From Curators to Creators: Navigating Regulatory Challenges for General-Purpose Generative AI in Europe

by **Gabriel Ernesto Melian Pérez ***

Abstract: This study examines the regulation of general-purpose generative AI (GPGAI) in the European Union, dividing the analysis into two parts. First, it explores whether GPGAI, by generating new content, qualifies as a content provider and thus falls outside the scope of 'safe harbour' protections. Drawing on case law from the CJEU and the Digital Services Act (DSA), the paper argues that GPGAI, by actively contributing to content creation, goes beyond the role of a mere intermediary and should therefore not benefit from safe harbour exemptions. Having established GPGAI's active role in content generation, the second part of the study addresses the broader regulatory implications, focusing on the AI Act and the revised Product Liability Directive. It contends that the AI Act's risk-based approach is insufficient to address the dynamic and unpredictable nature of GPGAI, potentially leading to ineffective regulatory obligations. The paper concludes by advocating for more tailored legal frameworks to ensure the responsible development of GPGAI, striking a balance between fostering innovation and safeguarding users.

Keywords: Safe Harbour, Curation AI (CAI), Generative AI (GAI), General-Purpose Generative AI (GPGAI), AI Act.

Recommended citation: Gabriel Ernesto Melian Pérez, IFrom Curators to Creators: Navigating Regulatory Challenges for General-Purpose Generative AI in Europe, 16 (2025) JIPITEC 201 para 1.

A. Introduction

1 Recent years have seen rapid technological advancements, resulting in the deployment of various types of AI with diverse functionalities. Some are designed for specialized applications in fields such as medicine, education, and defense, while others serve more general purposes aimed at non-specialist audiences. Among these AI tools, generative models stand out as particularly remarkable. They can create entirely new and original content based on the data they were trained on (Hacker et al., 2023), pushing the boundaries of creativity and innovation. However, these same capabilities carry the potential for misuse, as the content generated may inadvertently or intentionally be harmful, ranging from misinformation to offensive or defamatory material.

2 One of these AI tools is Meta's Imagine. In broad terms, Imagine is a generative AI that creates images, in the style of the already famous Midjourney, Stable Diffusion or DALL-E. Recently, Meta has announced that Imagine's functions will be incorporated into Facebook, Instagram and Messenger, so that the user can generate images to use them in the Facebook feed, in Stories, as comments, reactions or as profile pictures. This means that Meta would implement two different types of AI on its Facebook and Instagram social network: this generative AI (GAI)[1]

* LL.M. Göttingen, PhD fellow, Civil Law Department, Pompeu Fabra University. I express my deep gratitude to Professors Antoni Rubí Puig and Migle Laukyte for their valuable comments and feedback. I also appreciate the anonymous reviewers for their comments, which contributed significantly to the improvement of this work. This research has been developed within the framework of the research project "Responsabilidad contractual y extracontractual de las plataformas en línea", supported by the Ministry of Science and Innovation, the Agencia Estatal de Investigación and the European Regional Development Fund (PID2021-126354OB-I00).

1 Because of its broad capabilities and general scope of use, this generative AI (GAI) can be classified as a General-

and the AI that organizes and curates content (CAI). This seemingly innocuous distinction could have important legal consequences. Understanding the differences between these two types of AI and their respective levels of control over the content they generate and show is crucial for determining their responsibilities and potential liability.

3 This research compares the levels of control that GAI and CAI have over the content. The hypothesis proposed is that the two AIs have different levels of control over the content, and therefore, the same legal principles cannot be applied. As generative AI performs a substantial intervention in the creation of content, it could be considered that its role is too active to benefit from the safe harbour[2]. On this premise, GAI would then be subject to other EU[3] and national rules that will determine their level of liability for the content they generate. The most relevant norms include the AI ACT and the new defective products directive. However, a general review of them reveals a number of loopholes in the regulation of GPGAIs. The aim of this paper is to address these shortcomings and to propose some *ex ante* regulatory adjustments that would better clarify what the obligations of developers of these technologies would be.

4 The first part of this paper delves into the technical elements that distinguish the two types of AI at the core of this study, so that the reader has a clear understanding of the technological background before entering the more theoretical legal framework. The second section focuses on the classification of social networks within the broader landscape of media players, examining how the concepts of control and knowledge have shaped this model. The third section analyzes the evolution of these variables (knowledge and control) in the jurisprudence of the European Court of Justice, culminating in their consolidation within the recent Digital Services Act (DSA). Building on these theoretical insights, the next section discusses the relevance of technical differences in advocating for the exclusion of GAI from the benefits of the safe harbour provision. The fifth section focuses on the assessment of the current regulation of GPGAI, suggesting regulatory clarifications and changes aimed at reducing the generation of harmful content resulting from such systems. The last section concludes.

B. Technical Framework of GAI and CAI

5 Before discussing more specific issues, it is necessary to provide a general definition of AI. This article rests on the concept developed by the Organization for Economic Co-operation and Development (OECD) and supported also by G'sell (2024) that states: "*An AI system is a machine-based system that, for explicit or implicit objectives, **infers**, from the **input** it receives, how to **generate outputs** such as predictions, **content, recommendations,** or decisions that can **influence** physical or virtual environments. Different AI systems vary in their **levels of autonomy** and adaptiveness after deployment*" (OECD, 2024). The stressed aspects are the most important ones in the concept[4]. On this basis, let us proceed to analyze the typologies of interest to us: *recommendations* (CAI) and *content*

purpose generative AI (GPGAI).

2 There are arguments for (Henderson et al., 2023; Volokh, 2023) and against (Bambauer and Surdeanu, 2023; Miers, 2023), but they focus on US jurisdiction and Section 230. Therefore, it is necessary to settle this debate within the framework of European legislation, specifically the Digital Services Act, which is the norm that defines the criteria for enjoying Safe Harbour immunity.

3 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, OJ L, 2024/1689, 12.7.2024 (Artificial Intelligence Act, hereinafter AI Act). Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final. DIRECTIVE (EU) 2024/2853 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC.

4 This definition fits perfectly with that of the IA Act in Article 3(1) and Recital 12: "... A key characteristic of AI systems is their capability to infer. This capability to infer refers to the process of obtaining the outputs, such as predictions, content, recommendations, or decisions, which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data. The techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved. The capacity of an AI system to infer transcends basic data processing by enabling learning, reasoning or modelling...". According to de Graaf and Veldt (2022, p. 806) "it better expresses two common features of AI: self-learning and/or autonomous behaviour". For Hacker (2024, p. 9), however, "distinguishing AI from traditional software will be a challenge under this definition and require a good understanding of what it means to 'infer' the AI output from input. Furthermore, a purposive interpretation of the definition will need to posit a 'sufficient degree' of autonomy for models to qualify as AI".

(GAI).

I. CAI

6 Basically, a social network is an online platform that allows users to connect with one another and share content. However, what users can see, share and do on the Platform is not completely free, as it is subject, in principle, to the rules set by the platform (and of course, also to national legislation). These rules are usually set out in the "Terms and Conditions", "Community Standards" or "Content Policies" of each platform. The process by which the platform ensures that these rules are followed is known as "content moderation"[5]. This content moderation has two dimensions:

- platforms decide what content is suitable for publication, which York and Zuckerman (2019) call "*hard control*"[6];
- then, certain parameters determine what users see in their particular feed, which would be the "*soft control*" or curation.

7 Regarding the curatorial functions, social networks don't just give users a chronological set of information provided by everyone in their network[7]. Using specialized algorithms, content is displayed through intricate design parameters programmed into an AI[8] and complemented by the activity of the users themselves: interests shown, geographic location, what their "friends" like, etc.[9]. Therefore, these algorithms determine what content will be shown to users and in what order. According to the OECD definition, we could a priori classify it as a *recommendations* AI.

8 When one first creates an account, the content that is displayed can be quite random. However, as one engages with the content and other users, the algorithm will use this information to provide you with more tailored content. Essentially, the more you interact on the network, the more information the AI will obtain from you and the more personalized the experience will be (Chander and Krishnamurthy, 2018; Sylvain, 2021). Arguably, social network platforms differ from each other mainly by the content moderation they perform. This is why Gillespie (2018, p. 201) rightly argues that content moderation "*is central to what platforms do, not peripheral... is, in many ways, the commodity that platforms offer*"[10]. In fact, users opt for one platform or another mainly based on the choices made by these companies about the content they display.

II. GAI

9 Generative AI has been a revolution in the artificial intelligence landscape. It refers to "*a category of deep-learning models that are "trained" on extensive datasets and that can then be directed to generate content based on the data on which they have been trained*" (G'sell, 2024, p. 31) . Broadly speaking, generative AI usually works in response to an initial 'prompt', either a text

5 To Grimmelmann (2015, p. 47), moderation is "*the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse*".

6 This would be *ex ante* moderation, whereby the platform uses algorithms to determine whether the content to be uploaded complies with the content policies. For example, in a platform that only allows videos of pets, the system would prevent the uploading of videos about cars. Given today's information flow, it is impossible, or prohibitively costly, for humans to perform this function. To do so, implementation of a sufficiently competent AI to identify between pets and cars is required.

7 Although some platforms are currently implementing this functionality.

8 "Ranking algorithms often factor in machine-cognizable information about content, like whether machine learning models predict that an image includes nudity... Overall, the goal of ranking algorithms is to prioritize material according to content-based attributes like subject matter, relevance, or authoritativeness" (Keller, 2023a). However, they are not perfect and often tend to make mistakes when assessing these attributes (Llansó et al., 2020).

9 The displayed result (recommendation, ranking) is nothing more than the conjunction of design features chosen by the platform plus the behavior of the users (Llansó et al., 2020, p. 15). A recommender system is an algorithm designed to sift through a vast array of items and identify which ones to present to a user. These systems serve as essential tools in managing the overwhelming volume of content generated daily, assisting users in discovering relevant and personalized recommendations. *"Services like photo-sharing and community site Flickr, or Amazon.com's community ratings system, take inputs from millions of users in the form of ratings, tags, and engagement (e.g., via analyzing what and how much users click, comment on, or forward to their friends) to make the online experience better"* (Ziniti, 2008, p. 592). *"Internet platforms and services do not just show us information randomly—they organize, curate, and manage information for us... These platforms generally purport to be showing us information that we want to see based on a complex formula that takes into account our past information consumption habits combined with the habits and preferences of others... Because these formulas are proprietary and central to their business models, platforms do not share many details about how they make these decisions"* (Land, 2019, p. 290). It is difficult to know exactly how these systems work because the algorithms used by each platform are trade secrets. See (Thorburn, 2022).

