Assessing short-term and long-term risks of general-purpose artificial intelligence

An interdisciplinary approach to systemic risk assessment in the Artificial Intelligence Act

Author: D.F.P. (Daniël) de Weerd
Email address: daniel.de.weerd@student.uva.nl
Supervisor: prof. dr. J.V.J. (Joris) van Hoboken
Master track: Informatierecht

Word count: 13675 words
Submission date: 28/06/2024

UNIVERSITEIT VAN AMSTERDAM

Abstract

Chapter V of the Artificial Intelligence Act ("AIA") introduces a standalone regulatory regime for "general-purpose AI models" that is intended to regulate large language models. Providers of general-purpose AI models with "high-impact capabilities" are obliged to assess and mitigate systemic risks that may stem from their development, placing on the market, or use. This study aims to establish what kind of risks would qualify as such a "systemic risk". To this end, it combines a literature survey on the risks of large language models with a legal analysis of relevant provisions in the Artificial Intelligence Act. It further incorporates an analysis into the applicability of proposed systemic risk assessment frameworks for the Digital Services Act to the Artificial Intelligence Act.

Scientific work on the risks of large language models tends to distinguish between their short-term and their long-term risks. Short-term risks are mostly related to issues such as fairness, trust, accountability and sustainability, while long-term risks are mostly related to AI models causing harm through their advanced capabilities. The two perspectives are commonly viewed as competing rather than complementary approaches, with authors focusing on short-term risks often deriding long-term concerns as a bad-faith distraction. This study highlights bias, hallucination sustainability impact and disruption of democratic processes as key short-term risks and highlights knowledge barrier reduction and self-proliferation as key long-term risks.

This work further analyses the AIA's provisions on systemic risk in the light of its aims and legislative history. While Chapter V of the AIA provides a useful definition of a "general-purpose AI model", its definitions of "high-impact capabilities" and of systemic risk as "specific to" such capabilities remain unclear and raise questions on the scope of its systemic risk concept. Furthermore, the text of the AIA provides very little guidance on what exactly should be considered "systemic risks", with the main text providing no examples or risk categories at all and the recitals only providing a less-than-clear list of examples.

To inform our interpretation of the AIA's provisions on systemic risk, this work further studies proposed assessment frameworks for systemic risks relating to very large online platforms and search engines under the Digital Services Act. It argues that, while some proposed frameworks offer little to no substantive insights, work that extends concepts from financial systemic risk to platform regulation can be useful in distinguishing between sources of risk and the spread of risk. This argument, which states that platform systemic risk can be expressed in terms of propagated shocks, is viewed as potentially applicable to systemic risks of general-purpose artificial intelligence.

Finally, this work proposes capability, alignment and propagation as systemic risk assessment criteria for general-purpose artificial intelligence that can form the basis for a sociotechnical evaluation of large language models' systemic risks. It argues that, in addition to technically assessing their models' potential for harmful behaviour, GPAI model providers should assess the ways in which harm caused by such behaviour could propagate through society. Based on this framework, it concludes that risks relating to manipulation of democratic processes and knowledge barrier reduction are likely to qualify as "systemic risks" in the Artificial Intelligence Act, risks related to bias & discrimination, hallucination and self-proliferation could potentially qualify as such, and sustainability risks are unlikely to qualify as "systemic risks".

Table of Contents

1. Introduction

1.1 Problem statement

The EU Artificial Intelligence Act ("AIA") is world's first comprehensive approach to artificial intelligence ("AI") regulation, aiming to address AI's risks and harms to fundamental rights and public interests.[1] In addition to establishing a tiered regulatory regime in which AI systems are classified as having unacceptable, high or minimal risks based on their intended purpose, the AIA introduces a separate regime for general-purpose AI ("GPAI") models. This regime is intended to regulate the family of advanced language models known as large language models ("LLMs") that have seen a surge in adaptation since late 2022. While providers of GPAI models in general merely face some transparency obligations, providers of the most advanced such models are obliged to assess and mitigate systemic risks that may stem from their development, placing on the market, or use.[2]

Much remains unclear about the nature of the systemic risks that providers of general-purpose AI models should assess. As the regulatory regime for general-purpose AI models with systemic risk was only introduced in the final stretch of the legislative process, little official documentation exists on its key provisions' meaning or rationale. Similarly, academic legal work on the subject remains sparse. However, a wide body of academic and non-academic research exists on large language models and their associated risks. Drawing on such research could provide an original and relevant contribution to legal research on the AIA's regime for systemic risks associated with general-purpose AI models and could contribute to the upcoming codes of practice for GPAI model providers, which will include guidance on "the identification of the type and nature of the systemic risks at Union level".[3]

This work aims to offer such a contribution by combining a study of risks related to large language models with a legal doctrinal study into the Artificial Intelligence Act's relevant provisions. First, it studies the key technical properties of large language models and their associated risks, describes the often-made distinction between short-term and long-term artificial intelligence risks and highlights some key risks. Next, it sets out AIA's regulatory regime for general-purpose AI models in the light of its aims, legislative history and technical

[1] Regulation (EU) 2024/XXX of the European Parliament and of the Council of x month 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
[2] Article 55(1)(b) AIA.
[3] Article 56(2)(c) AIA.

background and further analyses relevant work on the systemic risk provisions of the Digital Services Act ("DSA"). Finally, it proposes a set of assessment criteria based on the preceding analysis that can be used to determine whether an LLM-related risk qualifies as a "systemic risk".

1.2 Research questions and methodology

This work uses the following main research question and sub-questions:

What risks associated with large language models could qualify as "systemic risks at Union level" under Article 55 of the Artificial Intelligence Act?

1. What are some risk factors and risks of large language models identified in scientific literature on safe and ethical artificial intelligence?
2. What requirements for a risk to qualify as a "systemic risk" can be derived from:
    a. The systemic risk regime imposed on providers of general-purpose AI model providers by the Artificial Intelligence Act;
    b. The systemic risk regime imposed on providers of very large online platforms and search engines by the Digital Services Act;
3. Based on the risk factors and requirements identified, what set of criteria can be developed to assess whether a risk associated with large language models should qualify as a "systemic risk at Union level"?
4. Applying the set of criteria developed, which of the key risks of large language models identified could potentially qualify as "systemic risks at Union level"?

The first sub-question will be answered through a literature survey on large language models and their associated risks. It is explicitly not intended to provide an exhaustive taxonomy of LLM-associated risks, but it rather seeks to give some context for the systemic risk analysis. Any harms best discussed in relation to other existing legal frameworks, such as risks relating to copyright, privacy or competition, are excluded from the scope of this research.

The second sub-question will be answered through a legal doctrinal investigation into the concept of "systemic risk" based on the relevant frameworks from the Artificial Intelligence Act and the Digital Services Act. It is in part descriptive, focusing on the text's key provisions, and in part evaluative, incorporating research on safe and ethical artificial intelligence and the

aims of the AIA as a normative framework to evaluate possible interpretations. The research question is of a legal nature, but makes use of insights gained from technical literature – what Schrama would describe as "unilateral interdisciplinary research".[4]

The third and fourth sub-questions will be answered by building on the analysis conducted for the first two sub-questions, identifying key overarching considerations, synthesizing them into a set of criteria that makes sense in the light of the risk factors and legal requirements derived, and applying the developed set of criteria to identified risks.

1.3 A note on (non-)academic sources and preprints

Most academic work on ethical AI and on the Artificial Intelligence Act is found in peer-reviewed journals and conference proceedings. A significant portion of work in the field of large language models' development and safety is however published through pre-prints and policy reports. Although such sources are not peer-reviewed, they account for some of the most influential work in artificial intelligence research and should be included in any work discussing the field's state of the art. Nevertheless, the nature of such sources should be made clear to the reader when citing them. Doing so is especially important when citing preprints or reports of which all authors are affiliated with the same organization, which in practice is often an AI developer or think tank with a material stake in the regulatory process.

To clarify the origin of pre-prints and reports, I cite them in a way that slightly deviates from the OSCOLA standard. Preprints of which all authors are affiliated with an academic institution are cited like journal articles, but with a clarification attached marking their status as preprints.

For example:

> Rishi Bommasani, Percy Liang and others, 'On the Opportunities and Risks of Foundation Models' [2023] *Preprint* <https://arxiv.org/pdf/2108.07258>

[4] Wendy Schrama, 'How to carry out interdisciplinary legal research: Some experiences with an interdisciplinary research method' [2011] 7(1) Utrecht Law Review 152

Preprints where authors are all affiliated with industry parties or think tanks are cited as policy reports mentioning their institution:

> Jared Kaplan, Sam McCandlish and others, 'Scaling Laws for Neural Language Models' (OpenAI, 2020) <https://arxiv.org/abs/2001.08361>
>
> Dan Hendrycks, Mantas Mazeika & Thomas Woodside, 'An Overview of Catastrophic AI Risks' (Center for AI Safety 2023) <https://arxiv.org/pdf/2306.12001>

This ensures both that relevant non-academic sources are included and that the reader is able to distinguish between academic and non-academic sources.

2. Conceptual and legal frameworks

This chapter outlines key conceptual and legal frameworks relevant to this work's analysis. It defines some key concepts related to general-purpose artificial intelligence and briefly outlines the regulatory framework of the Artificial Intelligence Act.

2.1 Key definitions related to general-purpose artificial intelligence

The field of artificial intelligence ("AI") can be understood as the study of understanding and building intelligent entities.[5] Since the inception of the field in the 1950's, AI researchers have made remarkable progress in developing systems that display superhuman performance on a large number of distinct tasks.[6] Given the massive economic incentives that come with automating tasks that would normally require human intellectual effort, the list of tasks that AI beats humans at grows by the day.[7]

Despite artificial intelligence approaches' success in outperforming humans on a wide range of tasks, their capabilities tend to remain limited to clearly specified domains and tasks. Most if not all existing artificial intelligence systems are what researchers call "narrow AI": systems designed to solve specific problems without the ability to generalize their capabilities. An illustrative example would be the Stockfish chess engine, which outperforms any human in determine a winning chess move from a valid set of inputs but has zero intelligent capabilities apart from evaluating chess positions. This limitation is inherent to nearly all artificial intelligence approaches that have been developed to date.

In contrast to narrow AI, researchers have long speculated about the possibility of creating artificial general intelligence ("AGI").[8] An AGI would be a human-made intelligent system "that can solve a variety of complex problems in a variety of different domains, and that controls itself autonomously, with its own thoughts, worries, feelings, strengths, weaknesses and

[5] Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach*, *Fourth Edition* (Pearson 2021), p. 1.
[6] Ibid, para. 1.4.
[7] Robin Hanson, 'Economic Growth Given Machine Intelligence' [2001] <http://mason.gmu.edu/~rhanson/aigrow.pdf> accessed May 6 2024
[8] Irving John Good, 'Speculations Concerning the First Ultraintelligent Machine' [1966] 6 Advances in Computers 31 88

predispositions".[9] The superlative of AGI would be an artificial superintelligence ("ASI"), an hypothetical artificial entity whose capabilities far surpass human intelligence.

Large language models ("LLMs") are artificial intelligence models with advanced language modelling capabilities that enable them to perform a large number of distinct tasks.[10] Since the 2022 release of OpenAI's LLM-powered web application ChatGPT, large language models have seen a major surge in adaptation. This surge has led to additional academic and regulatory scrutiny. While some AI developers argue that – their – LLMs constitute an early form of artificial general intelligence, such models are more often referred to by the term "general-purpose artificial intelligence" ("GPAI").[11]

Although the concepts of artificial general intelligence and general-purpose artificial intelligence might seem indistinguishable, the contexts in which they are used are markedly different. The phrase "AGI" is mostly used in relation to artificial superintelligence and existential risk, while the phrase "GPAI" is the customary term used in regulatory discussions on present-day large language models. While it is clear that present-day LLMs in no way qualify as superintelligence, whether they qualify as "AGI" is arguably a semantic discussion that adds little value to the debate on AI regulation.

