman Worked as a Babysitter: On Biases in Language Generation*, (2019), http://arxiv.org/abs/1909.01326 (last visited Jan 19, 2024).

[45] Jevan Cevik et al., *Assessment of the Bias of Artificial Intelligence Generated Images and Large Language Models on Their Depiction of a Surgeon*, n/a ANZ JOURNAL OF SURGERY, https://onlinelibrary.wiley.com/doi/abs/10.1111/ans.18792 (last visited Jan 19, 2024).

[46] Christoph Treude & Hideaki Hata, *She Elicits Requirements and He Tests: Software Engineering Gender Bias in Large Language Models*, *in* 2023 IEEE/ACM 20TH INTERNATIONAL CONFERENCE ON MINING SOFTWARE REPOSITORIES (MSR) 624 (2023), https://ieeexplore.ieee.org/document/10174028.

adapted to undertake new tasks with minimal effort. Box 1 presents relevant AI definitions.

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.[47]

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.[48]

Generative AI: AI systems that generate outputs more complex than a number, label, or recommendation (e.g., text, audio, video, images).[49]

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.[50]

Multi-modal AI: an AI system where the input or output includes more than one modality (e.g., images, video, audio, text, time-series).[51]

Transfer and meta-learning systems: systems designed to acquire a new capability with minimal additional learning.[52]

Foundation models are designed to conduct a broad variety of tasks.[53] 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 generative AI 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

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

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

[49] This definition was developed by the authors of this paper.

[50] This definition was initially developed by Claire Boine and Richard Mallah and subsequently published in Gutierrez, Carlos I., Anthony Aguirre, Risto Uuk, Claire C. Boine, and Matija Franklin. "A Proposal for a Definition of General Purpose Artificial Intelligence Systems." *Digital Society* 2, no. 3 (September 12, 2023): 36. It is the one endorsed by the authors of this paper. However, the AI Act presents an alternative definition of General Purpose AI systems as "an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems."

[51] This definition was developed by the authors of this paper.

[52] This definition was developed by the authors of this paper.

[53] Bommasani et al., *supra* note 48.

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.[54] 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 causal links in multicausal situations, detecting logical fallacies, understanding fables, and producing code for computer programs.[55] 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. 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.[56] Figure 1 shows different capabilities acquired at different levels of parameters.

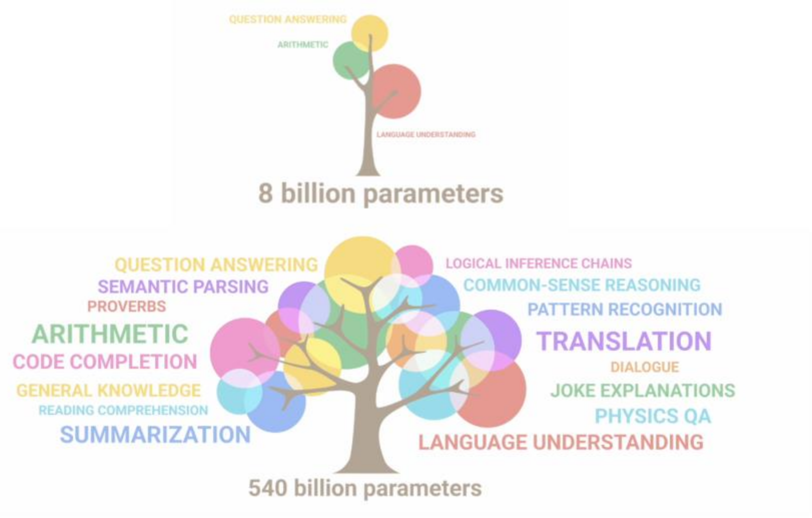
Figure 1. A visual representation of emergent capabilities (source: Narang, Sharan, and Aakanksha Chowdhery)[57]

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

[55] 137 emergent abilities of large language models, JASON WEI, https://www.jasonwei.net/blog/emergence.

[56] 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.

[57] 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.



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. To demonstrate how LLMs can be used for unexpected purposes, a group of researchers used one to reproduce the COMPAS recidivism prediction scores.[58] 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.