10 In the same line, see Elkin-Koren, De Gregorio and Perel, 2021 (p. 987).

sentence or an image. This prompt is the guidance that instructs the generative AI system to produce certain content, which can consist of text (such as those provided by OpenAI's ChatGPT or Google's Bard), images (such as those created by Stability AI's Stable Diffusion or Meta's Imagine AI) or even videos and music.

10 The fast-paced development we have been experiencing lately in generative AI is essentially driven by three key factors: the availability of big data, high computational power and the development of new models[11]. The confluence of these three critical factors has driven AI progress: big data provides extensive training information, increased computational power enables faster and more complex processing, and innovative AI models and architectures have led to breakthroughs in various domains, including natural language processing.

11 From a technical point of view, generative AI is based on machine learning and training on huge data sets. This training allows the system to learn patterns and relationships in the data, which it can then use to generate *new* content similar in style and structure to the data it was trained on. To do this, GAI makes use of artificial neural networks (ANNs), which are a key building block of many generative AI systems[12]. ANNs try to imitate human neural networks, a kind of digital brain, with interconnected nodes that process information. Each 'neuron' receives information, performs calculations and transmits the result to other neurons at the next level. Through training, these nodes are adjusted to learn patterns and relationships in the data[13]. These neural networks and deep learning are closely related; in fact, deep learning is a sub-area of machine learning that uses neural networks, especially deep neural networks. These additional layers allow networks to learn more complex representations of data, which is especially useful for tasks such as image recognition, natural language processing and text or audio generation. (Hacker et al., 2023).

12 This whole process benefits from big data, that is, the progressive accumulation of large amounts of data. The specific type of data employed in training a large model determines its functional model:

- **Generative text AI (Language Models):** The text generation process is based on the prediction of the next most likely word or phrase, given the previous sequence of text.
- **Generative image AI (Text-to-Image Models):** Image generation is more complex, as it involves translating a textual description into a visual representation. They can use techniques such as generative antagonistic networks (GANs)[14], diffusion models or transformers to create images that match the textual description (Noto La Diega and Bezerra, 2024).
- **Multimodal generative IA:** It is designed to process and generate multiple types of data, such as text, images, audio and video. For example, OpenAI's GPT-4o accepts combinations of text, audio, images and video as input and generate any of these formats as output. These capabilities make multimodal models highly versatile and useful in applications that require understanding and generating information in a variety of formats (G'sell, 2024).

13 The development of new artificial intelligences and their performance is also benefiting from technical developments, notably the introduction of the Transformer architecture by Google researchers in 2017 (Vaswani et al., 2017) and improvements in Generative Pre-Trained Transformer (GPT) models

11 "In sum, an AI model is a program trained on a large set of data with the ability to identify patterns in that data in order to produce relevant outputs in response to inputs without the need for human intervention" (G'sell, 2024, p. 32). "AI models include, among others, statistical models and various kinds of input-output functions (such as decision trees and neural networks)... AI models can be built manually by human programmers or automatically through, for example, unsupervised, supervised, or reinforcement machine learning techniques" (OECD, 2024, p. 8).

12 Although ANN-based models dominate the market, there are other types of AI that are not based on neural networks.

13 "Determining the model's size mainly involves determining the number of parameters or weights it will include. "Weights" are the numerical values that determine the strength of neural connections within a neural network and, thereby, help determine a model's output. During the training process, these weights are adjusted to optimize the model's performance, helping it produce more accurate and useful outputs. Furthermore, the relationship between the size of the model and its performance is mediated by the model's topology. "Topology" refers to the arrangement of neurons and layers within the neural network and how they are interconnected" (G'sell, 2024, p. 44). Although it is worth clarifying that more parameters do not mean that a model is better, with better performance, as it is currently more of a priority to synthesize those nodules to make it lighter. Models with many parameters tend to require more computational resources, both to train and to use.

14 These networks consist of a generator, which creates images, and a discriminator, which evaluates whether they are real or fake. During training, the two compete: the generator gets better at fooling the discriminator, and the discriminator learns to detect fake images. Over time, the generator produces images that are indistinguishable from the real ones. (Noto La Diega and Bezerra, 2024)

since around 2019 (Belcic and Stryker, 2024; Radford et al., 2019). Several companies have leveraged the use of transformer models[15], which have become popular because of their ability to process data streams and generate images with refined detail and contextual coherence.

14 Although each technology has its specific methods and approaches, the fundamental principles of pattern learning, progressive generation and fine-tuning are applicable to all image generative AI. In this sense, generative AI uses the elements and structures learned from the data to generate new combinations and variations[16].

C. Evolution of Intermediary Liability

15 Before the advent of the Internet, the most well-known media and intermediaries were broadcasters, newspapers, telephone networks and bookstores. These entities were placed into one of three traditional intermediary liability models: publishers/content providers (newspaper), distributors (libraries, bookstores), and conduits (telephone companies) (Patel, 2002, p. 651; Volokh, 2021, p. 454). In the 1990s, following the Internet boom, regulators were faced with the problem of assessing whether the liability of new web entities would be based on those traditional models or whether they were worthy of a new approach.

16 The basic principle underlying the question of whether an agent qualifies as a publisher, distributor or conduit is closely related to the idea of *control* exercised over the content; the more control is exercised, the more responsibility should be attributed to the agent. According to this model:

- Newspapers are subject to publisher liability because they have full editorial control over the content of their columns and articles.
- On the other hand, telephone, mail or courier companies are understood to have a very low share of responsibility. They have no control over what is discussed in a phone call or what is sent in a letter, therefore, it was understood that the fairest solution would be not to hold them liable for third parties' illegalities. Therefore, they fall into the category of common carriers or conduits, which refers to an entity acting as a passive conduit of illicit content (Candeub, 2020, p. 410).
- Meanwhile, libraries are somewhere in between the publisher and the conduit. They have "distributor liability" as they are entities that distribute third-party content, do not exercise editorial control over such content, but have some access to it. It has been understood that requiring distributors to review the content they distribute for illegalities would be an unjustifiably heavy burden[17]. Therefore, they are only liable for the illegal content they distribute if they become aware of such illegality.

17 To differentiate the distributor from the publisher, the latter is sometimes referred to as the primary publisher, since it is the publisher who exercises editorial control over the content, while the distributor is known as the secondary publisher or "re publisher", since its function is not to control the content, but to make it available to others without performing any creative or editorial function (Mirmira, 2000, p. 439; Patel, 2002).

18 With the emergence of the Internet, the rise of blogs first, then discussion forums and now with web 2.0 and social networks, it became clear that these new types of media were not easily categorized into the

15 "The transformer architecture marked a significant turning point for deep learning, particularly in the areas of natural language processing and computer vision. It enabled a huge leap in the amount of data that AI models could leverage and resulted in increased performance... The two most popular types of transformers are generative pre-trained transformers (GPT) and bidirectional encoder representations from transformers (BERT). OpenAI has used GPT to develop GPT-3 and GPT- 4, while Google has refined BERT to develop Bard (now called Gemini)" (G'sell, 2024, p. 34).

16 There is a large body of scholarship debating the impact of this issue on copyright. Due to the involvement of multiple parties in the outcome, it is difficult to establish from a copyright lens who should be considered the creator of the work. For example, Khosrowi et al. (2024) argue that "*GenAI outputs are created by collectives in the first instance. Claims to creatorship come in degrees and depend on the nature and significance of individual contributions made by the various agents and entities involved, including users, GenAI systems, developers, producers of training data and others*". Viewpoints such as these can contribute to reinforcing GAI's involvement in shaping content.

17 *Smith v. Cal.*, 361 U.S. 147, 153 (1959) ("*[I]f the bookseller is criminally liable without knowledge of the contents... he will tend to restrict the books he sells to those he has inspected*"). This conclusion, supported by Justice William Brennan, is reasonable since the distributor would try as much as possible to avoid any liability. This would have an undesirable impact on fundamental rights, specifically the right of access to information. If we were to expect every piece of content to pass through the distributor's filter, the content available to the public would be only that which the distributor has had time to review and approve, leading to the unavailability of legal content and a substantial risk of private censorship and false positives.

three traditional groups mentioned above. While publishers and social media platforms consist, broadly speaking, of making decisions on what content to show to users and in what order, some preliminary distinctions may include the following:

i Traditional media companies perform *ex-ante* moderation, based on editorial guidelines, before content is broadcast or published. Moderation in social networks generally operates *ex ante* and *ex post* and the review of such content is performed by computer tools, which do not understand the content in the same way that a human does (Keller, 2023a). "*... editors (human) and recommendation systems also differ in many regards, including that recommendation systems are automated and process third-party content, and as a result are generally less intentional or deliberate about overall outcomes... The effects of recommended content are highly unpredictable*" (Llansó et al., 2020, p. 16).

ii The moderation that is carried out on platforms is not comparable to an editorial process of those carried out in traditional media. Traditional media focus on keeping people informed on various topics, for which the information goes through several levels of fact-checking. In addition, this information is provided by licensed professionals who are guided by codes of conduct and ethics. This endows the information with a degree of reliability that is not associated with the content uploaded by users to social media. Traditional media "endorse" the content they publish after going through this editorial process. However, Internet platforms have never pretended to "endorse" the contents they host or claim authorship over them, because it is not "their speech", but that of the users (Zurth, 2020, p. 1145). The purpose of social networks is to share content provided by the users themselves, which are not platform employees. That content generally comes without centralized editorial oversight or planning (Elkin-Koren et al., 2021, p. 1033). The mere existence of a recommendation/curation system on the platforms does not necessarily mean that they have knowledge of the content of a particular item[18].

iii Traditional media only support one-way communication. On the contrary, social media lets people communicate in two-way. It means unlike traditional media, social media users can leave reactions, comments, etc. "*Twentieth-century print and broadcast media were not participatory media; the vast majority of people were audiences for the media, rather than creators who had access to and used the media to communicate with others. Twenty-first century model, by contrast, involves crowdsourcing and facilitating end user content. Social media host content made by large numbers of people, who are both creators and audiences for the content they produce.*" (Balkin, 2021, p. 75)

19 These aspects reveal an important element: social media platforms do not *control* content like an editorial desk would. The editor is responsible for knowing the content of any article that will be published and has the power and resources to control and approve the content before it is published. It can be said that content management in the case of newspapers and broadcasters is more *conscientious*, while on the platforms it is more superficial[19]. Platforms have no control over the accuracy or fairness of the content that users produce and upload.