2.2 The legal framework of the Artificial Intelligence Act

The Artificial Intelligence Act ("AIA") is the European Union's flagship approach to artificial intelligence regulation that seeks to promote the uptake of human-centric and trustworthy artificial intelligence while ensuring a high level of protection for health, safety and fundamental rights.[12] First proposed in April 2021, it saw significant changes between the original proposal and the final corrigendum as released in April 2024.[13]

[9] Cassio Pennachin and Ben Goertzel, 'Contemporary Approaches to Artificial General Intelligence' in Ben Goertzel and Cassio Pennachin (Eds.), *Artificial General Intelligence* (Springer 2007)
[10] Tom Brown and others, 'Language Models are Few-Shot Learners', 33rd Conference on Neural Information Processing Systems (NeurIPS 2020).
[11] On AGI claims see: Sébastien Bubeck and others, 'Sparks of Artificial General Intelligence: Early experiments with GPT-4' (Microsoft Research, 2023);
[12] Recital 1 AIA.
[13] Corrigendum to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/ ..... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 19 April 2024.

The AIA was proposed as part of the Commission's strategy on artificial intelligence as outlined in its White Paper on AI, which sets the "twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of such technology".[14] It is further based on the Ethics Guidelines for Trustworthy AI ("the Guidelines") as prepared by the High-Level Expert Group on AI and supported by the Commission.[15] The Guidelines specify that, to be deemed trustworthy, any artificial intelligence should be lawful, ethical and robust: complying with all relevant laws and regulations, respecting ethical principles and not causing unintentional harm.[16] The Guidelines derive seven technical and non-technical requirements from these key principles: human agency & oversight; robustness & safety; privacy & data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.

The Artificial Intelligence Act contains a multi-tiered approach consisting of several regulatory regimes, most of which apply to providers of AI systems. An AI system is defined as a machine-based system that infers its outputs from the inputs it receives; the AI system is the natural or legal person that develops an AI system and places it on the market under its own name or trademark.[17] AI systems are regulated based on their intended purpose, with four purpose-based risk tiers that can be categorized as unacceptable, high, limited, and minimal risk.[18]

Each of the risk tiers described is associated with a corresponding level of regulation. AI systems with unacceptable risk are prohibited. High-risk AI systems are subject to a product safety-based regime in which the provider faces a host of obligations regarding conformity assessment, risk and quality management and record-keeping.[19] Limited-risk AI systems face some transparency obligations, and minimal-risk AI systems are only subject to voluntary codes of conduct.[20]

[14] European Commission, 'Proposal for a Regulation of The European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' 2021/0106(COD), para. 1.1
[15] Ibid, para. 3.2.
[16] High-Level Expert Group on AI, 'Ethics Guidelines for Trustworthy AI' [2019] 5.
[17] Article 3(1) AIA; Article 3(3) AIA.
[18] Michael Veale & Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' [2021] 4 CRi 98
[19] Article 5; Article 16 AIA.
[20] Article 50 and Article 95 AIA.

AI systems with unacceptable risk include systems intended to subliminally manipulate or exploit the vulnerabilities of a natural person, systems related to predictive policing, social credit scoring and some systems related to biometric categorisation.[21] Two main categories of high-risk AI systems exist: AI systems intended as safety components in products already subject to conformity assessment and standalone AI systems with fundamental rights implications.[22] AI practices classified as "high-risk" include systems related to biometrics, the operation of critical infrastructure, education and vocational training, employment, enjoyment of essential services, law & border enforcement, administration of justice and democratic processes.[23]

As the majority of the AIA's provisions are related to high-risk AI systems, it is worth examining the characteristics of AI systems that lead them to be classified as high-risk. As "high-risk AI systems" are defined only in terms of specific intended purposes, several authors have attempted to identify a general set of criteria to determine whether an AI system qualifies as such. Golpayegani, Pandit and Lewis identify five such criteria: domain, purpose, capability, system user and AI subject.[24] Novelli et al. propose to assess AI systems' associated risks using an IPCC-adapted framework that expresses risk as a combination of hazard, exposure, vulnerability, and response.[25]

General-purpose artificial intelligence ("GPAI") models are subject to a separate regulatory framework that is distinct from the frameworks governing AI systems. This framework is analysed in detail in Chapter 4. While obligations imposed on high-risk AI providers are not directly applicable to providers of GPAI models, the AIA's regime for in particular high-risk AI systems can serve to inform us on the AIA's general aims and system.

[21] Article 5 AIA.
[22] AIA proposal, 5.2.3.
[23] Annex III AIA.
[24] Delaram Golpayegani, Harshvardhan Pandit, & Dave Lewis, "To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023)
[25] Claudio Novelli and others, "Taking AI risks seriously: a new assessment model for the AI Act" [2023], *AI & Society* 38 4

3. Large language models and their risks

While large language models offer many new possibilities, they are also associated with a large number of risks. This chapter first sets out the technical foundations of large language models and discusses how some of their technical features contribute to their inherent risks. Next, it describes the often-made distinction between their short-term and long-term risks and lists some key examples of both types of risks.

3.1 Technical foundations of large language models

Large language models are deep learning-based artificial intelligence models that are capable of fluently generating natural language. This section discusses their origins and inner workings, describes how they are used in practice, and further goes into their technical properties that contribute to their associated risks.

3.1.1 Deep learning and large language models

While early artificial intelligence approaches were mostly based on pre-defined rules, most AI advances in the past decade involve some form of self-learning, otherwise known as *machine learning*.[26] The most common machine learning paradigm is that of *supervised learning*, which is the practice of developing an AI system by supplying it with input, desired output, and a method of learning to compute such output from the given input. Such a system then "learns" to predict the desired output on the basis of the given input so that it can afterwards be used to predict such outputs for previously unseen input.[27]

Supervised learning methods often use so-called *neural networks* as a learning method, which are computational models loosely inspired by the human brain. A neural network is a structured set of numerical values that is capable of acquiring knowledge in a training process and modifying its values to store the knowledge acquired.[28] The values of a neural network are referred to as its *parameters*, the way its parameters are structured is referred to as its *architecture*, and its set of trained parameters is referred to as its *weights*. A neural network can be fully represented in terms of its architecture and weights.

[26] Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach*, *Fourth Edition* (Pearson 2021), ch. 19
[27] Ibid chapter 19.1
[28] Simon Haykin, Neural Networks: A Comprehensive Foundation (Pearson 1999) 24

Learning methods involving neural networks with a large number of parameters are referred to as *deep learning* methods. As deep learning enables the representation of abstract concepts within a neural network, deep-learning based AI systems have significantly progressed the state of the art in most if not all subfields of artificial intelligence.[29] Further advancements in deep learning have given rise to the field of generative artificial intelligence, which focuses on developing AI systems capable of generating realistic images, videos, and natural language.[30] The current state of the art in language generation is achieved by *large language models* ("LLMs"), which derive their sophisticated language modelling abilities from being trained on enormous, unfiltered language datasets.[31] In addition to their abilities in tasks like creative writing, translation and text summarization, the most advanced large language models exhibit significant complex reasoning abilities.[32] Such models are often referred to as *frontier models*.

A large language model is created in three steps. The first step is that of *pre-training*, in which a deep learning model is trained to predict the next word in a massive dataset of written text.[33] This training step results in the creation of an AI model with significant language modelling capabilities. The second step is that of *fine-tuning*, in which the model is trained on human-created instruction-response pairs so that it learns to follow user-provided instructions. The third and final step is that of *preference alignment*, in which the model's responses to instructions are evaluated by human annotators and the evaluations obtained are fed back into the model for further training.[34] This step seeks to ensure that the model's responses match users' preferences, are aligned with human values and do not contain inaccurate information: also referred to as responses that are "helpful, honest, and harmless".[35] Finally, the capabilities of the model obtained are evaluated through benchmarks, which are standardized sets of tasks on which large language models' performance can be measured. Capabilities measured by

[29] Yann LeCun, Yoshua Bengio & Geoffrey Hinton, 'Deep Learning' [2015] 512 Nature 436-444
[30] Leonardo Banh & Gero Strobel, 'Generative Artificial Intelligence' [2023] 33 63 Electronic Markets;
See also: Ashish Vaswani and others, 'Attention Is All You Need' [2017] *Preprint*
[31] Tom Brown and others, 'Language Models are Few-Shot Learners', 33rd Conference on Neural Information Processing Systems (NeurIPS 2020).
Humza Naveed and others, 'A Comprehensive Overview of Large Language Models' [2024] *Preprint*
[32] On the capabilities of frontier models as claimed by their developers, see:
'The Claude 3 Model Family: Opus, Sonnet, Haiku' (Anthropic, 2023)
'Gemini: A Family of Highly Capable Multimodal Models' (Gemini Team Google, 2023)
'GPT-4 Technical Report' (OpenAI, 2023)
'Llama 2: Open Foundation and Fine-Tuned Chat Models' (Meta, 2023)
[33] Strictly speaking: tokens, not words
[34] Long Ouyang and others, 'Training language models to follow instructions with human feedback', 36th Conference on Neural Information Processing Systems (NeurIPS 2022).
[35] Amanda Askell and others, 'A General Language Assistant as a Laboratory for Alignment' (Anthropic, 2021)

benchmarks can range from basic math and text comprehension to undergraduate-level or even graduate-level reasoning abilities.[36]

The capabilities of large language models have been observed to scale with the number of model parameters, the size of the data set, and the amount of computational resources used in training the model.[37] Some predict that LLMs' capabilities will continue to improve as long as the factors described are scaled up even further. Others believe that their capabilities will at some point plateau during scaling and that LLMs will need sensory grounding to develop greater intelligence.[38]

3.1.2 Distribution and deployment

Some large language models are open-source, while other such models are proprietary. As a large language model can be fully represented in terms of its architecture and its weights, publicly distributing both enables the model to be freely used, modified, and re-distributed. Nevertheless, developers of publicly released models are able to control legal usage of their models through licensing. While some developers release their models under highly permissive open-source licenses, other developers – such as Meta – release their work under more restrictive conditions. Most providers of frontier models, such as OpenAI, Google and Anthropic, keep their models proprietary by keeping both their models' architecture and weights private. Such providers allow for manual and automated access to their models through web apps and application programming interfaces ("APIs"), which allows users to query their models without having the ability to copy or modify them.[39]

---

[36] Dheeru Dua and others, 'DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs' [2019] *Preprint* <https://arxiv.org/pdf/1903.00161>
Karl Hobbe and others, 'Training Verifiers to Solve Math Word Problems' (OpenAI, 2021)
Dan Hendrycks and others, 'Measuring Massive Multitask Language Understanding', 9th International Conference on Learning Representations (ICLR 2021)
David Rein and others, 'GPQA: A Graduate-Level Google-Proof Q&A Benchmark' [2023] *Preprint* <https://arxiv.org/pdf/2311.12022>

[37] Jared Kaplan, Sam McCandlish and others, 'Scaling Laws for Neural Language Models' (OpenAI, 2020) <https://arxiv.org/abs/2001.08361>

[38] Yann LeCun, 'A Path Towards Autonomous Machine Intelligence' [2022] <https://openreview.net/pdf?id=BZ5a1r-kVsf>

[39] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023)

An ongoing debate exists among safety researchers on whether highly capable models should be open-sourced. Proponents of open-sourcing argue that publicly distributing such models' architecture and weights contributes to their further development, enhances transparency by improving our understanding of their capabilities, and democratizes decision-making on their desired behaviour.[40] Opponents of open-sourcing such models argue that freely distributing them harms efforts to counter misuse.[41]

3.1.3 Properties of large language models contributing to their risks

Despite their widespread adaptation in a relatively short time frame, scientific understanding of large language models' inner workings remains limited. Due to their extremely large number of parameters and emergent properties acquired during training, researchers currently still lack an understanding of "of how they work, when they fail, and what they are even capable of".[42] It can therefore be surprisingly difficult to predict their behaviour in advance, as both their exact capabilities and their value alignment can be hard to establish.