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 E.U. Commission and the public have been

[58] 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.

moving away from ascribing too much autonomy or agency to AI systems in their proposed definitions, GPAIS exhibit increasingly autonomous and agentic behaviors. Chan et al. identify four characteristics that determine how agentic an AI system is: 1) : "the degree to which the algorithmic system can accomplish a goal provided by operators or designers, without a concrete specification of how the goal is to be accomplished;" 2) "the degree to which the algorithmic system's actions affect the world without mediation or intervention by a human;" 3) "the degree to which the system acts as if it is designed/trained to achieve a particular quantifiable objective;" and 4) "the degree to which the algorithmic system is designed/trained to make decisions that are temporally dependent upon one another to achieve a goal and/or make predictions over a long time horizon."[59] They demonstrate that there is an increase in the deployment of increasingly agentic systems by companies like Google, Amazon, Spotify, Youtube and Meta.[60] Not all agentic systems are generative, but generative AI is also increasingly agentic. LLMs can now make entire plans based on a prompt. Today, a range of work is adapting LLMs to make them able to act in the world.[61] For instance, ACT-1 can undertake the prompt "find me a house for four people in Houston."[62] Integrated into robots, AI systems can also create a multi-step plan and act on it.[63] 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.[64] In a follow-up study to evaluate the level of autonomy of several systems, the same

[59] Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, (2023), http://arxiv.org/abs/2302.10329.

[60] *Id.*

[61] Richard Ngo, *Visualizing the Deep Learning Revolution*, MEDIUM (May 5, 2023), https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5.

[62] *Id.*

[63] *Id.*

[64] Elizabeth Barnes, *Update on ARC's Recent Eval Efforts - ARC Evals*, (2023), https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

research team got GPT-4 to make a plan to secure a stranger's log in information and carry it out.[65]

Chan et al. present potential harms from increasingly agentic systems. The first category is composed of delayed systemic harms. These include environmental risks, concentration of power, privacy infringements, fairness implications of decisions, financial risk, racism, misogyny, mental health issues, the amplification of political polarization, the spread of fake news and the manipulation of users' internal states.[66] The second category of harm has to do with collective disempowerment. It includes the diffusion of power away from humans and the exacerbation of the concentration of power among a coding elite.

III. THE ADDITION OF GENERATIVE AI IN THE AI ACT

A. *Shoehorning General purpose AI into the AI Act*

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

[65] Megan Kinniment et al., *Evaluating Language-Model Agents on Realistic Autonomous Tasks*, (2024), http://arxiv.org/abs/2312.11671 (last visited Jan 18, 2024).

[66] Chan et al., *supra* note 59.

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).[67] 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 the other order. In the same way, high-risk systems are supposed to achieve a high level of accuracy, robustness, and cybersecurity through their lifecycle, even though these metrics depend on what they are used for.[68] 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."[69] 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. It is thus impossible for providers of GPAI models to comply with all possible high-risk requirements so that the ensuing systems would comply.

However, it seems difficult for downstream users to comply as well. Currently, according to the Act, any person will be considered the provider of a high-risk system "if they modify the intended purpose of an AI system, including a general-purpose AI system, which has not been classified as high-risk and has already been placed on the market or put into service in such a way that the AI system concerned becomes a high-risk AI system."[70] For instance, if a school in the E.U. starts using an LLM to evaluate their students, they become considered a provider of a high-risk system and must comply with all requirements set forth for such systems, human oversight, robustness, including the establishment of a risk management system, governance measures, transparency requirements, record-keeping, and maintaining the technical documentation, etc.[71] It is highly

[67] AI Act, Article 9.
[68] AI Act, Article 15.
[69] Article 10 of the AI Act.
[70] AI Act, Article 25.
[71] AI Act, Articles 8-15.

unrealistic to expect downstream users who do not have the required training, information, resource, or technical expertise to be able to meet these requirements.

For instance, a GPAIS that was not specifically trained to be deployed in a certain context will not necessarily achieve the level of accuracy required by 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.[72] 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 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 E.U. (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.

While the initial creation of a legal framework that completely failed to include generative AI, despite the state of technology at the time can seem surprising, this is a typical case of path dependence. 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.[73] While GPT-3 had already come out in June 2020 when the E.U. Commission released the proposed AI Act in April 2021, 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: "[t]he 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

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

[73] 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.

significant risks, in particular from the viewpoint of protection of safety, consumer rights and fundamental rights."[74] 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. In his paper on the regulation of artificial intelligence (AI), Matthew U. Scherer wrote that: "[t]he 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."[75]

Another interpretation of the failure of the E.U. to adapt their proposed legislation to generative AI is the sociotechnical construction argument. Meg L. Jones argues that the law as a social and linguistic system constructs the meaning of new technologies and their uses with policy consequences. She writes that "[n]ot only does law not linearly follow technology, a great deal of legal work shapes technology and the way in which it will be understood in the future."[76] In this line of thinking, one could argue that E.U. policymakers constructed a definition of AI as statistical software that made sense to them in the already existing product safety law context.