20 So, what are social networks? Participatory networking platforms are, broadly speaking, *online platforms* that allow users to connect with each other to share content and communicate. These platforms also allow companies to connect with their customers and get feedback from them to improve their products and services. From this description,

18 See Keller (2023b, 2023a) and "Brief of Center for Democracy & Technology and 6 technologists as amici curiae in support of respondent", case Gonzalez v. Google LLC, 598 U.S. 617 (2023).

19 As said, the mere existence of a recommendation system on the platforms does not necessarily mean that they have knowledge about the content of an item. The algorithm makes decisions about what to do based on *signals* or elements it identifies in that content and the behavior shown by users (Leerssen, 2020). This process is perhaps sufficient to detect content that the user may like. However, it may not be sufficient to gather the necessary elements to determine the illegality of such content; this process is more complex and requires more information and intellectual capabilities. This curation process is not equivalent to "understanding" or "knowledge", at least not from the perspective of how a human would process a given item of content (Keller, 2023a, 2023b; Llansó et al., 2020). The algorithm can identify, for example, that an image contains nudity, however, it might be unable to differentiate between nudity occurring within the realm of artistic expression and that which signifies abuse. The inherent complexity of these situations underscores the need for human judgment and contextual understanding. This is explained very well by Keller (2023b): "*algorithms don't "know" what message a post conveys in the way a human would. That's why they make mistakes humans might not, like assuming any image with a swastika is pro-Nazi. In that narrower sense, one could perhaps argue that algorithms are not considering "content" but, rather, "data" or "signals" about the content*". That is why it is not the same whether a content is analyzed by a human editor of a magazine or by an algorithm within the framework of a platform.

one can conclude that, in principle, social networks do not create content themselves, but by means of AI, they organize and structure the information that third parties upload to the platform[20]. Based on what has been mentioned so far, we can place them in the broad category of hosting, specifically online platforms. Content hostings are generally associated with the liability of a distributor or secondary publisher, like that of a bookstore, which is triggered upon notification of illegal content[21]. In principle, the basic requirement to avoid liability would then be that the platform does not control or is responsible, in whole or in part, for creating or developing content and has no knowledge of the illegality of that content (Kosseff, 2022; Pagallo, 2011; Patel, 2002).

D. The Safe Harbour Doctrine

I. CJEU Case Law

21 In the 1990s, the European Union found itself in the same conundrum as other jurisdictions: it was unclear what standards of liability to apply to those new intermediaries that were emerging in the context of the Internet. The fact that each member state applied different standards was detrimental to the harmonization of the European internal market. Therefore, the European Union decided to harmonize the field by enacting the "*Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')*" or eCommerce Directive (ECD)[22]. One of the three pillars on which the ECD was based was the 'Safe Harbour' doctrine. This liability regime shields "online intermediaries" from liability for the content they transmit and host under specific conditions. Hosting firms are obligated to take down illegal content upon notification of their existence, that is, they are not liable for illegal content or activities unless they possess "actual knowledge" of them.

22 Section 4 (Arts. 12 to 15) shields certain online intermediaries (Mere conduit, Caching, Hosting) against claims that may arise from the transmission or storage of information *provided* or requested by their users. Although the goal of unifying the Internet intermediaries' liability rules was relatively achieved with the enactment of the ECD, the implementation and interpretation of this Directive in the different states was not homogeneous, perhaps due to its lack of clarity in some important points. For example: more precision was needed to establish when a platform played an "active role" and when exactly the "actual knowledge" was obtained. Over the years, the CJEU would clarify these issues, although in some cases it would make it even more blurred. The most relevant rulings could be the following:

i C-236/08 to C-238/08 *Google France*: Clarifies the applicability and preconditions for the liability exemptions of the ECD. In this case, the CJEU develops an argument focused on the active/passive role of online service providers, criterion that would permeate subsequent judgments. According to the court's interpretation, based on Article 14 and recital 42, a provider can only be exempt from liability if it has played a neutral and non-active role, meaning that it has no knowledge or control over the data stored. If its action is neutral, i.e. technical, automatic and passive, which indicates a lack of knowledge or control over the data it stores, then it could benefit from immunity.

ii C-324/09 *L'Oréal vs. eBay*: This is another case that involves determining whether the intermediary's actions are sufficiently active

20 This organization and structuring of content has been the most complex dimension of regulating social networks. Today's recommendation algorithms are so complex and advanced that they challenge the traditional distinction between publisher and mere distributor, as it is sometimes difficult to determine whether an intermediary cross the threshold of control, especially because of the intense content moderation work they perform. See Recommendation CM/Rec (2011)7 on a new notion of media which states in paragraph 25 that "*it should be noted that different levels of editorial control go along with different levels of editorial responsibility. Different levels of editorial control or editorial modalities (for example ex ante as compared with ex post moderation) call for differentiated responses and will almost certainly permit best to graduate the response".* The author recognizes the multiple challenges involved in moderating content on social networks. However, a critical assessment of this issue is beyond the scope of this article. The description of the content moderation process made here is meant to provide a framework to highlight the differences between the two types of AI examined in this paper.

21 "*An owner of a bookstore cannot be held responsible for the content of each and every book in her store. She does not read and inspect all the books. Similarly, it can be argued, an Internet provider should not be held accountable for content on its server. But if a bookstore owner is informed that a specific book contains child pornography, some other illegal material, or material that violates copyright, and she does not take the book out of the shelves, then the owner may be held legally responsible for violation of the law*" (Cohen-Almagor, 2010, p. 387)

22 Prior to the enactment of the ECD, there were two European countries that took the lead in regulating liability exemptions. In 1997, Germany adopted the IuKDG with a system based on knowledge. Sweden also did so shortly thereafter. (Husovec, 2023, p. 890)

(which allows him to have knowledge or control of the data stored) to deprive the intermediary of immunity. Accordingly, a platform can be held liable for trademark infringements committed by its users if it plays an active role that allows it to have knowledge of or control over the stored data. This is the case, for example, when the platform provides assistance to optimise the presentation of promotional offers or promotes them (Paragraph 123). In these cases, eBay does not limit itself to technical and automatic data processing but is actively involved in the management and presentation of the information. Another question that the court seeks to clarify is the scope of monitoring measures that can be imposed on intermediaries, a recurring matter in preliminary rulings. According to the CJEU, an active monitoring of all the data of each of its customers in order to prevent any future infringement would be precluded by EU law. Hence, the CJEU sets the boundaries of the notion of specific—and general—monitoring obligations by noting that ISPs can be only ordered to prevent further infringements by the same seller in respect of the same trademarks[23].

iii C-291/13 *Papasavvas*: This judgment further explores the role of the provider regarding the disputed content as the criterion for assessing whether it falls within the scope of Articles 12-14. The ruling draws an important distinction between the categories set out in articles 12-15 of the ECD and the *O Fileleftheros* newspaper, which, as *content provider*, has knowledge of the information it posts and exercises control over it. In this case, a newspaper company had a website that posted an online version of their articles. The Court ruled that the company had *knowledge* and *control* over the information posted on their website, making it ineligible to be considered a neutral intermediary service provider. If the articles on their website included illicit information, they should be held accountable for it. It would be a different matter if the newspaper provided a section on its online page for users to comment. Regarding these comments, the platform would not be considered a content provider but should be subject to the distributor regime. In this way, it clarifies the difference between a third-party content host and a newspaper that publishes and controls its own content, for which it is liable.

iv Joined Cases C-682/18 and C-683/18 *YouTube and Cyando*: In this case the court re-emphasizes that a platform cannot be compelled to introduce a screening system which entails general and permanent monitoring, because this would be contrary to Article 15 of the ECD. It also clarifies that in order for knowledge to materially arise, and immunity to be removed, the provider must be notified of an infringement in a concrete and precise manner, so that it can verify it without an in-depth legal and material examination. Therefore, a superficial notification is not enough to remove the intermediary's immunity.

II. The Digital Services Act (DSA)

23 Concerning liability, there was a certain consensus that hosting service providers should not be liable for illegal content shared through their services until they had actual knowledge. The DSA rightly maintains the principles of the ECD, reproducing in its articles 4, 5 and 6, content almost identical to the previous Articles 12, 13 and 14. While doing so, some adjustments are made considering the jurisprudence of the CJEU discussed in the previous sub-section[24].

24 One of the most relevant clarifications in the recitals are those related to the "active role", as the essential element to assess the liability of intermediaries. The DSA borrows the old formula of neutrality/passivity from recital 42 of the ECD. Recital 18 states that:

> "*The exemptions from liability established in this Regulation should not apply where, instead of confining itself to providing the services neutrally by a merely technical and automatic processing of the information provided by the recipient of the service, the provider of intermediary services plays an active role of such a kind as to give it knowledge of, or control over, that information. Those exemptions should accordingly not be available in respect of liability relating to information provided not by the recipient of the service but by the provider of the intermediary service itself, including where the information has been developed under the editorial responsibility of that provider*".

25 Therefore, collaboration or authorship implies not acting neutrally, in which case it would not be eligible for the safe harbour[25]. In order to nuance this statement and provide greater clarity, the regulation specifies in recital 22 that "*... the fact that the provider automatically indexes information uploaded to its service, that it has a search function or that it recommends information on the basis of the profiles or preferences of the recipients of the service is not a sufficient ground for considering that provider to have 'specific' knowledge of illegal activities carried out on that platform or of illegal*

23 On active content filtering measures, see also C-70/10 *Scarlet Extended vs. SABAM* and C-360/10 *SABAM vs. NetlogNV*.

24 See recital 16.

25 See also recital 20 and article 6(2).

content stored on it"[26]. This clarification is made to eliminate any uncertainty about the immunity that social media companies enjoy for their work in curating and displaying the information uploaded by users of the service.

III. "Active Role" as Threshold and How it Should be Understood.

26 In Europe, the case law discussed concludes that intermediaries could be held liable only if they actively and significantly contribute to the infringement. The fundamental premise is that a provider of services cannot govern the content that is transmitted and, provided it refrains from engaging in any editorial intervention, it should not be held liable for any unlawful content that individuals post via its services.