A first problem is that of accurately determining a large language model's capabilities. While the *general* capabilities of large language models tend to smoothly increase as they are trained using bigger data sets, more parameters or more computational resources, their *specific* capabilities abruptly and discontinuously change during scaling.[43] In practice, this means that large language models can learn to perform specific tasks during training which their developer did not predict or expect in advance. This creates uncertainty on LLMs' actual capabilities, which can further contribute to their risk of harm.[44]

---

[40] Sayash Kapoor, Rishi Bommasani and others, 'On the Societal Impact of Open Foundation Models' [2024] *Preprint* <https://crfm.stanford.edu/open-fms/paper.pdf>
Elizabeth Seger and others, 'Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives' (Centre for the Governance of AI and others 2023) <https://arxiv.org/pdf/2311.09227>
Annette Zimmermann, "The Generative AI Deployment Rush: How to Democratize the Politics of Pace" (Keynote, AIES AAII/ACM Conference on AI, Ethics, and Society, Montreal, 2023)
[41] Ibid.
[42] Rishi Bommasani and others, 'On the Opportunities and Risks of Foundation Models' [2021] <https://arxiv.org/pdf/2108.07258.pdf>
Usman Anwar and others, "Foundational Challenges in Assuring Alignment and Safety of Large Language Models" [2024] <https://arxiv.org/pdf/2404.09932.pdf>
[43] Deep Ganguli and others, 'Predictability and Surprise in Large Generative Models' (FAccT Conference on Fairness, Accountability, and Transparency, Seoul, 2022)
[44] Ibid.

A second key problem is that of value alignment, which is the problem of ensuring that the behaviour of a large language model is aligned with human values. Alignment involves both normative and technical considerations: an AI developer has to first establish the normative values that should be embedded in a system and then ensure that the system built actually reflects such values.[45] The problem of verifying an advanced AI model's value alignment – known as scalable oversight – remains an active area of research.[46] Scalable oversight can be particularly complex, including issues such as sycophancy, in which model behaviour aims to immediately please users rather than serve their actual interests, and strategic deception, in which an AI model hides its actual alignment from its users.[47] Finally, an LLM's alignment should be robust against adversarial inputs. Quite some large language models exist that will initially refuse to perform unethical requests, but will still fulfil them if creatively prompted to do so.

3.2 Risks associated with large language models

While large language models have a large number of potential uses and benefits, they also come with significant risks. This section introduces the commonly-made distinction between large language models' short-term risks and long-term risks, discusses why short-term and long-term approaches are not necessarily viewed as compatible with each other and highlights some key risks from both categories.

---

[45] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment' [2020] 30 Minds and Machines 411-437 https://cdn.openai.com/spec/model-spec-2024-05-08.html
See also: Piotr Mirowski, 'A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians' (forthcoming in: FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)
See e.g. OpenAI's Model Spec at https://cdn.openai.com/spec/model-spec-2024-05-08.html
[46] Collin Burns and others, 'Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision' (OpenAI, 2023) <https://arxiv.org/abs/2312.09390>
Samuel Bowman and others, 'Measuring Progress on Scalable Oversight for Large Language Models' (Anthropic, 2022) <https://arxiv.org/abs/2211.03540>
[47] Mrinank Sharma and others, 'Towards Understanding Sycophancy in Language Models' (Anthropic, 2023) <https://arxiv.org/abs/2310.13548>
Evan Hubbinger and others, 'Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training' (Anthropic, 2023) <https://arxiv.org/abs/2401.05566>
Jérémy Scheurer, Mikita Balesni & Marius Hobbhahn, 'Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure' (Apollo Research, 2023) <https://arxiv.org/pdf/2311.07590>

3.2.1 Types of risks

In public debate on the risks of artificial intelligence, a distinction has emerged between proponents of mitigating its short-term risks and proponents of mitigating its long-term risks.[48] Most work on short-term risks is rooted in the field of AI ethics, while most work on long-term risks is rooted in the field of AI safety. Where the field of AI safety is more narrowly concerned with designing AI systems in such a way that they behave as their developer intended, the field of AI ethics seeks to more generally ensure that AI systems respect principles such as fairness, accountability and transparency during their deployment and usage.[49] Both perspectives emphasize different risks related to general-purpose artificial intelligence, to the extent that they are commonly viewed as competing rather than complementary approaches.[50]

Ethics-based work on large language models tends to focus on their short-term risks, which are the immediate risks that exist in today's LLMs. Such work tends to highlight their limitations rather than their capabilities and emphasizes the harms of assuming advanced capabilities while they are not actually there.[51] Key short-term risks include GPAI models' potential for biased, toxic or incorrect outputs, potential violations of privacy and copyright law, and negative externalities such as their resource consumption.[52]

In contrast, safety-based work on large language models tends to focus on potential long-term risks that follow from such models' advanced capabilities.[53] Safety researchers tend to emphasize the risks of both LLMs and AI models in general becoming *too* intelligent, discussing highly capable models' potential for abuse by humans and their capability for autonomously inflicting harm. On the far end, some authors believe that the development of

[48] Henrik Skaug Sætra & John Danaher, 'Resolving the battle of short- vs. long-term AI risks' [2023] AI & Ethics
[49] Dario Amodei and others, 'Concrete Problems in AI Safety' [2016] <https://arxiv.org/pdf/1606.06565>
Tim Rudner & Helen Toner, "Key Concepts in AI Safety: An Overview" (*CSET Issue Brief*, 2021) <https://pdfs.semanticscholar.org/8a1c/cc06ddb9cc128e66c456a2f40d5f594c5c9f.pdf>
[50] ibid
[51] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (FAccT Conference on Fairness, Accountability, and Transparency, Canada, 2021)
[52] Antonella Maia Perini and others, 'Mapping the individual, social and biospheric impacts of Foundation Models' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)
[53] See e.g.: Dan Hendrycks, Mantas Mazeika & Thomas Woodside, 'An Overview of Catastrophic AI Risks' (Center for AI Safety 2023)
Markus Anderljung and others, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (Centre for the Governance of AI e.a., 2023)

general-purpose artificial intelligence comes with existential risks for humanity, arguing that a superintelligent agent is highly unlikely to align its behaviour with human values.[54]

As the short-term and long-term perspectives on AI risk are notably different, their divergence in approaches has led to significant debate. In particular, authors working on short-term risks have often derided long-term concerns as too narrowly technical, as an attempt to distract from the technology's societal risks, and as a means for "Big Tech" to keep dominating the conversation on AI regulation.[55] This work does not seek to reconcile short-term and long-term approaches or argue for one over the other: it rather aims to investigate how both perspectives can inform the way we should operationalize the concept of systemic risk for general-purpose artificial intelligence.

3.2.2 Examples of short-term risks

A first key short-term risk is that of bias: large language models can generate biased and discriminatory output that adopts and amplifies biases found in their training data. [56] Although problem of bias arises in any data-based language processing approach, large language models' novel capabilities can exacerbate the scale of the problem.[57] Hacker e.a. distinguish two main categories of discriminatory output in the context of content generation: demeaning and abusive

---

[54] Nick Bostrom, 'The Superintelligent Will' [2012] 22 Minds and Machines 71-85
Nate Soares, 'The Value Learning Problem' (IJCAI International Joint Conference on Artificial Intelligence, New York, 2016)
Andrew Critch & David Krueger, 'AI Research Considerations for Human Existential Safety' [2020] *Preprint*
[55] See e.g.: Seth Lazar & Alondra Nelson, 'AI safety on whose terms?' [2023] 381 6654 Science 138
The Editorial Board, 'Stop talking about tomorrow's AI doomsday when AI poses risks today' [2023] 618 Nature 885–886
[56] Vinitha Gadiraju and others, "'I wouldn't say offensive but...': Disability-Centered Perspectives on Large Language Models" (FAccT Conference on Fairness, Accountability, and Transparency, Seoul, 2022)
Moin Nadeem, Anna Bethke & Siva Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models" (Annual Meeting of the Association for Computational Linguistics, Online, 2022)
Abubakar Abid, Maheen Farooqi & James Zou, "Persistent anti- muslim bias in large language models" (AAAI/ACM Conference on AI, Ethics, and Society, New York, 2021)
Jesse Vig and others, "Investigating gender bias in language models using causal mediation analysis" (NeurIPS Conference on Neural Information Processing System, Vancouver, 2020)
Messi Lee, 'Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)
[57] Aylin Caliskan, Joanna Bryson & Arvind Narayanan, 'Semantics derived automatically from language corpora contain human-like biases' [2017] 356 Science 183
Nikhil Garg and others, 'Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes' [2018] 115 PNAS 3635
Justus Mattern and others, 'Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing' [2022] <https://arxiv.org/abs/2212.10678>
Emily Sheng and others, 'The Woman Worked as a Babysitter: On Biases in Language Generation' [2019] <https://arxiv.org/abs/1909.01326>

content and inadequate representation.[58] Furthermore, an LLM's inherent biases could lead allocative harms or quality-of-service harms when it is integrated into a decision-making procedure or user-facing application.[59] Risks of such harms are particularly high if large language models are used for fully automated decision-making, but still remain when a "human in the loop" exists due to automation bias and selective adherence to algorithmic advice.[60]

A second short-term risk is that of hallucination, in which large language models tend to generative plausible-sounding statements that are factually incorrect. As an LLM essentially performs next-word prediction based on its ability to model natural language, but is "not grounded in other domains of experience, such as video or real-world physical interaction", it lacks context about the world and its outputs are not directly informed by any source of truth.[61] For this reason, some argue that 'hallucination' is a misleading term, as it implies some relation to the truth while no such relation does in fact exist.[62] While attempts exist to mitigate or eliminate large language models' tendency to hallucinate, LLMs' tendency to make up falsehoods remains unresolved so far.

A third highlighted short-term risk is that of large language models' sustainability impact. As LLMs' development and deployment requires significant computational resources, their usage can lead to increased energy usage and carbon footprints.[63] Furthermore, the large-scale computational infrastructure they depend on can cause significant ecological damage both

---

[58] Philipp Hacker, Frederik Zuiderveen Borgesius, Brent Mittelstadt & Sandra Wachter, 'Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It' [2024] *Preprint* <https://media.licdn.com/dms/document/media/D4E1FAQH4ZFDb_cBLyA/feedshare-document-pdf-analyzed/0/1719418044729?e=1720051200&v=beta&t=v7MqDgvnlNYnQBLxJyHTgSWDVF0vZmNNwAoPDB1LxA0>

[59] See: Renee Shelby and others, 'Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction' (AIES Conference on Artificial Intelligence, Ethics, and Society, Montreal, 2023)
Valentin Hofmann and others, 'Dialect prejudice predicts AI decisions about people's character, employability, and criminality' [2024] <https://arxiv.org/abs/2403.00742>

[60] Linda Skitka, Kathleen Mosier & Mark Burdick, 'Does automation bias decision-making?' [1999] 51(5) HCS 991-1006
Irina Carnat, 'Human, all too human: accounting for automation bias in generative Large Language Models' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)
Saar Alon-Barkat & Madalina Busuioc, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' [2023] 33(1) JPART 153-169

[61] Tom Brown and others, 'Language Models are Few-Shot Learners', 33rd Conference on Neural Information Processing Systems (NeurIPS 2020).