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 GPAIS and generative AI. Given that the text was not drafted in a technology-agnostic manner and entirely built around end uses and intended purposes, adding these provisions was like trying to fit a square peg into a round hole.

There are three categories of additional relevant provisions that apply to GPAIS and generative AI in the AI Act. The first category is about all GPAI models (and not systems) although some of the provisions do not apply to those released under a free and open-source license. The second is for generative AI systems specifically. The third is for GPAIS classified by the AI Act as posing *systemic risks*.

The first category of provisions applies to providers of General Purpose AI models. The E.U. AI Act defines a GPAI model as such:

[74] EUROPEAN COMMISSION, *On Artificial Intelligence - A European Approach to Excellence and Trust*, (2020).

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

[76] Meg Leta Jones, *Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw*, (2017), https://papers.ssrn.com/abstract=2981855 (last visited Jan 24, 2024).

> AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.[77]

In short, the E.U. places the burden onto providers such as Google, OpenAI, Anthropic, and other big labs rather than onto European downstream deployers. Most of the obligations are transparency and record-keeping measures that downstream deployers would not have the necessary information to comply with. Providers must draw up two different kinds of technical documentation. One is to be kept at hand in case it is requested by the AI Office or national authorities, while the other one is for downstream providers that with to integrate the model into their AI system. The documentation for the authorities must contain various pieces of information including "the design specifications of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made; what the model is designed to optimise for and the relevance of the different parameters," "information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies (e.g. cleaning, filtering etc.), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases," "the computational resources used to train the model (e.g. number of floating point operations), training time, and other relevant details related to the training" and "known or estimated energy consumption of the model."[78] The documentation for downstream deployers must contain various elements including the acceptable use policies applicable, how the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, the architecture and number of parameters, the modality (e.g. text, image) and format of inputs and outputs and information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies.[79] Notably, it is supposed to include a description of "the tasks that the model is intended to perform,"[80] reminiscing of an intended purpose. The AI Act also includes a provision to address copyrights issues. It must be possible for authors to express a reservation of rights, so their material is not used to train AI systems. GPAI models providers also have

[77] AI Act, Article 3.
[78] AI Act, Annex XI.
[79] AI Act, Annex XII.
[80] *Id.*

"to draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model."[81]

The AI Act also lays out obligations that are specific to generative AI. For instance, "general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated."[82] AI systems that generate synthetic content, such as deepfakes, must also clearly label their outputs as artificially produced. Whenever an AI system generates or modifies content, this must be disclosed to users, unless the content is for legal purposes or falls within artistic or satirical contexts. The AI Office will collaborate on developing guidelines for identifying and labeling artificially generated content.

The AI Act also introduces obligations for AI systems carrying potential "systemic risks." A "systemic risk means a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain."[83] An AI model is considered as having systemic risk if the AI model has high impact capabilities, as assessed by technical tools and benchmarks, if it has similar capabilities or impact as decided by the Commission, or if the cumulative amount of computation used for its training measured in floating point operations was greater than 10^{25}.[84]

Providers of GPAI models with systemic risks have four main obligations.[85] First, they must evaluate AI models using state-of-the-art protocols and tools. This includes adversarial testing to identify and mitigate systemic risks. Second, they must assess and mitigate systemic risks from deploying and using these systems at the E.U. level. Third, they are required to document, track, and report serious incidents and corrective actions to the AI Office and relevant national authorities without undue delay. Finally, they must ensure the AI models, and their physical infrastructure, have adequate cybersecurity protection.

This current governance model poses significant issues. The first one is the absence of substantive regulation for most generative AI systems. The E.U. AI Act has created two risk-based parallel regimes. The first one is a risk classification based mostly on the sector an AI system is used in. That one implies that risk is correlated with area of application. This regime offers the most substantive safety requirements. However, this regime

[81] AI Act, Article 53.1(d).
[82] AI Act, Article 50.
[83] AI Act, Article 3.
[84] AI Act, Article 51.
[85] AI Act, Article 55.

would only apply to GPAI and generative systems deployed in the narrow use-cases considered high-risk. And even if they were deployed in those contexts, it is highly unlikely that their deployers could comply with the requirements.