27 I think we can agree that the implementation of recommendation algorithms in social networks cannot be considered as playing an "active role" or exclude them from the "safe harbour" (Angelopoulos, 2017; Arroyo Amayuelas, 2020; Pagallo, 2011; Sartor, 2017; Valcke et al., 2017; Van Eecke, 2011; Van Hoboken et al., 2018). The factor that makes an intermediary acquire an active role is whether the content is third-party or can be attributed to the platform because it had some involvement in its creation. When the task consists only of optimizing the presentation of the content uploaded by users employing algorithms, it should not be understood that an active role is acquired or that the platform automatically endorses or makes the content its own[27].

28 However, as proposed by Arroyo Amayuelas (2020, p. 817) "*it would be better to abandon this distinction between "active role" and "passive role" when qualifying hosting providers and replace these expressions with other more accurate terms, such as "degree of control", "performance of editorial functions", or "effective knowledge"*". The distinction between active and passive roles in the context of hosting providers' responsibilities may oversimplify this highly complex issue. This binary categorization fails to capture the nuanced spectrum of involvement and liability that service providers navigate in today's interconnected online environment. The suggested alternative terms offer a more granular approach to understanding the varied roles and responsibilities of these entities. By shifting the focus towards aspects such as control and editorial functions, regulators and policymakers can develop more comprehensive and adaptable frameworks.

IV. Comparison

29 As indicated by the legislation and case law analyzed, to assess whether an operator can benefit from the safe harbour, it would be essential to assess two key factors: (i) their level of knowledge and (ii) their degree of control over the content. Therefore, these are the two variables we should consider relevant to assess whether an operator has adopted an active role, enough to lose this benefit. "Knowledge" would refer to an entity's ability to know, be aware of and understand a piece of content. On the other hand, the concept of "control" implies the ability of a system to directly influence or determine such content. It implies that an agent has the ability to modify or adjust the resulting content through instructions, rules or configurations. On this basis, let's analyze how this works in each context.

1. Knowledge and Predictability of the Outcome

30 CAI works after the content is created. It organizes, filters, and selects pre-existing content without altering or modifying its original form or meaning. Here, AI does not have semantic understanding[28] of the content beyond the parameters used to classify and suggest it. In fact, its results depend on metrics of relevance, popularity and personalization. Although the arrangement of content has a certain impact on its visibility[29], this is not considered to constitute modification of the content itself. For instance, rearranging search results or grouping articles by topic affects their visibility but does not change their substantive content. Therefore, an AI that merely curates content qualifies for safe harbour protections.

26 This notion seems to have been borrowed from the CJEU judgment on YouTube and Cyando, para. 114.

27 A good example is described in the Amicus curiae of *Gonzalez v. Google*, presented by some internet law scholars in support of Google: "*The more apt analogy, which supports Respondent in this case, would be the difference between YouTube simply saying "Here are the videos we have picked and chosen for you based on your interests" (or a shortened version of that, such as "You might like . . . .") and one that consisted of the words "John Smith is a Murderer, Watch this Video to Learn More!" The former involves just the statutorily protected filtering, picking, and choosing, with a statement that YouTube has filtered, picked, and chosen. The latter involves the software adding defamatory material of its own, and not just filtering, picking, and choosing.*"

28 See 6.

29 It's important to note that while content organization AI may not modify the content itself, the way it arranges and presents information can significantly impact user perception and interpretation.

31 Meanwhile, GAI interprets and applies patterns learned from training data to create new content. However, it cannot be claimed that as a computational tool, it 'knows' or is aware of what it is generating, as current AI systems lack self-awareness or comprehension. This is due to the essentially stochastic nature of how these tools work, meaning it involves an element of randomness or unpredictability. This stochastic nature is a fundamental characteristic of generative AI models, which allows them to produce diverse and creative outputs. The inherent randomness allows generative models to produce different outputs from the same input. Generative AI can also produce incorrect or nonsensical information, a phenomenon often referred to as "hallucinations" (Noto La Diega and Bezerra, 2024).

32 Since the variable 'knowledge' is not so relevant to the argument, it is argued that the degree of control over the outcome should be considered as a defining factor.

2. Degree of Control and Influence on the Content

33 In CAI, control is limited to technical aspects, such as sorting or prioritizing according to general criteria. Thus, AI does not control the content itself but its visibility or availability. In contrast, GAI represents a significant shift. It plays a substantial role in content creation, depending on the model and its design[30]. Generative AI tools, such as Imagine AI, contribute to the creation of new content by combining, transforming, and synthesizing data.

34 When viewed on a spectrum, this collaboration in content creation positions generative AI closer to a content provider than a mere distributor. Even though this type of AI does not 'understand' what it creates in a conscious sense, its intervention is active: it responds to user input and generates information that did not previously exist[31]. Consequently, GAI exercises greater control over the final outcome compared to CAI.

35 Its influence comes in several ways. Developers shape AI behaviour through model design, training processes and data selection (Henderson et al., 2023). Ultimately, they control the dataset that serves as the core for the model's content generation[32]. This gives them, to some extent, the ability to limit certain outcomes or influence the likelihood of specific results emerging[33]. The influence of developers extends beyond initial model creation and data selection, as they can also implement safeguards and filtering mechanisms to further refine AI outputs[34]. These measures can help mitigate potential biases or undesirable content, though their effectiveness may vary. Additionally, developers and deployers can continuously update, and fine-tune models based on user feedback and emerging ethical considerations. However, their control is limited, as outputs depend heavily on user prompts, and adversarial attacks can bypass filters.

36 To summarize, GAI and CAI operate with different purposes and capabilities. Having analyzed how each works, we can conclude that GAI can produce *new* content based on patterns learned from large volumes of data. However, AI that organizes and curates content in social networks does not create new information; instead, it classifies, filters and recommends existing content using algorithms based on user history, relevance, trends, or similar parameters. In other words, in the case of social network, AI could be responsible for the organization and arrangement of content, but not for the content itself, which is created by the users of the network[35].

30 Here we anticipate that not all generative AIs are the same, so the benefit of the safe harbour will depend on the specific case. This issue will be further elaborated in the next section.

31 "*This task combines forecasting and recognition tasks. However, the output often combines several existing elements such as images, text and audio to produce an object that was never seen before*". (OECD, 2022, p. 52)

32 As recognised by the OECD (2024, p. 6), "*Human intervention can occur at any stage of the AI system lifecycle, such as during AI system design, data collection and processing, development, verification, validation, deployment, or operation and monitoring*". In its 2022 version they explain that "*The lifecycle encompasses the following phases that are not necessarily sequential: planning and design; collecting and processing data; building and using the model; verifying and validating; deployment; and operating and monitoring*" (OECD, 2022, p. 7). As will be explained in Section 5, each of these phases should be subject to specific obligations.

33 "So just as these base models might identify associations that do not exist, they might successfully recover harmful associations present in the training data. Major training datasets have been shown to include websites with harmful hate speech and disinformation". (Henderson et al., 2023, p. 603)

34 See Section 5.2.

35 It should be noted that, despite the conceptual simplification made here to support the argument, also a CAI could exceed the passivity threshold defined by the CJEU. This could arise if: a recommendation algorithm systematically prioritizes illegal or infringing content, especially when the platform knows or should know that such content is problematic, or when the platform already has actual knowledge that certain sources are "of a dubious nature" (i.e. known for breaching rights) and still allows or encourages their visibility. This could be interpreted as an active role. In essence, not all CAI can be considered automatically passive under the current

Whereas in GAI, the model is much more involved in the configuration of the content created. It can be said that it acts as a kind of co-creator or contributor to the outputs and therefore could be considered a content provider.

E. GAI does not Fit into the Safe Harbour

37 Having analyzed all the historical, legal and technical aspects, it only remains to assess whether generative AI can benefit from the 'Safe Harbour'.

38 There are arguments in favor. Since generative AI lacks intention or knowledge in the human sense, some might argue that it could benefit from safe harbour protections, similar to CAIs, on the basis that it is merely a tool responding to user instructions without directly understanding the content. Authors such as Botero Arcila (2023), Stalla-Bourdillon (2023), Miers (2023), Bambauer and Surdeanu (2023), suggest that GAI products like ChatGPT share functional similarities with tools such as search engines and predictive technologies like autocomplete. These similarities stem from the foundational purpose and operation of these systems: to process user input and generate output aligned with their queries or prompts. In essence, they argue that the entire process is grounded in probability calculations, what could be described as "statistical inference". Both CAI and GAI rely on identifying patterns in existing content and producing outputs consistent with user needs, whether it be a social media feed or a generated image[36].

39 However, this argument is open to counterarguments based on the qualitative difference between organizing existing content and creating new content, which increases the likelihood of liability. DSA' Recital 18 states: "*... Those exemptions should accordingly not be available in respect of liability relating to information provided not by the recipient of the service but by the provider of the intermediary service itself, including where the information has been developed under the editorial responsibility of that provider*". Editorial responsibility implies that the provider makes active decisions about the content, its creation, development or modification[37]. In other words, they will not be able to invoke the safe harbour for information that they themselves have helped to create or develop.

40 As Perault (2023) rightly points out, but on the basis of section 230, that relevant question will be whether GAI 'develop' content, at least 'in part' to the extent that it can control or influence the outcome. Although drafted differently, the DSA and section 230 appear to follow the same line of reasoning[38]. "*The immunity extends to those who merely host or pass on information created by others. That's not necessarily true of generative AI*" (Henderson et al., 2023, p. 622). Content creation can then be considered a more 'proactive' act compared to organizing pre-existing content, which may make it difficult to argue that generative AI operates in a neutral way.

41 This paper does not argue that GAI alone is the author of the content; clearly, the user's prompt plays a significant role. However, a relevant contribution to the creation of content may suffice to disqualify a provider from safe harbour protections (because it has some level of control over the content). This would likely apply to a tool like Meta's Imagine, which transforms text prompts into entirely new images. As Perault (2023) observes, Twitter does not draft tweets for its users and thus qualifies for legal immunity. By contrast, using a *contrario sensu* interpretation, if Twitter were to assist in drafting tweets, it might lose that immunity[39]. This is precisely the scenario with Meta's Imagine function, which actively contributes to the creation of content by generating images based on user input.

42 While this is the general notion, this conclusion should also be nuanced on a case-by-case basis and

legal framework. Depending on the design and operation of the algorithm, a CAI could be held liable if it crosses the thresholds discussed here.