[62] Michael Townsen Hicks, James Humphries & Joe Slater, 'ChatGPT is bullshit' [2024] 26 38 EIT

[63] Mattias Rillig and others, 'Risks and Benefits of Large Language Models for the Environment' [2023] 57(9) EST 3464-3466 <https://pubs.acs.org/doi/10.1021/acs.est.3c01106>
Steven Monserrate, 'The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage' (MIT Case Studies in Social and Ethical Responsibilities of Computing 2022) <https://s3.amazonaws.com/assets.pubpub.org/2l4m6ikahjxvjn3wgr6aj498hji0t0cg.pdf>

locally and globally.[64] Notably, the risk of sustainability impact differs from other short-term risks in that it is not directly related to the functionality of large language models. Instead, I would argue that it is best viewed as an externality that is a consequence of their large-scale deployment.

A final short-term risk of large language models is that of threats to democratic processes, with malicious actors potentially abusing their capabilities to generate propaganda or otherwise influence public opinion.[65] As large language models can generate highly persuasive content on a previously unseen scale, they are likely to lower the cost of propaganda and disinformation campaigns and could make such campaigns much harder to discover.[66]

3.2.3 Examples of long-term risks

A first long-term risk highlighted is that of knowledge barrier reduction, in which advanced LLMs' capabilities could lower the barrier for malicious actors to conduct cyber-attacks or develop chemical, biological, radiological or nuclear weapons.[67] The risk on knowledge barrier reduction are grounded in concerns of large language models becoming both highly capable, yet not aligned enough with human values to refuse to perform such tasks. Furthermore, worries exist that open-sourcing highly capable models poses greater risks of knowledge barrier reduction, as doing so could give malicious actors much more control over their behaviour.[68]

A second, even more hypothetical risk is that of self-proliferation. Self-proliferation is the risk of an intelligent artificial agent creating and training a new, even more intelligent agent, which in turn creates a new even more advanced agent. Bostrom hypothesizes that this cycle of

---

[64] Antonella Maia Perini and others, 'Mapping the individual, social and biospheric impacts of Foundation Models' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)

[65] Josh Goldstein and others, 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations' [2023] *Preprint* <https://arxiv.org/pdf/2301.04246>

[66] ibid 72

Kobi Hackenburg and others, 'Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues' [2023] *Preprint* <https://www.benmtappin.com/papers/Role-play_Human_Experts_PREPRINT.pdf>

Hu Bai and others, 'Artificial Intelligence Can Persuade Humans on Political Issues' [2023] *Preprint* <https://osf.io/preprints/osf/stakv>

[67] Miles Brundage and others, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (Future of Humanity Institute 2018) <https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf >

Toby Shevlane and others, 'Model evaluation for extreme risks' (Google DeepMind 2023)

[68] Anjali Gopal and others, 'Will releasing the weights of future large language models grant widespread access to pandemic agents?' [2023] *Preprint* <https://arxiv.org/pdf/2310.18233>

"recursive self-improvement" could lead to the creation of an artificial superintelligence.[69] Concerns on self-proliferation are mostly related to the problem of human control over a superintelligent agent, with some authors arguing that such an agent's rational, self-interested behaviour is highly unlikely to conform to human values and therefore creates existential risks for humanity.[70] While an LLM-based application such as ChatGPT would not have the technical means to self-proliferate as it does not have the means to manipulate digital or physical systems, a large language model with access to computational resources and storage space could – theoretically – self-proliferate. Such an LLM would however require much more advanced capabilities than today's state-of-the-art models possess.[71]

[69] Irving John Good, 'Speculations concerning the first ultraintelligent machine' [1965] 6 Advances in Computers, 31, 33
Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014) 29
[70] Nick Bostrom, 'The Superintelligent Will' [2012] 22 Minds and Machines 71-85
Nate Soares, 'The Value Learning Problem' (IJCAI International Joint Conference on Artificial Intelligence, New York, 2016)
Andrew Critch & David Krueger, 'AI Research Considerations for Human Existential Safety' [2020] *Preprint*
[71] Mary Phuong and others, 'Evaluating Frontier Models for Dangerous Capabilities' (Google DeepMind 2024)

4. General-purpose artificial intelligence and systemic risk

Chapter V of the Artificial Intelligence Act introduces a novel, standalone regulatory regime for general-purpose AI models ("GPAI models"), under which providers of such models with "high-impact capabilities" are obligated to assess associated systemic risks. This chapter outlines the background of the general-purpose AI regulatory regime, sets out its legal framework regarding models with systemic risk and discusses open questions on the interpretation of some of the frameworks' key provisions. Finally, it discusses existing work on the systemic risk provisions found in the Digital Services Act and argues that some of this work can be meaningfully applied to the AIA's general-purpose artificial intelligence regime.

4.1 Background

As the AIA's original proposal predates the widespread emergence of large language models, its regulatory regime for general-purpose artificial intelligence was only introduced in the final stretches of the legislative process. During drafting, the regime for GPAI models – referred to in earlier drafts as "foundation models" – shifted from first qualifying them as high-risk AI systems by default, then qualifying them as such based on their downstream applications, and finally introducing the distinct regime as we know it from the final version.[72] This departure from the AIA's high-risk regime was made necessary by the fact that a GPAI model does not have a specific intended purpose, which in the AIA is the key determinant of an AI system's qualification as high-risk.[73]

The AIA's regulatory regime for general-purpose artificial intelligence is often understood as a political compromise favouring industry interests over ethical safeguards. During the final round of interinstitutional negotiations, questions on rules for general-purpose artificial intelligence were the last obstacle to a political agreement after some Member States came out as opposing all binding regulation on such models.[74] Additionally, lobbying work from the German and French AI industries is viewed as having played a significant part in softening the

[72] Claudio Novelli and others, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' [2024] *Preprint* <https://arxiv.org/pdf/2401.07348> accessed 10 May 20204
[73] See: Article 6 resp. Annex III AIA;
Natali Helberger & Nicholas Diakopoulos, 'ChatGPT and the AI Act' [2023] 12 1 Internet Policy Review <https://policyreview.info/essay/chatgpt-and-ai-act>;
Philipp Hacker, Andreas Engel & Marco Mauer, "Regulating ChatGPT and other Large Generative AI Models" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023) 1114.
[74] Luca Bertuzzi, 'EU's AI Act negotiations hit the brakes over foundation models' (*Euractiv*, 10 November 2023) <https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/> accessed May 9 2024

obligations imposed on GPAI model providers.[75] For this reason, some academics and NGOs have cast the final agreed text as a failure for effective fundamental rights protection.[76]

4.2 Legal framework

The AIA's systemic risk assessment and mitigation obligations apply to providers of general-purpose AI models with systemic risks, where such models qualify as having systemic risk if they have high-impact capabilities. This section outlines the key concepts of "general-purpose AI model", "high-impact capabilities" and "systemic risk" and explains how they are interconnected in the context of the AIA's regulatory framework.

4.2.1 General-purpose AI models

A general-purpose AI ("GPAI") model is defined as an artificial intelligence model "that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market".[77] Such generality and capable performance is presumed for models with at least one billion parameters.[78] A GPAI model is distinct from an AI system and is not an AI system on its own.[79] Just like with AI systems, the provider of a GPAI model is the natural or legal person that places the model on the market under its own name.[80]

The definition of general-purpose artificial intelligence has been subject to change during the AIA's drafting. An early draft defined a GPAI system as an "AI system (…) able to perform generally applicable functions such as image/speech recognition, audio/video generation, pattern detection, question answering, translation etc".[81] As this definition merely listed some specific AI tasks, some authors argued that it failed to effectively demarcate general-purpose

---

[75] See e.g.: 'Trojan horses: how European startups teamed up with Big Tech to gut the AI Act' (*Corporate Europe Observatory* 11 March 2024) <https://corporateeurope.org/en/2024/03/trojan-horses-how-european-startups-teamed-big-tech-gut-ai-act> accessed May 9 2024
Jacha Bareis, 'BigTech's Efforts to Derail the AI Act' (*Verfassungsblog,* 5 December 2023) <https://verfassungsblog.de/bigtechs-efforts-to-derail-the-ai-act/> accessed May 9 2024
[76] Ljubiša Metikoš, 'De AI Wet: Zwak, Zwakker, Zwakst' [2024] 3 Mediaforum 73
'The EU AI Act: a failure for human rights, a victory for industry and law enforcement' (*Access Now*, 13 March 2024) <https://www.accessnow.org/press-release/ai-act-failure-for-human-rights-victory-for-industry-and-law-enforcement/> accessed May 9 2024
[77] Article 3(63) AIA.
[78] Recital 98 AIA.
[79] Recital 97 AIA.
[80] Article 3(3) AIA.
[81] https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

AI from fixed-purpose AI systems.[82] To resolve this issue, Gutierrez e.a. propose to define a GPAI system as "an AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained".[83] Uuk e.a. propose four different ways to assess an AI models' generality: the *quantity* of tasks a model can perform, their *performance* in distinct tasks, their *adaptability* to perform new distinct tasks and their potential for developing *emergent abilities*.[84]

The final text of the AIA opts for a definition based on quantity and performance that is very close to the one proposed by Gutierrez e.a., but adds a distinction between "models" and "systems". This distinction reflects the difference between large language models and the applications in which they are integrated. In my opinion, the AIA's final definition adequately addresses the concerns by Gutierrez and Uuk by clarifying that an AI model only qualifies as "general-purpose" if it is capable of performing a wide range of tasks. Moreover, as present-day large language models are undeniably "general-purpose" in nature, the specifics of defining general-purpose artificial intelligence might be of limited relevance to their regulation.

4.2.2 High-impact capabilities

High-impact capabilities are defined as "capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models".[85] Whether a model has high-impact capabilities is evaluated "on the basis of appropriate technical tools and methodologies, including indicators and benchmarks".[86] Such evaluation should take into account the number of model parameters, data set size and quality, the amount of computation used for training the model, input and output modalities, benchmark performance, number of business users and number of registered end-users.[87] High-impact capabilities are presumed for models trained with at least 10^25 floating point operations ("FLOPs"), which is a metric for computational

[82] Carlos Gutierrez and others, 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems' [2023] 2 36 Digital Society
Philipp Hacker, Andreas Engel & Marco Mauer, "Regulating ChatGPT and other Large Generative AI Models" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023) 1114
[83] Ibid 5.
[84] Risto Uuk, Carlos Ignacio Gutierrez & Alex Tamkin, 'Operationalising the Definition of General Purpose AI Systems: Assessing Four Approaches' [2023] *Preprint* <https://arxiv.org/pdf/2306.02889>
[85] Article 3(64) AIA.
[86] Article 51(1)(a) AIA.
[87] Annex XIII AIA.

resources used.[88] The Commission may amend the thresholds for "high-impact capabilities" by delegated act.[89]

General-purpose AI models essentially qualify as having high-impact capabilities if they are what AI researchers would refer to as *frontier models*: models whose capabilities match or improve on the state of the art in artificial intelligence research. The listed criteria for high-impact capabilities seem to be derived from scientific work on scaling laws, general industry practices and a wish to take into account user adaptation. The criteria regarding model parameters, data set size and computation used are directly taken from OpenAI's work on scaling laws, which cites the same criteria as key explanatory variables for large language models' performance.[90] The wording on input and output modalities emphasizes the – potentially greater – risks of models other than large language models that are capable of processing visual information or generating biological sequences.[91] The inclusion of benchmark performance matches standard industry practice in evaluating large language models' capabilities, and the number of registered business users and end users serves as a proxy for model adaptation.