At the same time, there is another set of obligations for providers of GPAI models that are considered to have high levels of capability. A GPAI model is assumed to have such level of capability if it was trained on more than 10^{25} flops. Therefore, this other regime relies on the assumed correlation between capability—captured imprecisely through the flops level—and risk. Currently, most generative AI systems on the market are below this threshold, except for GPT-4. Thus, most generative AI systems are neither considered high risk according to the AI Act nor above the flops threshold. Except for the obligation to label synthetic data as such, these generative AI systems are not subject to substantial safety requirements beyond mere record keeping duties on the part of their providers.

Another issue in this regime has to do with the fact that open-source GPAI models are not subject to the record-keeping obligations. Some generative AI systems 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. 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 an advantage, and that the open-source community is getting ahead.[86] 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 targeted algorithms to achieve similar results without incurring the same costs through brute force. Large companies then learn from the research and code published online by the open-source community. They benefit from a highly skilled workforce for free.[87] They can use and learn from the research outputs and software published online; but also have the resources to take them further internally. Simultaneously, some academics and researchers are asking to restrict open-source models because they believe those pose serious dangers. For instance, many systems available for free can be used for criminal purposes such as to malware or biochemical weapons. Yet, those releasing them will not be subject to the highest standards of transparency and disclosure.

[86] 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.

[87] Will Knight, *The Myth of 'Open Source' AI*, WIRED, https://www.wired.com/story/the-myth-of-open-source-ai/.

B. Improving the governance of generative AI

1. The relation between capability and risk

Several countries have enacted standards based on the assumption that systems that are generally more capable are also riskier. Policymakers in the U.S. and the E.U. have adopted a threshold of floating points operations (flops)—respectively 10^{26} and 10^{25}—to categorize GPAIS as higher risk.[88] This choice contains two assumptions: that flops and capabilities are highly correlated, and that capability and risks are highly correlated. Therefore, the number of flops used to train a system is correlated with the system's risk level. The question of a relation between capabilities and risk can be divided into two components: 1) whether a more generally capable system is correlated with more risk; 2) whether certain specific capabilities of an AI system are correlated with more risk.

AI systems are made of three ingredients: data (for which both quantity and quality matter), the model (which refers to the choice of algorithm, the architecture of the system and the number of parameters), and computing power (which can be quantified in the number of flops used to train the system). Research has shown that increasing the size of the dataset, the number of parameters, or the computing power used to train an AI system lead to significant increase in performance and generalization abilities, a phenomenon labeled the "scaling laws."[89] Diaz and Madaio have contested the scaling laws when it comes to data, showing that as the size of the dataset increases, the performance of the AI system might become worse for certain communities.[90] On average, practitioners have found that scaling one of the AI ingredients leads to AI systems acquiring emergent abilities that were not foreseen by their producers,[91] although such emergent capabilities have sometimes been overestimated.[92] As a result, it

[88] While the E.U. was adopting the 10^{25} threshold that only includes GPT-4, the U.S. was adopting the 10^{26} that did not include any current system. See Joseph R. Biden, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, EO 14110 (2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[89] Wei, *supra* note 55; Aakanksha Chowdhery et al., *PaLM: Scaling Language Modeling with Pathways*, ARXIV:2204.02311 [CS] (2022), http://arxiv.org/abs/2204.02311 (last visited Apr 12, 2022).

[90] Fernando Diaz & Michael Madaio, *Scaling Laws Do Not Scale*, (2023), http://arxiv.org/abs/2307.03201 (last visited May 27, 2024).

[91] Sanjeev Arora & Anirudh Goyal, *A Theory for Emergence of Complex Skills in Language Models*, (2023), http://arxiv.org/abs/2307.15936 (last visited Mar 18, 2024); Ganguli et al., *supra* note 58; Wei, *supra* note 55.

[92] Rylan Schaeffer, Brando Miranda & Sanmi Koyejo, *Are Emergent Abilities of Large Language Models a Mirage?*, (2023), http://arxiv.org/abs/2304.15004 (last visited Jun 9, 2023).

is assumed that the computing power used to train an AI system correlates with its capability level.