36 Authors from Europe who argue that GAIs are similar to search engines, such as Botero Arcila (2023) or Stalla-Bourdillon (2023), generally seek to have the due diligence obligations of the DSA, namely transparency and systemic risk assessment and mitigation, extended to them.

37 It should be emphasized that we are not referring here to basic system design decisions (e.g., how the interface is structured or how the model responds to prompts). This does not necessarily imply that the provider has editorial control over the generated content. Control neither is merely selecting or organizing third party information (what CAIs do). Editorial implies a certain degree of involvement in the outcome.

38 This analysis would perhaps be easier if the DSA had introduced a definition of 'content provider' as Section 230(f)(3) does: "*The term "information content provider" means any person or entity that is responsible, in whole or in part, for the creation or development of information provided through the Internet or any other interactive computer service*". In the United States, the analysis is facilitated by the availability of the 'material contribution test', developed by the case law to determine what degree of contribution to the content makes you cross the threshold, turning you into a content provider.

39 At this point, X (formerly Twitter) has arguably crossed that threshold by deploying Grok, its AI model that answers user queries. As a GAI, Grok produces new content based on patterns learned from large volumes of data.

not be treated in absolute terms, as there are different types of GAIs. In this respect, the distinction made by Henderson et al. (2023) and G'sell (2024) between *extractive* models, which are based on extracting information from third party sources, and *abstractive* models, is particularly relevant. *Extractive* models directly use and reproduce content from third-party sources without significant alteration. Their output is a direct reflection of the input data, making it easier to trace the origin of the information. *Abstractive* AI models, on the other hand, are designed to generate new content based on a deeper understanding of the input data. Unlike extractive models, abstractive systems create their own summarized representation of information, enabling them to reformulate, restructure, or combine ideas. As a result, extractive models are more likely to qualify for safe harbour protections, whereas abstractive models, due to their transformative nature, are less likely to do so (Volokh 2023, p. 496).

43 Another element to consider would be the use of synthetic data created by the company itself to train its AI. If the company is found to have played a substantial role in shaping the content on which the model is trained, its impact on the outcome of the model could be considered to be even higher. The manner in which data is incorporated into the model's training through the use of proprietary data sources could strengthen the argument that the company is actively shaping the content. In essence, the more a company is involved in the curation, editing, or creation of data, the more difficult it becomes for immunity to be applied (Henderson et al., 2023).

44 In summary, under current EU regulations, GAI would have to be excluded from the safe harbour benefit because of its active role in content creation. It would be a matter of degrees, the more 'expressive' and 'creative' the AI is, the more influence it can be said to have on the final content. An AI that only reproduces third party excerpts might have a stronger argument for immunity.

45 This even seems to have been understood by the companies themselves, which can be inferred from their reaction to some lawsuits in the US[40]. In *Walters v. OpenAI*[41], Mark Walters, a radio talk show host, filed a defamation lawsuit against OpenAI after ChatGPT generated false information about him. The AI system erroneously described Walters as a defendant in a separate lawsuit and falsely accused him of fraud. Walters claims that these fabricated statements caused significant reputational damage, particularly affecting his career and audience. In *Battle v. Microsoft*[42], the plaintiff, Jeffery Battle, a U.S. Air Force veteran, filed a defamation lawsuit against Microsoft after discovering that searching his name on Bing resulted in a false description linking him to the "Portland Seven", a group of U.S. citizens who attempted to join the Taliban after 9/11. This mistake arose from the AI-assisted Bing search, which conflated his biography with that of a similarly named individual. Battle claims the false information caused significant reputational harm and seeks both monetary compensation and permanent removal of the erroneous data from Bing search results. What is interesting about these two cases is that defendants have not relied on the Section 230 defense so common in previous cases where social networks have been sued. It can be inferred that these companies may also have anticipated that immunity does not apply to the type of business they operate[43].

F. How to Limit the Proliferation of Harmful Content from General-Purpose GAI (GPGAI)?

46 The "easier" question has been clarified in the previous sections: GIA does not receive immunity based on European legislation and case law. However, we must still decide how to regulate these tools that have the potential to generate so much harmful content. To this end, the article now focuses on European law's responses to this issue. The goal must be to find solutions that both empower these new technologies and incentivize the implementation of reasonable security measures[44].

47 In the European Union (EU), known for its proactive and comprehensive approach to technology regulation, GPGAI[45] poses unique challenges that

40 To date, I am not aware of similar processes in Europe.

41 L.L.C., 1:23-cv-03122, (N.D. Ga.).

42 No. 1:2023cv01822 - Document 48 (D. Md. 2024)

43 Keep in mind that this is only an author's inference, since other factors may have also influenced this defense.

44 To some extent this may contribute to decreasing the presence of illegal content on social media platforms, i.e. as less illegal content comes out of these IAGs, less illegal content will be published on the platforms. These tools should be designed for legitimate uses, not as tools to achieve harmful outcomes. "*Lawmakers could directly ex ante regulate the AI's risk-creating behaviour. Namely, regulatory agencies could ex ante set detailed standards for the behaviour, employment, operation and functioning of any AI".* (Kovač, 2021, p. 109)

45 The previous sections dealt with generative artificial intelligence from a broader perspective, however, this section 5 focuses specifically on one type of generative artificial intelligence, namely general-purpose generative artificial intelligence (GPGAI), given the unique challenges it poses.

expose some loopholes. Therefore, this section discusses the main legal uncertainties in the European regulation of GPGAI, specifically how due diligence duties and transparency obligations are structured. This critical examination aims to contribute to the debate on the need for a more robust legal framework adapted to the realities of this emerging technology.

I. Loopholes in European Law regarding GPGAIs

1. AI Act

48 The EU AI Regulation (Artificial Intelligence Act), considered a pioneering global framework, seeks to establish risk categories for AI applications and set reasonable standards for their implementation. This represents a significant shift in policy by adopting a risk-based, preventive approach to regulate AI *ex-ante*, focusing on preventing harmful outcomes using safety principles. This "ecosystem of trust" aims to provide legal certainty for innovation while ensuring AI operators fulfill obligations proportionate to the risks their systems pose. The regulation seeks to prevent both material harm (e.g., threats to health, safety, or property) and immaterial harm[46], such as violations of fundamental rights (e.g., privacy, freedom of expression, dignity) and societal concerns like disinformation (de Graaf and Veldt, 2022, p. 804; Kretschmer et al., 2023, p. 3).

49 Four risk categories are distinguished:

- **Unacceptable risk**: implementation of AI in these areas will be forbidden. Examples include social scoring systems, facial recognition in public spaces, and manipulative AI. (Chapter 2, Article 5)
- **High-risk AI**: Most of the regulation focuses on this category. These systems are recognized in Annex III: Remote biometric identification systems, critical infrastructure, education, employment, access to and enjoyment of essential public and private services, law enforcement and Administration of justice[47].
- **Limited risk**: In theory this is the relevant category for General Purpose AI models and systems, requiring transparency in cases of, for example, chatbots or generation that may constitute deepfakes.
- **Minimal risk**: This level includes all other AI systems that do not fall under the above-mentioned categories.

50 The AI Act imposes specific obligations depending on the type of technology; therefore, it is crucial to accurately identify the technology in question to establish the applicable responsibilities. In this regard, to understand the obligations of a GPGAI, such as Meta's Imagine or Midjourney, it is necessary to differentiate between *general-purpose AI models*, *general-purpose AI systems* and *AI systems*. This distinction is essential, as different responsibilities and obligations apply to each of these categories, that in the end represent different stages and actors in the chain of operation of an AI.

51 ***General-purpose AI models*** (Article 3(63) and Recital 97) are characterized by their ability to perform a wide variety of tasks in a competent manner[48]. They can be distributed in various forms, such as libraries, APIs, direct downloads or hard copies, and are commonly modified or tweaked to create new models[49]. It is important to note that while these AI models are essential components of AI systems, they do not constitute systems per se. To become AI systems, they need additional elements, such as a user interface. In general, AI models are often integrated as part of larger AI systems. The EU IA Act classifies general purpose IA models (GPAIM) according to their level of risk: *systemic* or *non-systemic*. A GPAIM is considered systemic risk if it has 'high capabilities', that is, if it has considerable calculation capacity, which implies that it has required more than 10^25 floating point operations per second (FLOPS).

52 According to Article 3(1) an ***IA system*** is "*a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*". This is the most encompassing category for which the AI act is meant from the outset.

53 Meanwhile, a ***general-purpose AI system*** (Article 3(66) and Recital 85) "*means an AI system which is based on a*

46 As will be discussed later, this is an essential difference with the new Defective Products Directive.

47 Although these were the activities initially listed, the plan is for the list to be periodically reviewed and updated.

48 "The term "general purpose" is indicative of the models' abilities to be adapted to a variety of tasks outside of those for which they were specifically trained". (G'sell, 2024, p. 34)

49 In these cases, the question of accountability can be complicated by the fact that another actor is involved in the production chain, which may incorporate functionalities unforeseeable by the creator of the original model.

general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems". Recital 100 is more specific and indicates that "*when a general-purpose AI model is integrated into or forms part of an AI system, this system should be considered to be general-purpose AI system when, due to this integration, this system has the capability to serve a variety of purposes*"[50].

54 Using our case study to better illustrate these differences, the *system* and the *model* can be distinguished as follows:

- **LLama** is the underlying model on top of which Imagine runs (as ChatGPT runs on top of Open AI's GPT model). If we focus on the definition of the AI Act, LLama has the key characteristics of a GPAI model, namely generality. It can be tuned or adapted to specific applications, be it in the field of health, education, etc. On its own, LLama is not a system, as it lacks the necessary components to interact with users directly (such as user interfaces or specific input/output mechanisms). In other words, this is the general purpose model, designed for a wide range of tasks but not linked to a specific use case until it is integrated into a system.

- **Imagine** is the system that uses LLama as its core but incorporates other elements that make it a functional and specific tool: an interface that allows user interaction and additional components designed to handle the outputs (tuning and adjustments) to meet the specific objectives of the system and facilitate its practical use. In other words, it is the complete AI system, which, based on LLama, adds components needed to solve specific problems and provides an interface for users to interact with the underlying model.

55 Considering the above distinction, we shall now consider what obligations the AI Act establishes for such general-purpose GAI tools as Imagine, Midjourney or ChatGPT. In the case of Imagine and some other AIs, a single company concentrates control over both the model and the system. That is, Meta owns both the model (LLama) and the system (Imagine). This means that Meta will have to take into account the obligations that the regulation establishes for both tools[51].