The FLOPs-based high impact capability presumption threshold has drawn criticism, with Novelli arguing that it does not capture GPAI model's multidimensional nature, that the initial threshold of 10^25 FLOPs is too high and that advances in building more efficient models may soon make it redundant.[92] I think the first objection is most relevant, as the latter two objections can be remedied through a delegated act by the Commission. As the FLOPs threshold is seemingly derived from OpenAI's scaling laws research, one can wonder why the amount of computation used is highlighted at the expense of model parameters and data set size. While I would further agree that the threshold is somewhat arbitrary, it is important to keep in mind that the threshold is a rebuttable presumption rather than an iron rule. Therefore, any particularities not captured by the 10^25 FLOPs threshold can still be addressed on a case-by-case basis.

---

[88] Article 51(2) AIA.
[89] Article 51(3) AIA.
[90] Jared Kaplan, Sam McCandlish and others, 'Scaling Laws for Neural Language Models' (OpenAI, 2020) <https://arxiv.org/abs/2001.08361>
[91] On biological design tools, see: Jonas Sandbrink, 'Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools' [2023] *Preprint* <https://arxiv.org/pdf/2306.13952> accessed 10 May 2024
[92] Claudio Novelli and others, "Generative AI in EU Law: Liability, Privacy, Intellectual Property and Cybersecurity" [2024]

Authors have generally assumed that the Commission's power to adapt the high-impact capability threshold merely is about the FLOPs threshold. Interestingly however, the text gives the Commission the power to "amend the thresholds listed in paragraphs 1 and 2 of this Article", which would not only cover the FLOPs definition but also the definition of high-impact capabilities in general. This raises the question whether the Commission is merely empowered to amend the FLOPs threshold, or whether the term "thresholds" should be read more broadly as to allow for a broad redefinition of the "high-impact capabilities" concept.

4.2.3 Systemic risk

A systemic risk is defined as follows:

> "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" [93]

This definition contains three key requirements. The first – and obvious – requirement is that there should be a "risk", which is the combination of the probability of an occurrence of harm and the severity of that harm.[94] The second requirement is that the risk should be "specific to the high-impact capabilities of general-purpose AI models", which implies that the risk should in some way be linked to frontier models' advanced capabilities. The third requirement is that the risk should have a "significant impact on the Union market", either due to general-purpose AI models' reach, or due to their negative effects on a wide range of domains that can be propagated at scale. This requirement specifies that the term "systemic risk" primarily describes societal-scale risks, with individual risks most likely falling outside of the concept's scope.

Recital 110 provides an outline of potential systemic risks posed by general-purpose AI models. It describes two main categories of systemic risk, names eight risk factors influencing systemic risk and describes eight examples of potential systemic risks. The first described category of systemic risk describes the risk of "major accidents, disruptions of critical sectors and serious

[93] Article 3(65) AIA.
[94] Article 3(2) AIA.

consequences to public health and safety". The second category describes negative effects on "democratic processes, public and economic security and the dissemination of illegal, false or discriminatory content".

Risk factors influencing systemic risk include conditions of misuse; model reliability, model fairness and model security; the level of autonomy of the model; its access to tools; novel or combined modalities; release and distribution strategies; the potential to remove guardrails; and other factors.[95] Potential systemic risks include alignment issues related to control and malicious use; barrier-of-entry risks for chemical, biological, radiological and nuclear weapons development; models' potential for offensive cybersecurity operations; risks related to models interacting with digital and physical systems ("tool use"); models self-replicating or training other models; bias & discrimination at the individual, collective or societal level; disinformation & privacy harms; and "the risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community."

Some listed examples of systemic risks reflect short-term risks, while others reflect long-term risks.[96] Short-term risks described include risks regarding bias, discrimination, disinformation and privacy, all of which well-documented in literature on ethical artificial intelligence. Specifically, the mention of bias "at the individual, collective or societal level" appears to be derived from Smuha's work.[97] Long-term risks described include barrier-of-entry reduction, concerns relating to alignment, control and malicious use, self-replication, and the risk of a "chain reaction". The risk of barrier-of-entry reduction for e.g. chemical or biological weapons development has not received a lot of attention in academic literature, but has been the subject of a significant number of safety think-tank reports. The phrasing "alignment, control and malicious use" seems to refer to alignment issues for artificial superintelligence, apparently citing the possibility that we might lose control over advanced general-purpose AI models. The

[95] On novel & combined modalities, see:
Sophie Rose & Cassidy Nelson, 'Understanding AI-Facilitated Biological Weapon Development' (Centre for Long-Term Resilience 2023) <https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio>

[96] The examples listed seem to be directly taken from the G7-established Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. See: 'Commission welcomes G7 leaders' agreement on Guiding Principles and a Code of Conduct on Artificial Intelligence' <https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-g7-leaders-agreement-guiding-principles-and-code-conduct-artificial>

[97] Nathalie Smuha, "Beyond the individual: governing AI's societal harm" [2021] 10 Internet Policy Review 3

risk of "self-replication" evokes similar concerns: the phrase seems to reflect existential risk concerns such as Bostrom's "intelligence explosion" in which an highly intelligent AI system recursively creates progressively more intelligent versions of itself.[98] Finally, the described risk of a chain reaction is not quite specified and leaves its meaning to the imagination of the reader.

4.3 Discussion

4.3.1 High-impact capabilities as a static or dynamic threshold

As discussed in Section 4.2.2, the AIA essentially defines high-impact capabilities as capabilities that match or exceed the state of the art ("the frontier") in AI development. With the release of newer, more capable models, the state of the art is however certain to shift over time. This raises the question whether the high-impact capabilities threshold should be adapted upon the release of new frontier models. In other words: will GPT-4 still qualify as having high-impact capabilities as soon as GPT-5 is released?

The definition of high-impact capabilities as "capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models" seems to suggest a dynamic threshold that should be revised upward with every release of a newer, more capable GPAI model. This interpretation would however not make sense in the light of sensible risk management strategy, as the release of a newer frontier model clearly does not negate the systemic risks of frontier models that already exist. Therefore, I would advocate a static threshold that is based on the state of the art around the time at which the general-purpose AI regime becomes applicable. Such a threshold would ensure that highly capable models are consistently and predictably regulated regardless of unexpected advancements.

4.3.2 Short-term and long-term risks as "specific to" advanced models' capabilities

Article 3(65) AIA defines systemic risks associated with general-purpose AI models as "specific to" to their high-impact capabilities. Although it seems clear that the phrase "specific to" requires some link between a risk and GPAI models' high-impact capabilities, it is unclear how exactly this link should be interpreted and what the implications of this requirement would be for the scope of the AIA's systemic risk concept.

[98] Nick Bostrom, *Superintelligence* (Oxford University Press 2014) 29

To illustrate the relevance of this issue, it is worth remembering that not all risks related to large language models are generally considered to be specific to their advanced capabilities. This especially applies to most short-term risks, which I would instead view as specific to LLMs' limitations or to their externalities. The risk of an LLM providing biased or discriminatory output, for example, exists regardless of the model's capabilities and does not increase or decrease with its performance on a given task. The same applies to the risk of a large language model producing false or misleading output: this is not a risk linked to an LLM's "high-impact capabilities" but a risk linked to presuming such capabilities when they might not be there. I would further argue that LLMs' risks of environmental harm are not at all a consequence of their capabilities, with such risks mostly depending on the hardware on which they are hosted.

On the contrary, most if not all long-term risks related to large language models can be understood as a direct consequence of their advanced capabilities. The described risks of knowledge barrier reduction and self-proliferation are all inherently linked to such models' – potentially – advanced intelligence, with the former as an example of abuse of such capabilities by malicious actors and the latter as an example of an advanced intelligence acting autonomously.

A restrictive interpretation of Article 3(65)'s specificity requirement could therefore reduce the scope of the AIA's systemic risk concept to predominantly long-term AI risks. One could read the "specific to" phrase to imply that an LLM-associated risk can *only* qualify as a systemic risk if it exists in highly capable models, but does not exist in models without high-impact capabilities.[99] This reading would remove most short-term risks, which appear both in highly capable and less capable models, from the systemic risk concept's scope.

Several arguments could be made against such a restrictive interpretation. First of all, relevant recitals specifically mention short-term risks related to discrimination, disinformation, and privacy harms as potential systemic risks.[100] Additionally, a teleological argument can be made that excluding most short-term risks goes against the purpose of the AIA, which is to ensure a high level of fundamental rights protection. This argument could be further supported by an

[99] See also: Cornelia Kutterer, 'Regulating Foundation Models in the AI Act: From "High" to "Systemic" Risk' (AI-Regulation Papers 2024) 7 <https://ai-regulation.com/wp-content/uploads/2024/01/C-Kutterer-Regulating-Foundation-Models-in-the-AI.pdf>
[100] Recital 110 AIA, see also Section 4.3.2.

analogy with the General Data Protection Regulation ("GDPR"), of which the Court of Justice has consistently interpreted its provisions in the light of the GDPR's fundamental aim of ensuring a high level of personal data protection for natural persons.[101]

While the AIA's systemic risk concept should therefore arguably include risks not directly related to GPAI models' high-impact capabilities, doing so still raises the question of how to conceptualize such risks as *specific to* such capabilities. A reasonable way of doing so could be to view short-term risks as "systemic" insofar they are *amplified* by models' advanced capabilities. Consider for example two large language models A and B, where both models exhibit a similar bias against minority groups. Assume that A is a highly capable model, while B is not. Due to its advanced capabilities, A will be much more frequently used and relied upon compared to B. Therefore, A has much greater potential for collective or societal harm, as its high-impact capabilities amplify its inherent bias to such an extent that it becomes a systemic risk.

A real-world example of this amplification argument would be a large language model performing the task of applicant screening for a job application procedure. LLMs could very well be prone to undervaluing applications by members of minority groups. If a non-capable model is used for this task, providing nonsensical recommendations regardless of its bias, the most likely outcome would be that the end user discards its outputs, undoing any harms introduced by the model's bias. If a highly capable model with similar biases is however used, the odds of real-world allocative harm being propagated become much greater, as the end user of the model is much more likely to accept its recommendations.

Although this amplification argument could be useful in broadening the scope of the AIA's systemic risk concept, I do not believe that it can be used to qualify *any* risk as a systemic risk, as some link between the risk and a GPAI model's high-impact capabilities still needs to exist.

---

[101] See e.g.: Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317 ("Google Spain")
Case C 362/14 Maximillian Schrems v Data Protection Commissioner [2015] ECLI:EU:C:2015:650 ("Schrems I")
Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779 ("Breyer")
Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH [2016] ECLI:EU:C:2018:388 ("Wirtschaftsakademie")
Case C 311/18 Data Protection Commissioner v Facebook Ireland Ltd & Maximillian Schrems [2020] ECLI:EU:C:2020:559 ("Schrems II")
Case C-634/21 OQ v Land Hessen [2023] ECLI:EU:C:2023:957 ("Schufa")

For example, I would view risks related to LLMs' sustainability impact as too far removed from a model's functionality and therefore not capable of qualifying as a "systemic risk". This underscores the limitations of conceptualizing systemic risks as specific to advanced capabilities.