The International Standard Organization considers that "AI systems have a spectrum of risk, determined by the severity of the potential impact of a failure or unexpected behavior."[93] Relevant factors to assess the level of risk of an AI system include:

- the type of action space the system is operating in (e.g. recommendations vs direct action in an environment)
- the presence or absence of external supervision
- the type of external supervision (automated or manual)
- the ethical relevance of the task or domain
- the level of transparency of decisions or processing steps
- the degree of system automation.[94]

It is likely that a more generally capable AI system operates in a wider action space with more automation. It is also less likely to be subject to high levels of human supervision given that: 1) a higher proportion of a task or plan might be automated and conducted by a single system, reducing the number of points of potential control; and 2) as AI systems become more capable, they are more likely to be used to supervise and control other systems. It is also possible that more capable systems will be deployed in wider use cases and higher stakes contexts, making their potential mistakes more harmful. Moreover, as the level of capability of GPAIS increase, they become more complex, worsening the black box effect and making their level of transparency lower.[95] Finally, if a highly capable system acquires dangerous goals, for instance through goal misspecification, it could be capable of achieving it and causing widespread harm.[96] It is thus sensible to expect the number of AI accidents to increase as more capable systems are deployed. In addition, more capable GPAIS also increase the risk of intentional harm as AI systems that are more capable make AI-assisted crimes easier (e.g. phishing and scamming, attacks on cyber and real infrastructures, etc.).[97] It is also expected that certain specific capabilities (e.g. reasoning or planning) increase the risk posed by AI systems because it might both increase the number of actions an AI system can take and decrease the level of supervision they receive.

[93] INTERNATIONAL STANDARD ORGANIZATION, *Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology*, (2022).

[94] *Id.*

[95] PASQUALE, *supra* note 14.

[96] Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, DEEPMIND (Apr. 21, 2020), https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

[97] CHECK POINT RESEARCH, *OPWNAI : Cybercriminals Starting to Use ChatGPT*, (2023), https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/ (last visited Mar 13, 2024); ChatGPT - the impact of Large Language Models on Law Enforcement, EUROPOL, https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement; Kinniment et al., *supra* note 57.

It is thus important to both increase and improve risk prevention measures for more capable systems. However, many experts agree that a flops threshold is an imperfect measure. First, scaling other ingredients such as the number of parameters and the size of the dataset can also lead to highly capable systems. Second, this threshold can drive certain companies to build more capable systems while avoiding reaching the number of flops that would lead to their models being regulated. For instance, they might create systems from multiple smaller models as opposed to a single bigger one. In fact, this is currently one of the most promising ways to make more capable and more agentic systems. Research shows that you can improve the capabilities of LLMs by combining them with other AI models and using programs built around them. For instance, using a software framework (called a scaffolding program) that supports and guides the learning, reasoning, and decision- making processes of an AI agent shows promising results. A scaffolding program can do the following: 1) task decomposition (breaking down complex, high-level tasks into smaller, more manageable subtasks that the AI agent can solve independently); 2) prompt engineering (generating optimal prompts for the AI agent); 3) memory retention (maintaining a history of actions, decisions, and results); and 4) coordination (managing the interaction between the AI agent and other components of the system, such as external APIs, databases, or simulated environments). In the study referred to earlier in which GPT-4 created and acted on a phishing plan, a scaffolding program was used to make calls to the LLM API and to run code in a virtual machine.[98] Another approach to create AI systems capable of reasoning and planning is to combine symbolic-based models such as solvers and verifiers with LLMs. For example, Kambhampati et al. built an LLM-Modulo Framework in which an LLM provides broad approximate knowledge of the problem domain, while the external verifiers bring in formal reasoning capabilities and help ensure the correctness of the generated plans or solutions.[99]

The current definition of GPAI models with systemic risk is thus under-protective of consumers.

2. Addressing systemic issues

The fact that GPAI models and not systems are regulated in the AI Act is problematic. AI systems are sociotechnical compounds that encompass not only the AI model but also all the components required to deploy and operate the model in a real-world environment. This includes data pipelines, user interfaces, hardware infrastructure, and any other software or tool. Importantly, it also includes the ways humans interact with the AI,

[98] Kinniment et al., *supra* note 65.

[99] Subbarao Kambhampati, *Can LLMs Really Reason and Plan?*, https://cacm.acm.org/blogs/blog-cacm/276268-can-llms-really-reason-and-plan/fulltext.

such as through user interfaces or feedback mechanisms. An AI system can even contain several AI models interacting with one another.

There are different ways a generative AI system can be harmful. First, an AI system can be harmful from not performing like is expected. This is what lawyers might consider a defect. Yet, it is almost impossible to foresee how an AI system will behave because generative AI is highly unpredictable. A same prompt might lead to different outputs at different times. An AI system might also perform very well in a certain context and not in another one, so it would be harmful when introduced in a new context. As a result, it is impossible for a generative AI company to foresee how a system will perform in all the contexts it will be deployed and with all the prompts it will receive. Therefore, ensuring that generative AI systems are safe requires continuously monitoring the way they interact with users and what impacts they are causing, even and especially after deployment. It involves stress testing systems in different contexts and for outliers. This is true for all generative AI systems, and not solely the ones trained on more than 10^{25} flops.