56 As mentioned earlier, as far as the *system* is concerned, this type of AI seems to fit, in principle, in the Limited risk group, which means that the law establishes for them some specific obligations. According to Article 50:

- Providers should ensure that AI systems that interact directly with individuals inform them that they are dealing with an AI, unless this would be obvious to a reasonably informed and observant person in the context.

- Providers must mark the generated synthetic content (audio, image, video or text) in a machine-readable format and detectable as artificial[52]. Technical solutions must be effective, interoperable, robust and reliable, taking into account technical constraints, costs and recognized standards.

- Deployers of AI systems that generate or manipulate content (image, audio or video) as deepfakes must disclose that it is artificial[53]. If text is generated to report on matters of public interest, it must be disclosed as artificial, except if it is authorized by law to combat crime or if the content has been reviewed and is under human editorial responsibility.

57 Regarding the model, its obligations are recognized in Chapter 5. It first sets out the obligations of general-purpose AI models and later those that present systemic risk. Regarding the former, Article 53 states that providers must (a) maintain detailed technical documentation of the model, including its training, testing and results; (b) provide information and documentation to AI system providers that integrate the model, ensuring understanding of the model's capabilities and limitations; (c) implement a policy to comply with copyright laws, using state-of-the-art technologies to identify rights reservations in accordance with Directive (EU) 2019/790; (d) publish a detailed summary of the content used to train the model. As for the latter, article 55 states that they must follow the above obligations and in addition must (a) conduct assessments following standardized protocols and advanced tools, including documented adversarial testing to identify and mitigate systemic risks; (b) assess and mitigate systemic risks in the European Union, including their possible sources during the development, marketing or use of the model; (c) timely record, document and report relevant information on

50 "The majority of generative AI users do not engage directly with a generative AI model. Rather, they interact through an interface with a generative AI system that incorporates the model". (G'sell, 2024, p. 36)

51 Although this is the case in other scenarios such as Open IA with GPT or Google with Gemini, it does not always necessarily be so, as in other cases a company could only be in charge of the model, while another company is developing the system. In these cases, each will have to respect their respective set of obligations.

52 This would make it easier for social networks to detect them.

53 This would facilitate the moderation of AI-generated content in social networks.

serious incidents and corrective actions to the IA Office or to the authorities; (d) ensure an adequate level of cybersecurity protection for the model and its infrastructure.

58 It should be mentioned that these obligations became part of the Regulation very late in the process. This is because the GAI explosion happened at a stage when the general structure of the regulation was already relatively advanced. This situation explains why some aspects of the law may seem insufficient or not fully adapted to the techno-social realities of these systems[54]. One of the most problematic issues is when these general-purpose AIs end up being used in the context of activities that qualify as high risk. If the interface allows Imagine to be used in areas considered high-risk according to Annex III of the regulation, such as education or justice[55], would the whole system be classified as a high-risk AI system and be subject to the corresponding obligations? It is a plausible scenario, one that puts in tension the coherence of the whole norm with techno-social reality.

59 As Helberger and Diakopoulos (2023, p. 2) rightly note, generative AI systems, such as Imagine or Midjourney, have key differences from the traditional AI systems for which the AI Act was originally intended. These differences are their dynamic context and scale of use. It should be stressed that these generative AI systems are not designed for a specific purpose or context, but rather they are adaptable for later use in a wide range of fields, and their accessibility allows for a massive and diverse use, and is not aimed at a specific audience. This broad scope is partly a result of the massive scale of data used for training. These characteristics pose significant challenges to the AI Act's core approach, especially in terms of classifying these systems into high/low risk categories and the inherent unpredictability of future risk. Therefore, this classification criterion may not be the most appropriate for AI of general and widespread use.

60 According to the current logic of the AI Act, the classification of an AI system as unacceptable, high or minimal risk depends on the intended use of the system. Systems intended for areas specified in Annex III are considered high risk, while in other cases they are classified as minimal or no risk. However, in the case of general-purpose AI, it is the deployer who decides how to use the system, meaning that it is the deployer who ultimately determines whether the system falls into the high or low risk category[56] or even whether a use prohibited by Article 5 is made[57]. In most cases, therefore, the risks to society stem from the use of these systems by deployers. However, the personal scope of the regulation must be considered here. In the definition of deployer it states: "'*deployer means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity*"; which means that the obligations of this rule do not extend to ordinary users. The problem is that the regulation is ignoring the millions of users who use General-Purpose AI Systems, when these users are the ones who introduce the prompts to generate illegal content and on whom a certain share of responsibility should perhaps fall. It therefore seems that the regulation does not fully capture the particularities of GPGAI that have the capacity to generate a massive and unpredictable impact.

61 The unpredictability of future risks associated with GPGAIs also relates to their widespread use and to the versatility of these systems. According to the AI Act, providers of high-risk AI systems should be able to identify and analyze all potential risks associated with their use in areas such as health, security and fundamental rights. This includes anticipating all high-risk uses that may arise, including those that were not initially foreseen[58]. For each of these possible scenarios, suppliers would have to develop and implement mitigation strategies to reduce risks (Recitals 65, 114, and article 9). However, this is extremely costly and difficult to implement, as risk

54 "*While this scheme has worked relatively well for tangible products, the division of duties seems much more questionable in a world of (a) AI as a service which learns and changes, (b) 'AI as a service' or 'upstream' AI services, (c) general purpose AI and (d) AI as part of the services of a platform (the 'AI lifecycle').*" (Edwards, 2022, p. 5)

55 GPGAI can potentially be used to assess the learning outcomes of individuals (see recital 56) or to assist in the drafting of judicial documents (see recital 61).

56 This is a reality that the regulation seems to recognize in Recital 85: "*General-purpose AI systems may be used as high-risk AI systems by themselves or be components of other high-risk AI systems*" and Recital 84: "*To ensure legal certainty, it is necessary to clarify that, under certain specific conditions, any distributor, importer, deployer or other third-party should be considered to be a provider of a high-risk AI system and therefore assume all the relevant obligations. This would be the case if that party ... modifies the intended purpose of an AI system, including a general-purpose AI system, which has not been classified as high-risk and has already been placed on the market or put into service, in a way that the AI system becomes a high-risk AI system in accordance with this Regulation*".

57 The use of generative AI to create misleading or manipulative advertising or propaganda.

58 "AI systems can be general purpose, meaning the same system can be applied to different contexts and raise different impacts for different individuals and groups. For example, a developer of a facial recognition system could sell their product to authenticate entry to prisons or to surveil customers for targeted advertising. Holistically evaluating the risk of such a system in the abstract is an impossibility". (Edwards, 2022, p. 6)

assessments would have to be based on hypothetical scenarios and mitigation measures would depend on specific conditions of use, which have not yet occurred at the time of the assessment[59]. This is a shortcoming of the regulation, as *ex ante* risk assessments may not adequately capture all of the multiple real-life scenarios of use of a GPGAI. (Hacker et al., 2023, p. 1114).

2. Defective Products Directive (EU) 2024/2853

62 Product safety is not only achieved with *ex ante* and preventive standards. Ex post liability rules also operate as strong normative elements to moderate certain activities and to prevent damage. It could be said that *ex ante* rules such as the AI Act try to prevent the event of harm from occurring in the first place, but sometimes the harm does occur anyway, and someone has to compensate the victim. This is where tort law and its deterrence function intervene.

63 In the field of AI, it is the new Product Liability Directive (EU) 2024/2853 that regulates compensation for certain damage caused by AI systems[60]. The fundamental goal of this new directive is to harmonize the laws of the Member States of the European Union regarding the producer's liability for damage caused by defective products. This new directive repeals and replaces the previous Directive 85/374/EEC of 1985, thus updating the EU legal framework to adapt it to current technological and legal realities. This harmonization seeks to ensure better consumer protection in the European internal market. The directive addresses the challenges and opportunities posed by recent technological developments, especially in the following areas: Digital products, Artificial intelligence and Software (now explicitly considered as a product). Member States will have to transpose this new directive into national law by 9 December 2026 at the latest.

64 The first issue to highlight here, in the words of de Graaf and Veldt (2022, p. 823), is that "*product liability is, in short, limited to damage to property (other than the product itself) and damage resulting from death or personal injury caused by a defect in the product. In any case, liability for pure economic loss is left to national law*". Therefore, here we already face the first obstacle: the outputs of generative AI could be illegal, but the directive is unlikely to apply given the nature of the harm that such technology can generate (child abuse, defamation, intellectual property breaches, non-consensual sexual deepfakes or hate speech)[61]. This regulation is therefore of little relevance in assessing the damage caused by GPGAI. This has been criticized by Hacker (2024, p. 12) in the following way: "*instances of discrimination or violation of personality rights equally, and in some cases perhaps even more strongly, impact fundamental rights as the typical PLD scenarios of damage to property or health".*

65 In addition to the above, there is a CJEU ruling that may make it difficult to hold GAIs liable for the outputs they produce under the new defective product Directive. In the *Krone*[62] case, the Court determined that a newspaper containing erroneous health advice could not be deemed a defective product because the defect lay in the information, not the newspaper itself. Information, as a service, was considered out of the scope of the Directive. Applying this reasoning to generative AI outputs remains contentious[63]. Van Staalduinen (2024) argues that, unlike the newspaper in *Krone*, an AI system is not merely a carrier of information but its source. The outputs of AI are integral to their design and purpose. Unlike external advice published in a newspaper, the output of a GAI reflects its functionality and adaptability. This perspective hinges on the notion that the AI's design and operation inherently shape its outputs, linking any defects in those outputs to the product itself. Therefore, if a GAI produces erroneous or harmful outputs, it could be classified as defective under the Directive. This distinction highlights that not all information-providing products are the same; while the newspaper in *Krone* served as a medium for external information, GAI directly create the information they provide, warranting potential liability under the DPLD for resulting damages.

59 "... it is simply impossible to predict if, and if so, what the risks are that we can expect from unleashing extremely powerful AI models on society". (Helberger and Diakopoulos, 2023, p. 4)

60 This paper focuses on the Defective Products Directive, as the draft of the AI Liability Directive remains unofficial and lacks a definitive agreement on its provisions.

61 Directive' Recital 24: "*Types of damage other than those provided for in this Directive, such as pure economic loss, privacy infringements or discrimination, should not by themselves trigger liability under this Directive. However, this Directive should not affect the right to compensation for any damage, including non-material, under other liability regimes*".