In conclusion, the interpretation of Article 3(65) AIA's specificity requirement could have significant implications for the scope of its systemic risk concept. While a restrictive interpretation could limit the concept to risks directly linked to models' advanced capabilities, sensible textual and teleological arguments exist for a broader interpretation that would include other key risks. Such risks could be argued to qualify as systemic risks insofar they are amplified by an advanced model's high-impact capabilities. Nevertheless, I would argue that the AIA's specificity requirement still does exclude some risks from qualifying as "systemic", as some link between the risk and a model's advanced capabilities still needs to exist, illustrating the limits of the AIA's systemic risk definition.

4.4 Insights from the Digital Services Act as a means for systemic risk assessment

As much questions remain on the AIA's systemic risk provisions, work on the similar provisions in the Digital Services Act ("DSA") might provide some useful guidance. This section briefly describes the DSA's legal framework, discusses proposed risk assessment frameworks for the DSA and briefly discusses their implications for our interpretation of the Artificial Intelligence Act.

4.4.1 Legal framework

Under the Digital Services Act, providers of very large online platforms ("VLOPs") and search engines ("VLOSEs") are obliged to assess and mitigate systemic risks that arise from their services.[102] These obligations are intended to address the risk of social harm caused by design choices incorporated in such platforms and search engines, which is amplified by the "systemic role in amplifying and shaping information flows online" inherent to their scale.[103] Systemic risk assessment should at least consider risks related to the dissemination of illegal content, the exercise of fundamental rights, civil discourse and electoral processes, and "gender-based violence, protection of public health and minors and serious negative consequences to the

[102] Articles 33-35 DSA.
[103] European Commission, 'Impact Assessment on the Digital Services Act' (SWD(2020) 349 final, 15 December 2020, para. 33.

person's physical and mental well-being".[104] As of 2024, entities subject to the DSA's systemic risk obligations include a wide array of social media apps, online shopping concerns, search engines, video services, and miscellaneous online platforms. Well-known examples include Facebook, LinkedIn, TikTok, X, Google Search, Bing, Wikipedia and YouTube.[105]

4.4.2 Proposed systemic risk assessment frameworks

Loi offers a mathematical formalization of the concept of systemic risk.[106] She proposes a framework in which risks are viewed as consisting of a risk event A, the probability of a risk event p, the negative consequence $\boldsymbol{C}$, the – negative – value of the consequence $v(C)$ and the conditional probability of a negative consequence p'. In this framework, the systemic risk produced by a given service S would be quantifiable as the product $R_s = p_s * p'_s * v(C)$. Relevant probabilities would be empirically measured by counting the frequency of certain types of events.

I believe that this framework does not meaningfully contribute to our understanding of systemic risk. While describing risk as $R_s = p_s * p'_s * v(C)$ instead of "the combination of the probability of an occurrence of harm and the severity of that harm" might be more readable for the mathematically inclined, doing so does not yield any new insights into the nature of online platforms' systemic risks. Furthermore, the framework fails to address the *systemic* part of "systemic risk": it treats a systemic risk like any other risk without attempting to understand what properties would make a risk "systemic".

Broughton Micova & Calef offer a more useful framework, applying some concepts from the study of financial systemic risk to systemic risk assessment for very large online platforms.[107] In this framework, systemic risk is defined as the risk of experiencing *systemic* events, where a systemic event consists of *shocks* and *propagation mechanisms*.[108] A shock is a negative event, while a propagation mechanism is the way through which the consequences of this event

---

[104] Article 34(1) DSA.
[105] European Commission, 'Supervision of the designated very large online platforms and search engines under DSA' <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses> accessed May 10 2024
[106] Michele Loi, "Making sense of the Digital Services Act: How to define platforms' systemic risks to democracy" [2023] *Algorithm Watch* <https://algorithmwatch.org/de/wp-content/uploads/2023/08/AlgorithmWatch_Risk_Assessment-DSA.pdf>, accessed 22 February 2024
[107] Sally Broughton Micova & Andrea Calef, 'Elements for Effective Systemic Risk Assessment Under the DSA' (CERRE 2023)
[108] Olivier de Bandt & Philipp Hartmann, 'Systemic Risk: A Survey' (ECB Working Paper Series no. 35, 2000) 10-11

negatively impact other actors. Shocks can be either *idiosyncratic* or *systematic*, where idiosyncratic shocks initially affect a single actor while systematic shocks affect a large number of actors at once. Furthermore, shocks can be *endogenous* or *exogenous*, where an endogenous shock originates within the system and an exogenous shock originates outside of the system.

The value of a shock-propagation systemic risk assessment framework is further demonstrated by a recent European Commission study, which develops a risk assessment framework for online disinformation campaigns that is notably similar to the framework Micova & Calef propose.[109] The study describes a systemic risk as a risk reaching a certain degree of severity, defining this severity as "a function of the relationship between the qualitative assessment of the risk posed by the content in context and a quantitative measure of the reach and/or intensity of exposure of audiences to that content".[110] Based on this definition, the study proposes a four-step assessment framework: qualitatively evaluating harmful content, using quantitative metrics to evaluate its potential for spread and amplification, qualitatively determining whether a platform has established mitigation measures and quantitatively measuring the effectiveness of such measures in reducing amplification and spread.

Apart from the risk assessment framework described, the Commission has sent requests for information under the DSA to key very large online platforms regarding their mitigation measures related to "so-called 'hallucinations' where AI provides false information, the viral dissemination of deepfakes, as well as the automated manipulation of services that can mislead voters" and "impact of generative AI on electoral processes, dissemination of illegal content, protection of fundamental rights, gender-based violence, protection of minors, mental well-being, protection of personal data, consumer protection and intellectual property".[111] To my knowledge, this is the first and so far only formal Commission action which describes large language models' risks related to hallucination and election manipulation as potential systemic risks.

---

[109] European Commission, Directorate-General for Communications Networks, Content and Technology, 'Digital Services Act – Application of the risk management framework to Russian disinformation campaigns' (Publications Office of the European Union 2023) <https://data.europa.eu/doi/10.2759/764631>

[110] p. 15

[111] Press release, 'Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act' (European Commission, 14 March 2024) <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very> accessed May 11 2024

4.4.3 Lessons for our interpretation of the Artificial Intelligence Act

Existing work on the DSA's systemic risk assessment provisions provides some useful lessons for our interpretation of the AIA's similar concept. In particular, distinguishing between sources and spread of risk could be well suited for general-purpose AI systemic risk assessment and the enforcement priorities of the Commission suggest that it will not settle for an overly narrow interpretation of the AIA's systemic risk concept.

First of all, I believe that a clear distinction between sources of risk and potential for spread of a risk, as exemplified by Broughton Micova & Calef's shock-propagation framework and the Commission's risk assessment framework, could be a valuable contribution to systemic risk assessment for general-purpose artificial intelligence. As risks related to large language models are determined both by properties inherent to the model and the way such models are used in practice, distinguishing between 'shocks' and 'propagation' could potentially be useful in assessing their systemic risks. Potentially harmful behaviour by a large language model could reasonably be viewed as a 'shock', while the way in which such behaviour causes societal harm could be characterized as the propagation of a shock.

Furthermore, the limited formal Commission actions so far indicate that it is aware of a broad array of risks related to generative AI and large language models, and that it is especially concerned about large language models' potential for disrupting elections. Such enforcement priorities could lead the Commission to take a broad view on the scope of "systemic risk" in the AIA, and would make it unlikely to merely investigate hypothetical long-term risks. However, predictions regarding Commission priorities should always be taken with a grain of salt, as different teams may be responsible for different enforcement actions and their priorities can shift over time with the political climate.

5. Systemic risk assessment for general-purpose artificial intelligence

Insights from studying large language models' risks and the legal frameworks of the AIA and the DSA can be used to develop assessment criteria for general-purpose AI models' systemic risks. This chapter builds on insights gained from the preceding chapters to propose three systemic risk assessment criteria for general-purpose AI models and argues that they are consistent with key risk factors and legal requirements identified. Finally, it applies the proposed criteria to the risks of large language models identified in Chapter 3 and discusses whether they could qualify as "systemic risks" in the sense of Article 55 AIA.

5.1 A capability-alignment-propagation systemic risk assessment framework

This section proposes a systemic risk assessment framework for general-purpose artificial intelligences that is based on three criteria: capability, alignment, and propagation. First, it outlines the proposed assessment criteria step by step and lists examples of how they can be applied. Second, it argues that the criteria proposed adequately represent most risk factors identified in academic work on large language models' risks and the Artificial Intelligence Act.

5.1.1 Criteria for systemic risk assessment

I propose to assess general-purpose AI models' potential for systemic risk on the basis of three criteria: the capabilities of the model, the alignment of the model and the potential for a risk to propagate through society. The capability and alignment criteria are inherent properties of a model determining its behaviour, while the potential for a risk to propagate is determined by the way in which model behaviour impacts society. Taken together, the three criteria can form the basis for a sociotechnical evaluation of general-purpose AI models' risks suitable for determining whether they qualify as "systemic".

The model capability and alignment criteria are essentially about what a model *can* do and what it *will* do: what novel, complex tasks can it perform? Are there tasks which it can clearly perform given its reasoning capabilities, but will nevertheless refuse as unethical? How can the combination of a models' capability and value alignment lead to potentially harmful behaviour? Also note that any assessment of a model's alignment should measure both its explicit and its implicit alignment: it should refuse to perform harmful tasks, but it also should not exhibit more subtle biases while performing a task. Mandating general-purpose AI model developers to

assess their models' risks through measuring their capabilities and alignment would be nothing new, as all such developers already perform extensive safety research into exactly those two properties.

In contrast to the capability and alignment criteria proposed, the propagation criterion would involve assessing the societal consequences of the distribution, deployment and usage of a general-purpose AI model. Model providers should assess how the way their products are used affects society at large: how do their models influence their users' behaviour? Are they used to make important decisions? Additionally, the scale at which harmful model behaviour could propagate through society is highly relevant, as a merely small-scale risk would be unlikely to qualify as a systemic risk. Mandating such a risk evaluation would further be a significant step up from existing practice, in which general-purpose AI developers mostly tend to perform safety research into their models' technical properties rather than into their societal implications.

5.1.2 Connecting the proposed criteria to risk factors identified

The proposed criteria are derived from and further build on the work by Broughton Micova & Calef on platform systemic risk. I apply the distinction they make between the sources and spread of risk as "propagated shocks" to systemic risk in the Artificial Intelligence Act, describing the potential for a shock through a model's capability and alignment and using "propagation" in largely the same way.

Furthermore, the criteria proposed match the risk factors of large language models identified in academic work on their risks. The capability criterion aligns with the increasing abilities of LLMs, the ways in which such abilities can nevertheless fall short and the well-known problems in assessing such capabilities. The alignment criterion addresses issues raised in research on value alignment, bias, and the general challenge of ensuring that LLM-based systems behave ethically. Finally, the propagation criterion broadly reflects concerns on large language models' deployment, distribution and use, covering the ways in which they are made accessible to the public and the ways in which their behaviour can cause real-world societal harm.

Finally, assessing LLM's risks by their capability, alignment and potential to propagate harm is consistent with Article 3(65) AIA's definition of a systemic risk. Article 3(65) defines a systemic risk as a risk specific to general-purpose AI models' high-impact capabilities with a

significant impact at Union level, where a risk is the combination of the probability and the severity of harm.[112] Assessing a model's capabilities allows one to determine how a risk is specific to its capabilities and what the potential severity of harm could be, assessing a model's alignment allows for determining the probability of a given harm, and assessing the way in which harmful behaviour can propagate could determine whether a risk has a "significant impact at Union level". Further recall that Article 3(65) specifically cites potential negative effects "that can be propagated at scale across the value chain", which further stresses the need to assess the scale at which harmful behaviour could spread.