Second, an AI system might also be harmful because it is performant and being used in malicious ways. While generative AI systems do not create new harms in that regard, they make it possible to automate and scale up pre-existing ones. For instance, while troll farms have existed for a long time, all the steps of disinformation campaigns can now be automated, making it possible to scale them up to unprecedented levels.[100] Other criminal activities such as fraud, targeted fishing and malware deployment can be automated and scaled up as well. This is the case with current generative AI systems trained on less than 10^{25} flops. Preventing these harms requires conducting risk assessments including red teaming exercises meant to see if it is possible to use these systems in harmful ways.[101]

Third, AI systems might be harmful by virtue of being used regardless of their performance. This is the case, for instance, if they create overreliance and lead to human losing skills or meaning, if they lead to exploitative labor practices, or if they cause environmental impacts. This can happen whether the AI system has been trained over more than 10^{25} flops or not. Assessing these harms can only be done at society's level and in collaboration with users and civil society. This requires governments to

[100] CENTER FOR SECURITY AND EMERGING TECHNOLOGY ET AL., *AI and the Future of Disinformation Campaigns: Part 1: The RICHDATA Framework*, (2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/ (last visited Jun 23, 2024).

[101] ANTHROPIC, *Frontier Threats Red Teaming for AI Safety*, (2023), https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety; CHRISTOPHER A. MOUTON, CALEB LUCAS & ELLA GUEST, *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach*, (2023), https://www.rand.org/pubs/research_reports/RRA2977-1.html).

step up, coordinate such dialogue and assessment and adopt policies that can counteract these social harms.

In all cases, it is the systems that should be assessed: including the interactions between the different models, the scaffolding, any relevant interface, and the interaction with humans. It is our hope that the EU Commission, pursuant to articles 51 and 97, will adopt new thresholds and rules to include all generative AI systems into the systemic risk category.

3. Required disclosures

As we have seen, it is impossible to predict all possible harms from generative AI systems, since they are inherently unpredictable and that their output is highly context-dependent. Therefore, a safety approach that is exclusively pre-market will necessarily fail. This is why certain disclosures from providers and deployers of AI systems should be mandated on a regular basis. These should include all the incidents that incurred with the system and all the hand patches done.

For the incident disclosure, the victims of certain harms from AI systems should have a way to file an incident report with their national body to open an investigation. The types of incidents that must be reported should include defect-related harms, criminal harms, and societal-level harms. Given that AI systems can deeply affect society through small effects on many individuals, it is essential to include such incidents in the database. The incident database should be made public to enable academics and civil society to analyze it 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.

Hand patches should also be released publicly by AI companies. 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. 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.[102] 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. For providers to simply hand patch their systems would not truly fix the issue. This is why providers of generative AI systems 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.

CONCLUSION

There is much to applaud about the AI Act, which represents the first substantial attempt by an advanced economy to regulate the complex challenges raised by artificial intelligence technologies. But like all good first drafts, there is much that can be improved upon reflection. It may well be too late for the E.U. to learn from the mistakes of the AI Act. However, in this case, Europe's loss may well be the rest of the world's gain.

What we have learned from the AI Act is that it is essential to build a legal regime that is adapted to generative AI and general-purpose AI systems. As the United States, Canada, and other jurisdictions follow in the footsteps of the EU, this paper has sought to argue that it is essential to reconsider how we define AI and to recognize that the potential harms of AI extend far beyond flawed datasets. Particularly with respect to generative and general-purpose AI technologies, it is important to reject an EU-style product safety approach and to acknowledge that these systems often lack a prior intended purpose, making traditional risk classification frameworks inadequate.

The project of regulating AI will be a lengthy one; one that will not be solved by a single enactment from any jurisdiction, no matter how well-meaning. It is a testament to the rapidly evolving nature of AI that the E.U. AI Act is in some senses obsolete even before it has come into effect. Future regulatory efforts should focus on regulating systems over models, implementing universal risk assessments and red-teaming exercises for all generative AI systems, and establishing robust disclosure requirements for

[102] 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/.

incidents and interventions. Only by learning from the ways that the AI Act has failed to succeed can future AI laws have any chance of effectively addressing the complex challenges posed by generative AI.