62 ECJ, Krone, 1 June 2021, C-65/20, ECLI:EU:C:2021:471.

63 For arguments in favor of recognizing these AI tools as defective products see Spindler (2023) and Camacho Clavijo (2024). "*Contrary to what is established in the Case C-65/20 VI v KRONE, there is in our case a fundamental difference that may justify a different qualification in the context of product liability for software and printed information. In our case software does not simply convey information but constitutes an entity that can be used for the specific purpose for which it has been designed. The inaccurate medical assessment or prediction/information issued by the AI system constitutes an intrinsic element of the purpose-built system itself and is therefore implicit in its use and could therefore qualify as a defective product*" (Camacho Clavijo, 2024). For the argument against see (Borges, 2023, p. 39)

66 However, the application of the Directive to GAI remains contentious. While Van Staalduinen's interpretation offers a compelling argument[64], it must be acknowledged that this is still a matter of controversy and as such, some may still consider GAI to be outside the scope of the defective product Directive. Courts may need to address this regulatory gap by clarifying the scope of the DPLD regarding harmful information/outputs provided by GAI systems.

67 Even if GAI outputs are deemed outside the DPLD's scope, this does not mean such activities are exempt from liability. In these cases, national law, often based on principles of negligence, would likely govern. Here, the existence of a duty of care, a breach of that duty, and causation of harm would need to be established. For instance, if a developer fails to adequately train or monitor the AI system, and this negligence results in harm, liability could still arise under national frameworks, either civil, criminal or administrative.

II. Clarifying GPGAI Obligations.

68 While the AI Act and the new Defective Products Directive seek greater regulatory harmonization across the EU, the reality is that the current framework does not fully address the particularities and risks of GPGAI, limiting its effectiveness in both preventing and redressing harm. This legal exposure may have major implications for the future deployment of generative AI products and the public at large.

69 Thus, a more dynamic approach to continuous risk monitoring and mitigation is advocated here. In order to do so, however, amendments to the current AI act would have to be made. For example, the duties under Article 55[65], namely (i) to conduct adversarial testing to identify and mitigate systemic risks and (ii) to assess and mitigate risks during the development, marketing or use of the system, should apply to all GPGAI systems without having to subject them to high/low risk categorization. These duties are best performed not by the model developer[66], but by the system developer, who has a sounder awareness of how the system is being used by the users. It would be something like GPGAI with systemic risks or "general risk GAI".

70 Helberger and Diakopoulos (2023, p. 4) suggest drawing inspiration from Article 34 of the Digital Services Act (DSA)[67], already obliging very large *intermediaries* (online platforms and search engines) to regularly monitor the negative effects of their algorithmic systems on fundamental rights and social processes. This approach could be extended to providers of large-scale generative models, requiring them to assess and mitigate systemic risks on an ongoing basis. Hacker et al. (2023) go a step further and propose to amend the DSA to incorporate a fourth category: GPGAIs as *content providers*[68]. This is due to what has already been discussed: the DSA's scope of application is restricted to intermediaries, it does not apply to content providers *per se*. However, this does not prevent this norm could be expanded in the future with some amendments to regulate also some aspects of the GPGAIs[69]. Certainly, this technology could also benefit from the implementation of some established solutions developed in the DSA, including notice and action, trusted flaggers systems or compulsory dispute resolution[70], as none of these

64 Van Staalduinen (2024) also argues that if software, alarm systems, smart watches or other measurement devices are recognized as products by the DPLD impact assessment, and after all, their function is essentially to provide information, there is no reason to exclude AI which ultimately performs a similar function.

65 Obligations of *providers* of general-purpose AI *models* with systemic risk.

66 The duties of the model developer should actually focus on ensuring the quality and security of the training data and knowing as much as possible about the capabilities and limitations of the model, which should be communicated to the system developer.

67 These generative models are special because they produce content that can support human communication, which raises new challenges and questions about how to regulate the use of AI to ensure that this communication is ethical and responsible. The DSA regulates digital spaces where much of this human communication happens, establishing a framework to make it safer, more transparent and more respectful of fundamental rights. Although the DSA is designed to apply only to intermediaries, as argued at the beginning of this article, its general objectives and purposes are so broad that, in principle, it could cover a wide variety of electronic services.

68 "*For example, to extend the DSA to LGAIMs in specific ways, one would have to update the DSA or include a reference in the AI Act. Both modifications require concurring decisions by the EP and the Council (Art. 289 TFEU)*" (Hacker et al., 2023, p. 1120). As can be noted, these authors only consider it possible to find a solution through the modification of one of the two norms. This seems to be a sound solution compared to writing a completely new standard.

69 This would require a number of modifications to the DSA, in particular to nuance the active/passive division, which is not representative of GPGAIs. A different activity criterion should be introduced for GPGAIs, one that measures rather the degree of involvement, influence or contribution to the content, based on the differences explained between extractive and abstractive models.

70 "*While the notice and action mechanism applies to all hosting services, instruments like trusted flaggers, obligatory dispute resolution, and risk management systems are reserved for the*

are recognized as such in the AI Act. It would seem a good idea to require developers and operators of GPGAIs to adopt notice-and-action mechanisms, prioritizing alerts sent by trusted moderators[71]. They could then adjust systems to block problematic prompts and avoid loopholes exploited by malicious actors. Therefore, this article argues that effective regulation of the GPGAI necessarily requires the amendment of at least one of the two regulations currently in force: the DSA or the IA Act (Hacker et al., 2023, p. 1120). A revision is required to establish a bridge between these two regulations, avoiding a vacuum that would be filled by opaque self-regulatory measures. While it is certainly plausible to adopt expansive interpretive approaches (Botero Arcila, 2023; Stalla-Bourdillon, 2023), such a strategy would provide only partial solutions and does not comprehensively address the existing regulatory gap. Legal certainty for market operators and users must be ensured through the development of clear, tailored, and predictable legal obligations.

71 In summary, to limit the massive generation of illegal content, the norm should at a minimum have the following obligations in place for model providers and system deployers, regardless of the level of risk involved:

- Obligations relating to the developer of the general-purpose GAI *model*: safeguards related to the training of the model, i.e.: curation of data used to train the model (inspired by Article 10), collaboration with potential deployers to create operational synergies that improve the security of the system (Article 11), conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks (Article 55(1)(a)), ensuring an adequate level of protection for cybersecurity and for the physical infrastructure (Article 55(1)(d))

- Obligations relating to the deployers of general purpose GAI *systems*: as this is the system that is implemented around the model and is the one with which the public interact directly, the duties are more focused on fine-tuning, adding security layers and functionalities, for example: collaborating with the model developers to create operational synergies to improve the security of the system, conducting and documenting adversarial testing of the model in order to identify and mitigate systemic risks, implementing external systems that analyze inputs and outputs to identify and block problematic content, implementing a DSA-inspired notice and action system, involving trusted flaggers.

72 Some companies are already incorporating some of these security mechanisms on a voluntary basis. In a recent paper, Chi et al. (2024) present Llama Guard 3 Vision, a model that improves the safety of multimodal AI conversations by addressing harmful content in both inputs (prompts) and outputs (responses) involving images. Unlike previous versions of Llama Guard, which only analyzed text, this version is specifically designed to support image reasoning use cases. Llama Guard 3 Vision can therefore analyse images in conjunction with text to identify harmful content that previous versions of Llama Guard might miss. For example, it can detect an inappropriate request based on the image provided, even if the text itself is not problematic. The model is trained to predict safety labels based on the 13 risk categories of the MLCommons taxonomy. These categories include violent crimes, sexual content, hate speech, election misinformation, and more. While Llama Guard 3 Vision offers an additional layer of protection, it is not immune to adversarial attacks. It is important to be aware of its limitations and to continue to explore ways to improve its robustness against malicious users.

G. Conclusion

73 The technical differences between generative AI (GAI) and content curation AI (CAI) are central to defining their potential liability regarding the content they create or organize. Based on the European law and case law approach, this differences are key to deciding whether they can benefit from the safe harbour doctrine. The European Union (EU), through the Digital Services Regulation (DSA), focuses on the

narrower group of "online platforms"". (Hacker et al., 2023, p. 1118)

71 Developers of generative AI systems usually implement a variety of measures to prevent their product from being used for malicious purposes. These measures consist of stress-testing the system in an attempt to identify potential vulnerabilities in advance. This approach is known as "red-teaming" (Ahmad et al., 2024). This generally consists of asking a network of external human testers to try to bypass security safeguards in an attempt to identify vulnerabilities. This is one way to understand how users could potentially interact with the system. However, no amount of testing can completely rule out unwanted or harmful behavior due to the complexity of language models and the ways in which users interact with them. That is why this article advocates borrowing approaches from the DSA as a way to improve the system on a more continuous basis and with the involvement of more stakeholders beyond "red-team". It suggests the implementation of a "notice and action" system, allowing continuous monitoring based on real-time detection and active response to incidents. This approach overcomes the limitations of relying solely on ad hoc tests such as those carried out by the "red-teams".

concept of active role to determine the liability of online intermediaries. An entity has an active role when it has a direct involvement (control) in the creation of content. As discussed, the *abstractive* GAI has a more active role, since it produces new information, something that goes beyond curation or mere recommendation. Therefore, the *abstractive* GIA loses the benefit of safe harbour in the EU, as its crucial contribution to content creation makes it to some extent responsible for the outcomes.

74 This raises the need for specific rules to define the appropriate scope of accountability expected for these technologies. However, current EU legislation presents loopholes in the regulation of GPGAI. With its massive usage capacity, versatility and unpredictability of potential risks, GPGAI challenges the current risk-based approach of the AI Act. The difficulty of anticipating and mitigating all risks, together with the exclusion of ordinary users from legal obligations, limits the effectiveness of the law in preventing and redressing harm. It could be important to draw inspiration from Article 34 of the Digital Services Act (DSA), which obliges large search engines and platforms to regularly monitor the negative effects of their algorithmic systems on fundamental rights. It is also particularly important to define the responsibilities of the actors involved in the supply chain: essentially those who design and train the models and those who put the system into operation for the public. Specifically, the latter, due to their closer understanding of user actions, should focus on fine-tuning and implementing functionalities like adversarial testing, input/output monitoring, and notice-and-action systems to mitigate risks and ensure safety. The implementation of safeguards in GPGAI should be continuous, based on real-time detection and active response to incidents, rather than depending solely on adversarial *ex-ante* testing. However, in order to materialize these obligations, it is necessary to bridge the gap between the DSA and the IA act, since it seems that neither of them succeed in capturing the real essence of the GPGAIs.