A limitation of the proposed assessment criteria is that, by emphasizing the behaviour of a general-purpose AI model and the way in which such behaviour impacts society, they leave little to no room to account for GPAI models' negative externalities such as their resource usage. This limit however follows from the AIA's conceptualization of systemic risk as specific to models' advanced capabilities, which largely excludes externalities from the scope of the concept.

5.2 Applying the assessment criteria proposed

5.2.1 Short-term risks

The first short-term risk highlighted in this work is large language models' potential for biased or discriminatory output. While this issue is not inherently related to LLMs' advanced capabilities, highly capable models are more likely to be relied upon and would therefore be more likely to propagate bias in some cases. Propagation assessment would further involve assessing LLM's real-world usage: does their widespread usage perpetuate representational harm? Are they used on a large scale to make decisions that could negatively affect individuals, thereby creating some form of systemic discrimination? Do their users put unwarranted trust in them to be 'objective' and bias-free? Although I believe that large-scale propagation of bias through LLM's could – and should – be a key systemic risk to be assessed by GPAI model providers, this will ultimately depend on the way Article 3(65) AIA's specificity criterion is interpreted.

[112] Article 3(2) AIA.

The second short-term risk highlighted is the issue of a large language model 'hallucinating', i.e. making up plausible but false statements. This risk is arguably a consequence of its poor factual capabilities, combined with its users' tendency to rely on its statements. Most potential for harm due to LLMs' hallucinations follows from the gap between a model's perceived and actual capabilities, as perceived by its end user. While it is clear that hallucinations might lead to misunderstandings that could negatively affect individuals, I would argue that they do not inherently pose a systemic risk unless some additional mechanism exists that propagates their harm. The risk of hallucination might for example become systemic if large language models end up replacing online platforms and search engines in mediating online information flows, as this could lead to a further decrease in the truthfulness of online content and the accompanying level of online trust.

The third highlighted risk is that of large language models' energy consumption, resource usage and sustainability in general. As this issue is not at all related to large language models' capabilities or alignment, it is highly unlikely to qualify as a systemic risk. This serves as an example of the limits of both the proposed criteria and the AIA's systemic risk definition. As Article 3(65) AIA defines a systemic risk as "specific to" a GPAI model's capabilities, it leaves little room for risks that are not at all a consequence of capabilities but results from their negative externalities instead.

The fourth and final highlighted short-term risk is that of threats to democratic processes. This risk follows from large language models' advanced persuasion capabilities, poor value alignment and the possibility to effectively propagate created propaganda and disinformation through online platforms. As such risks are already a key focus of the European Commission and could plausibly become reality with current generation of frontier models, I would view it as highly likely to qualify as a systemic risk under the AIA. Interestingly, as mitigating large language models' risks to democratic processes would involve addressing both the creation and the online spread of propaganda and misinformation, the systemic risk provisions of the AIA and the DSA would apply in tandem. For such content, I would argue that its creation is primarily a systemic risk under the AIA, while its spread– presumably through online platforms – would qualify as a systemic risk under the DSA. In other words: the AIA's provisions would entail preventing shocks and the DSA's provisions would entail mitigating the propagation of shocks.

5.2.2 Long-term risks

The first long-term risk highlighted in this work is that of knowledge barrier reduction, in which large language models' advanced capabilities could hypothetically aid malicious actors in obtaining the knowledge needed to inflict catastrophic harm through for example biological weapons or cyberweapons. This is another key example of a risk following from models' advanced capabilities and poor alignment: their advanced capabilities enable them to be of use for highly complex issues, while their lack of alignment results in them not refusing to actually help for malicious purposes. For providers of highly capable open-source LLMs, propagation assessment would further entail mapping the implications of publicly distributing their architecture and weights. While the risk of knowledge barrier reduction is for now hypothetical, its potential harm is sufficiently several that I believe that GPAI model providers should reasonably be expected to assess for such risks. I therefore believe that it could qualify as a "systemic risk" under the AIA.

The second and final long-term risk highlighted is that of self-proliferation, in which an advanced large language model could autonomously develop superintelligence by recursively creating even more intelligent versions of itself. While this is certainly the most speculative risk addressed in this work and remains squarely outside of the mainstream of EU policy thinking, the text of the AIA and the risk assessment criteria proposed do in fact suggest that it could qualify as a systemic risk. It is a key example of a risk specific to a capable model's advanced capabilities and poor alignment and additionally involves propagation concerns, as it depends on a models' abilities to interface with real-world digital or physical systems. Furthermore, the AIA's recitals explicitly describe the risk.[113] Nevertheless, including self-proliferation as a potential systemic risk would have little consequences for GPAI model developers as they tend to already assess for self-proliferation risks. They would perhaps be required to think twice before allowing an advanced GPAI model to control significant computational resources, which in general would be a sensible idea. Developers should however not be allowed to devote the majority of their risk management resources to self-proliferation instead of focusing on more pressing risks.

[113] Recital 110 AIA.

6. Conclusion

This chapter briefly summarizes this work's findings to answer its main research question:

> **What risks associated with large language models could qualify as "systemic risks at Union level" under Article 55 of the Artificial Intelligence Act?**

Chapter 2 introduced key concepts in artificial intelligence regulation, outlining long-existing concerns on superintelligence, the recent development of AI systems that can competently perform a large number of distinct tasks, and the relation between modern "general-purpose" AI and the older notion of "artificial general intelligence". It further discussed the Artificial Intelligence Act's key aims and general framework, highlighting its risk-based approach intended to provide robust safeguards for public safety and fundamental rights.

Chapter 3 elaborated on technical aspects and risks of large language models, which represent the technological state of the art in general-purpose AI development. It discussed their key technical properties and the way in which they contributes to their risks and highlighted short-term risks and long-term risks as the two main categories of LLM-related risks. It further cited bias & discrimination, hallucinations, and sustainability concerns as key examples of short-term risks and cited threats to democratic processes, knowledge barrier reduction and self-proliferation as key long-term risks.

Chapter 4 focused on unravelling the Artificial Intelligence Act's legal framework for providers of general-purpose AI ("GPAI") models with systemic risk. It briefly described the AIA's legislative background and discussed the origin of the GPAI regime as a political compromise between "light-touch" regulation and fundamental rights protection. Next, it set out the AIA's regulatory framework for general-purpose AI models exhibiting systemic risk through their high-impact capabilities and discussed some open issues regarding its key definitions. criteria to identified LLM-related risks to discuss whether they could qualify as "systemic risks". Finally, it argued that existing DSA-related work distinguishing between sources and spread of risk could be applied to the AIA and that Commission enforcement priorities on the DSA suggest that it is be likely to take a broad view of the scope of the AIA's systemic risk concept.

Chapter 5 proposed capability, alignment and propagation as systemic risk assessment criteria for general-purpose AI models, arguing that the criteria described cover most technical and societal risk factors of large language models and are consistent with the system of the AIA and the DSA. Applying the criteria proposed, it concluded that risks relating to manipulation of democratic processes and knowledge barrier reduction are likely to qualify as "systemic risks" in the Artificial Intelligence Act, risks related to bias & discrimination, hallucination and self-proliferation might qualify as such, and sustainability risks are unlikely to qualify as systemic risks.

7. Sources

7.1 Primary sources

7.1.1 Legislation

Regulation (EU) 2024/XXX of the European Parliament and of the Council of x July 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
Version used: Corrigendum to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/ ..... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 19 April 2024.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

7.1.2 Official publications

European Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts', COM(2021) 206 final, 21 April 2021.

European Commission, 'Impact Assessment on the Digital Services Act' SWD(2020) 348 final, 15 December 2020.

European Commission, 'White Paper On Artificial Intelligence - A European approach to excellence and trust', COM(2020) 65 final, 19 February 2020

European Commission, 'Supervision of the designated very large online platforms and search engines under DSA' <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses> accessed May 10 2024

European Commission, Directorate-General for Communications Networks, Content and Technology, 'Digital Services Act – Application of the risk management framework to Russian disinformation campaigns' (Publications Office of the European Union 2023) <https://data.europa.eu/doi/10.2759/764631>

High-Level Expert Group on AI, 'Ethics guidelines for trustworthy AI', 8 April 2019

Press release, 'Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act' (European Commission, 14 March 2024) <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very> accessed May 11 2024

7.1.3 Case law

Case C-131/12 Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González [2014] ECLI:EU:C:2014:317 ("Google Spain")

Case C-362/14 Maximillian Schrems v Data Protection Commissioner [2015] ECLI:EU:C:2015:650 ("Schrems I")

Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779 ("Breyer")

Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH [2016] ECLI:EU:C:2018:388 ("Wirtschaftsakademie")

Case C-311/18 Data Protection Commissioner v Facebook Ireland Ltd & Maximillian Schrems [2020] ECLI:EU:C:2020:559 ("Schrems II")

Case C-634/21 OQ v Land Hessen [2023] ECLI:EU:C:2023:957 ("Schufa")

7.2 Academic sources

7.2.1 Books and contributions in books

Bostrom N, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014)

Haykin S, *Neural Networks: A Comprehensive Foundation* (Pearson 1999)

Pennachin C & Goertzel B, 'Contemporary Approaches to Artificial General Intelligence' in Goertzel B & Pennachin C (Eds.), *Artificial General Intelligence* (Springer 2007)

Russell S & Norvig P, *Artificial Intelligence: A Modern Approach*, *Fourth Edition* (Pearson 2021)

7.2.2 Journal articles

Alon-Barkat S & Busuioc M, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' [2023] 33(1) JPART 153-169

Banh L & Strobel G, 'Generative Artificial Intelligence' [2023] 33 63 Electronic Markets

Bengio Y, Hinton G & LeCun Y, 'Deep Learning' [2015] 512 Nature 436-444

Bostrom N, 'The Superintelligent Will' [2012] 22 Minds and Machines 71-85

Bryson J, Caliskan A & Narayanan A, 'Semantics derived automatically from language corpora contain human-like biases' [2017] 356 Science 183

Burdick M, Mosier K & Skitka L, 'Does automation bias decision-making?' [1999] 51(5) HCS 991-1006

Danaher J & Sætra HS, 'Resolving the battle of short- vs. long-term AI risks' [2023] AI & Ethics

Diakopoulos N & Helberger N, 'ChatGPT and the AI Act' [2023] 12 1 Internet Policy Review

Gabriel I, 'Artificial Intelligence, Values, and Alignment' [2020] 30 Minds and Machines 411-437

Garg N and others, 'Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes' [2018] 115 PNAS 3635

Good IJ, 'Speculations concerning the first ultraintelligent machine' [1965] 6 Advances in Computers, 31, 33

Gutierrez C and others, 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems' [2023] 2 36 Digital Society

Huang C and others, 'An Overview of Artificial Intelligence Ethics' [2023] 4 4 IEEE Transactions on Artificial Intelligence 800

Lazar S & Nelson A, 'AI safety on whose terms?' [2023] 381 6654 Science 138

Metikoš L, 'De AI Wet: Zwak, Zwakker, Zwakst' [2024] 3 Mediaforum 73

Novelli C and others, "Taking AI risks seriously: a new assessment model for the AI Act" [2023], AI & Society 38 4

Rillig M and others, 'Risks and Benefits of Large Language Models for the Environment' [2023] 57(9) EST 3464-3466

Schrama W, 'How to carry out interdisciplinary legal research: Some experiences with an interdisciplinary research method' [2011] 7(1) Utrecht Law Review 152

Smuha N, "Beyond the individual: governing AI's societal harm" [2021] 10 Internet Policy Review 3

Townsen Hicks M, Humphries J & Slater J, 'ChatGPT is bullshit' [2024] 26 38 EIT

The Editorial Board, 'Stop talking about tomorrow's AI doomsday when AI poses risks today' [2023] 618 Nature 885–886

Veale M & Zuiderveen Borgesius F, 'Demystifying the Draft EU Artificial Intelligence Act' [2021] 4 CRi 98

Wei J and others, 'Emergent Abilities of Large Language Models' [2022] Transactions on Machine Learning Research

7.2.3 Conference papers

Abid A, Farooqi M & Zou J, "Persistent anti-muslim bias in large language models" (AAAI/ACM Conference on AI, Ethics, and Society, New York, 2021)

Bender EM, Gebru T, McMillan-Major A, and Shmitchell S, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (FAccT Conference on Fairness, Accountability, and Transparency, Canada, 2021)

Brown T and others, 'Language Models are Few-Shot Learners', 33rd Conference on Neural Information Processing Systems (NeurIPS 2020)

Carnat I, 'Human, all too human: accounting for automation bias in generative Large Language Models' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)

Gadiraju V and others, "'I wouldn't say offensive but...': Disability-Centered Perspectives on Large Language Models" (FAccT Conference on Fairness, Accountability, and Transparency, Seoul, 2022)

Ganguli D and others, 'Predictability and Surprise in Large Generative Models' (FAccT Conference on Fairness, Accountability, and Transparency, Seoul, 2022)

Golpayegani D, Pandit H, & Lewis D, "To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023)

Hacker P, Engel A & Mauer M, "Regulating ChatGPT and other Large Generative AI Models" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023)

Hendrycks D and others, ‘Measuring Massive Multitask Language Understanding’, 9th International Conference on Learning Representations (ICLR 2021)

Lee M, 'Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans' (forthcoming in: FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)

Mirowski P, 'A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)

Nadeem M, Bethke A & Reddy S, "Stereoset: Measuring stereotypical bias in pretrained language models" (Annual Meeting of the Association for Computational Linguistics, Online, 2022)

Ouyang L and others, 'Training language models to follow instructions with human feedback', 36th Conference on Neural Information Processing Systems (NeurIPS 2022)

Perini AM and others, 'Mapping the individual, social and biospheric impacts of Foundation Models' (FAccT Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, 2024)

Shelby R and others, 'Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction' (AIES Conference on Artificial Intelligence, Ethics, and Society, Montreal, 2023)

Sheng E and others, 'The Woman Worked as a Babysitter: On Biases in Language Generation' [2019] https://aclanthology.org/D19-1339/

Soares N, 'The Value Learning Problem' (IJCAI International Joint Conference on Artificial Intelligence, New York, 2016)

Solaiman I, "The Gradient of Generative AI Release: Methods and Considerations" (FAccT Conference on Fairness, Accountability, and Transparency, Chicago, 2023)

Steed R and others, 'Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models'https://aclanthology.org/2022.acl-long.247/

Vig J and others, "Investigating gender bias in language models using causal mediation analysis" (NeurIPS Conference on Neural Information Processing System, Vancouver, 2020)

Weidinger L and others, 'Taxonomy of Risks posed by Language Models' (FAccT Conference on Fairness, Accountability, and Transparency, Seoul, 2022)

Zimmermann A, "The Generative AI Deployment Rush: How to Democratize the Politics of Pace" (Keynote, AIES AAII/ACM Conference on AI, Ethics, and Society, Montreal, 2023)

7.2.4 Pre-prints

Anwar U and others, "Foundational Challenges in Assuring Alignment and Safety of Large Language Models" [2024] https://arxiv.org/pdf/2404.09932.pdf

Bereska L & Gavves E, 'Mechanistic Interpretability for AI Safety: A Review' [2024]

Bommasani R and others, 'On the Opportunities and Risks of Foundation Models' [2021] https://arxiv.org/pdf/2108.07258.pdf

Critch A & Krueger D, 'AI Research Considerations for Human Existential Safety' [2020] https://arxiv.org/abs/2006.04948

Critch A & Russell S, "TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI" [2023]
https://arxiv.org/abs/2306.06924

Dua D and others, 'DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs' [2019]
https://arxiv.org/pdf/1903.00161

Hacker P, Zuiderveen Borgesius F, Mittelstadt B & Wachter S, 'Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It' [2024]
https://media.licdn.com/dms/document/media/D4E1FAQH4ZFDb_cBLyA/feedshare-document-pdf-analyzed/0/1719418044729?e=1720051200&v=beta&t=v7MqDgvnlNYnQBLxJyHTgSWDVF0vZmNNwAoPDB1LxA0

Hanson R, 'Economic Growth Given Machine Intelligence' [2001]
http://mason.gmu.edu/~rhanson/aigrow.pdf

Hofmann V and others, 'Dialect prejudice predicts AI decisions about people's character, employability, and criminality' [2024]
https://arxiv.org/abs/2403.00742

Kapoor S, Bommasani R and others, 'On the Societal Impact of Open Foundation Models' [2024]
https://crfm.stanford.edu/open-fms/paper.pdf

LeCun Y, 'A Path Towards Autonomous Machine Intelligence' [2022]
https://openreview.net/pdf?id=BZ5a1r-kVsf

Mattern J and others, 'Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing' [2022] https://arxiv.org/abs/2212.10678

Monserrate S, 'The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage' (MIT Case Studies in Social and Ethical Responsibilities of Computing 2022)
https://s3.amazonaws.com/assets.pubpub.org/2l4m6ikahjxvjn3wgr6aj498hji0t0cg.pdf

Naveed H and others, 'A Comprehensive Overview of Large Language Models' [2024]
https://arxiv.org/abs/2307.06435

Novelli C and others, Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity' [2024]
https://arxiv.org/pdf/2401.07348

Rein D and others, 'GPQA: A Graduate-Level Google-Proof Q&A Benchmark' [2023]
https://arxiv.org/pdf/2311.12022

Sandbrink J, 'Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools' [2023]
https://arxiv.org/pdf/2306.13952

Uuk R, Gutierrez CI & Tamkin A, 'Operationalising the Definition of General Purpose AI Systems: Assessing Four Approaches' [2023]
https://arxiv.org/pdf/2306.02889

Vaswani A and others, 'Attention Is All You Need' [2017]
https://arxiv.org/abs/1706.03762

7.3 Other sources

7.3.1 Industry & policy reports

Amodei D and others, 'Concrete Problems in AI Safety' (Google Brain and others, 2016)
https://arxiv.org/abs/1606.06565

Anderljung M and others, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety' (Centre for the Governance of AI e.a., 2023)
https://arxiv.org/abs/2307.03718

Askell A and others, 'A General Language Assistant as a Laboratory for Alignment' (Anthropic, 2021)
https://arxiv.org/abs/2112.00861

Bandt O & Hartmann P, 'Systemic Risk: A Survey' (ECB Working Paper Series no. 35, 2000) 10-11
https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp035.pdf

Bowman S and others, 'Measuring Progress on Scalable Oversight for Large Language Models' (Anthropic, 2022)
https://arxiv.org/abs/2211.03540

Broughton Micova S & Calef A, 'Elements for Effective Systemic Risk Assessment Under the DSA' (CERRE 2023)
https://cerre.eu/wp-content/uploads/2023/07/CERRE-DSA-Systemic-Risk-Report.pdf

Bubeck S and others, 'Sparks of Artificial General Intelligence: Early experiments with GPT-4' (Microsoft Research, 2023)
https://arxiv.org/abs/2303.12712

Burns C and others, 'Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision' (OpenAI, 2023)
https://arxiv.org/abs/2312.09390

Gabriel I and others, "The Ethics of Advanced AI Assistants" (Google DeepMind, 2024) https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

Hendrycks D, Mazeika M & Woodside T, 'An Overview of Catastrophic AI Risks' (Center for AI Safety 2023) https://arxiv.org/abs/2306.12001

Karl Cobbe and others, 'Training Verifiers to Solve Math Word Problems' (OpenAI, 2021) https://arxiv.org/pdf/2110.14168

Hubbinger E and others, 'Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training' (Anthropic, 2023) https://arxiv.org/abs/2401.05566

Kaplan J, McCandlish S and others, 'Scaling Laws for Neural Language Models' (OpenAI, 2020) https://arxiv.org/abs/2001.08361

Kutterer C, 'Regulating Foundation Models in the AI Act: From "High" to "Systemic" Risk' (AI-Regulation Papers 2024) https://ai-regulation.com/wp-content/uploads/2024/01/C-Kutterer-Regulating-Foundation-Models-in-the-AI.pdf

Loi M, "Making sense of the Digital Services Act: How to define platforms' systemic risks to democracy" [2023] Algorithm Watch https://algorithmwatch.org/de/wp-content/uploads/2023/08/AlgorithmWatch_Risk_Assessment-DSA.pdf, accessed 22 February 2024

Phuong M and others, 'Evaluating Frontier Models for Dangerous Capabilities' (Google DeepMind 2024) https://arxiv.org/abs/2403.13793

Rose S & Nelson C, 'Understanding AI-Facilitated Biological Weapon Development' (Centre for Long-Term Resilience 2023) https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio

Rudner T & Toner H, "Key Concepts in AI Safety: An Overview" (CSET Issue Brief, 2021) https://pdfs.semanticscholar.org/8a1c/cc06ddb9cc128e66c456a2f40d5f594c5c9f.pdf

Scheurer J, Balesni M & Hobbhahn M, 'Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure' (Apollo Research, 2023) https://arxiv.org/pdf/2311.07590

Seger E and others, 'Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives' (Centre for the Governance of AI and others 2023) https://arxiv.org/pdf/2311.09227

Sharma M and others, 'Towards Understanding Sycophancy in Language Models' (Anthropic, 2023) https://arxiv.org/abs/2310.13548

Shevlane T and others, "Model evaluation for extreme risks" (Google DeepMind, 2023) https://arxiv.org/pdf/2305.15324.pdf

Weidinger L and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (Google DeepMind, 2024) https://arxiv.org/pdf/2310.11986

'The Claude 3 Model Family: Opus, Sonnet, Haiku' (Anthropic, 2023) https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

'Llama 2: Open Foundation and Fine-Tuned Chat Models' (Meta, 2023) https://arxiv.org/abs/2307.09288

'Gemini: A Family of Highly Capable Multimodal Models' (Gemini Team Google, 2023) https://arxiv.org/abs/2312.11805

7.3.2 Blogs and news articles

Bareis J, 'BigTech's Efforts to Derail the AI Act' (*Verfassungsblog,* 5 December 2023) <https://verfassungsblog.de/bigtechs-efforts-to-derail-the-ai-act/> accessed May 9 2024

Bertuzzi L, 'EU's AI Act negotiations hit the brakes over foundation models' (*Euractiv*, 10 November 2023) <https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/> accessed May 9 2024

'Trojan horses: how European startups teamed up with Big Tech to gut the AI Act' (*Corporate Europe Observatory* 11 March 2024) <https://corporateeurope.org/en/2024/03/trojan-horses-how-european-startups-teamed-big-tech-gut-ai-act> accessed May 9 2024

'The EU AI Act: a failure for human rights, a victory for industry and law enforcement' (*Access Now*, 13 March 2024) <https://www.accessnow.org/press-release/ai-act-failure-for-human-rights-victory-for-industry-and-law-enforcement/> accessed May 9 2024