75 To conclude, it is worth remembering that any legislative strategy in the field of AI must acknowledge the global nature of this market and the intense competition for technological development but also restate the commitment to the protection of democratic values and fundamental rights. Any successful European regulatory reform cannot afford to ignore the strategies adopted by other jurisdictions, especially the United States and China. However, neither can it uncritically replicate their approaches, which often prioritize innovation at the expense of safety, transparency and accountability. Instead, Europe must go its own way, it must aspire to become a global benchmark in AI development and deployment, not only for its technological capabilities, but also for its commitment to ethics and security. Ultimately, the European approach must be based on the conviction that technological development and the protection of fundamental rights are not incompatible, but complementary. We are confident that in the long term this is the right way forward for sustainable leadership in the age of artificial intelligence.

H. REFERENCES

Ahmad, L., Agarwal, S., Lampe, M., Mishkin, P., 2024. OpenAI's Approach to External Red Teaming for AI Models and Systems. Technical report, OpenAI, November 2024. URL: https://cdn.openai.com/papers/openais-approach-to-external-red-teaming.pdf

Angelopoulos, C., 2017. "On Online Platforms and the Commission's New Proposal for a Directive on Copyright in the Digital Single Market". Available at SSRN 2947800.

Arroyo Amayuelas, E., 2020. "La responsabilidad de los intermediarios en internet: ¿Puertos seguros a prueba de futuro?" Cuadernos de Derecho Transnacional, 2020, num. 1, p. 808-837.

Balkin, J.M., 2021. "How to regulate (and not regulate) social media". J. Free Speech L. 1, 71.

Bambauer, D.E., Surdeanu, M., 2023. "Authorbots." J. Free Speech L. 3, 375.

Belcic, I., Stryker, C., 2024. "What is GPT (generative pre-trained transformer)?" | IBM [WWW Document]. URL https://www.ibm.com/think/topics/gpt (accessed 11.25.24).

Borges, G., 2023. "Liability for AI Systems Under Current and Future Law: An overview of the key changes envisioned by the proposal of an EU-directive on liability for AI". Computer Law Review International 24, 1–8.

Botero Arcila, B., 2023. "Is It a Platform? Is It a Search Engine? It's ChatGPT! The European Liability Regime for Large Language Models Symposium: Artificial Intelligence and Speech." J. Free Speech L. 3, 455–488.

Camacho Clavijo, S., 2024. "AI assessment tools for decision-making on telemedicine: liability in case of mistakes". Discover Artificial Intelligence 4, 24.

Candeub, A., 2020. "Bargaining for Free Speech: Common Carriage, Network Neutrality, and Section 230". Yale JL & Tech. 22, 391.

Chander, A., Krishnamurthy, V., 2018. "The myth of platform neutrality". Geo. L. Tech. Rev. 2, 400.

Chi, J., Karn, U., Zhan, H., Smith, E., Rando, J., Zhang, Y., Plawiak, K., Coudert, Z.D., Upasani, K., Pasupuleti, M., 2024. "Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations". arXiv preprint arXiv:2411.10414.

Cohen-Almagor, R., 2010. "Responsibility of and Trust in ISPs". Knowledge, Technology & Policy 23, 381–397.

de Graaf, T., Veldt, G., 2022. "The AI Act and Its Impact on Product Safety, Contracts and Liability". European Review of Private Law 30.

Edwards, L., 2022. "Regulating AI in Europe: four problems and four solutions". Ada Lovelace Institute 15, 2022.

Elkin-Koren, N., De Gregorio, G., Perel, M., 2021. "Social Media as Contractual Networks: A Bottom Up Check on Content Moderation". Iowa L. Rev. 107, 987.

Gillespie, T., 2018. "Platforms are not intermediaries". Geo. L. Tech. Rev. 2, 198.

Grimmelmann, J., 2015. "The virtues of moderation". Yale JL & Tech. 17, 42.

G'sell, F., 2024. "Regulating under Uncertainty: Governance Options for Generative AI". Stanford Cyber Policy Center. Freeman Spogli Institute. Stanford Law School.

Hacker, P., 2024. "Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence: Complementary impact assessment". EPRS | European Parliamentary Research Service.

Hacker, P., Engel, A., Mauer, M., 2023. "Regulating ChatGPT and other large generative AI models", in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1112–1123.

Helberger, N., Diakopoulos, N., 2023. "ChatGPT and the AI Act". Internet Policy Review 12.

Henderson, P., Hashimoto, T., Lemley, M., 2023. "Where's the Liability in harmful AI Speech?" J. Free Speech L. 3, 589.

Husovec, M., 2023. "Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules". Berkeley Technology Law Journal 38.

Keller, D., 2023a. "What the Supreme Court Says Platforms Do". Lawfare. URL https://www.lawfaremedia.org/article/what-the-supreme-court-says-platforms-do (accessed 11.30.23).

Keller, D., 2023b. "Carriage and Removal Requirements for Internet Platforms: What Taamneh Tells Us". Journal of Free Speech Law 4, 87–138.

Khosrowi, D., Finn, F., Clark, E., 2024. "Engaging the many-hands problem of generative-AI outputs: a framework for attributing credit". AI and Ethics 1–19.

Kosseff, J., 2022. "A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not)". Berkeley Technology Law Journal 37.

Kovač, M., 2021. "Autonomous Artificial Intelligence and Uncontemplated Hazards: Towards the Optimal Regulatory Framework". European Journal of Risk Regulation 13, 94–113.

Kretschmer, M., Kretschmer, T., Peukert, A., Peukert, C., 2023. "The risks of risk-based AI regulation: taking liability seriously". arXiv preprint arXiv:2311.14684.

Land, M.K., 2019. "Regulating Private Harms Online: Content Regulation under Human Rights Law", in: Human Rights in the Age of Platforms. Cambridge, MA : The MIT Press, p. 285.

Leerssen, P., 2020. "The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems". European Journal of Law and Technology 11.

Llansó, E., van Hoboken, J., Leerssen, P., Harambam, J., 2020. "Artificial Intelligence, Content Moderation, and Freedom of Expression". Transatlantic High Level Working Group, Working Group on Content Moderation Online and Freedom of Expression. Annenberg Public Policy Center.

Miers, J., 2023. "Yes, Section 230 Should Protect ChatGPT And Other Generative AI Tools". Techdirt. URL https://www.techdirt.com/2023/03/17/yes-section-230-should-protect-chatgpt-and-others-generative-ai-tools/ (accessed 11.20.24).

Mirmira, S., 2000. "Lunney v. Prodigy Services Co". Berk. Tech. LJ 15, 437.

Noto La Diega, G., Bezerra, L.C., 2024. "Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive". International Journal of Law and Information Technology 32, eaae021.

OECD, 2024. Explanatory memorandum on the updated OECD definition of an AI system (No. 8),

OECD Artificial Intelligence Papers. OECD Publishing, Paris.

OECD, 2022. OECD Framework for the Classification of AI systems (No. 323), OECD Digital Economy Papers. OECD Publishing, Paris.

Pagallo, U., 2011. "ISPs & Rowdy web sites before the law: Should we change today's safe harbour clauses?" Philosophy & Technology 24, 419–436.

Patel, S.K., 2002. "Immunizing Internet Service Providers from third-party Internet defamation claims: How far should courts go". Vand. L. Rev. 55, 647.

Perault, M., 2023. "Section 230 Won't Protect ChatGPT." J. Free Speech L. 3, 363.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019." Language models are unsupervised multitask learners". OpenAI blog 1, 9.

Sartor, G., 2017. "Providers Liability: From the eCommerce Directive to the future", IP/A/IMCO/2017-07. ed. European Parliament's Committee on the Internal Market and Consumer Protection.

Spindler, G., 2023. "Different approaches for liability of Artificial Intelligence–Pros and Cons", in: Liability for AI. Nomos Verlagsgesellschaft mbH & Co. KG, pp. 41–96.

Stalla-Bourdillon, S., 2023. "What if ChatGPT was much more than a chatbox? What if LLM-as-a-service was a search engine?" Peep Beep! URL https://peepbeep.blog/2023/04/03/what-if-chatgpt-was-much-more-than-a-chatbox-what-if-llm-as-a-service-was-a-search-engine/ (accessed 3.31.25).

Sylvain, O., 2021. "Platform Realism, Informational Inequality, and Section 230 Reform". Yale LJF 131, 475.

Thorburn, L., 2022. "How Platform Recommenders Work. Understanding Recommenders". URL https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a (accessed 12.1.23).

Valcke, P., Kuczerawy, A., Ombelet, P.-J., 2017. "Did the Romans get it right? What Delfi, Google, eBay, and UPC TeleKabel Wien have in common", in: The Responsibilities of Online Service Providers. Springer, pp. 101–116.

Van Eecke, P., 2011. "Online service providers and liability: A plea for a balanced approach". Common Market Law Review 48.

Van Hoboken, J., Pedro Quintais, J., Poort, J., Van Eijk, N., 2018. Hosting Intermediary Services and Illegal Content Online: An analysis of the scope of article 14 ECD in light of developments in the online service landscape, A study prepared for the European Commission DG Communications Networks, Content & Technology. URL: https://www.ivir.nl/publicaties/download/hosting_intermediary_services.pdf

van Staalduinen, J.H., 2024. "European Product Liability for AI-based Clinical Decision Support Systems", in: Digital Governance: Confronting the Challenges Posed by Artificial Intelligence. Springer, pp. 15–40.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. "Attention Is All You Need". Presented at the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA. https://doi.org/10.48550/arXiv.1706.03762

Volokh, E., 2023. "Large libel models? liability for AI output". J. Free Speech L. 3, 489.

Volokh, E., 2021. "Treating social media platforms like common carriers?" J. Free Speech L. 1, 377.

York, J.C., Zuckerman, E., 2019. Moderating the public sphere, in: Human Rights in the Age of Platforms. Cambridge, MA: The MIT Press, p. 137.

Ziniti, C., 2008. "Optimal liability system for online service providers: How Zeran v. America online got it right and web 2.0 proves it". Berkeley Tech. LJ 23, 583.

Zurth, P., 2020. "The German NetzDG as role model or cautionary tale? Implications for the debate on social media liability". Fordham Intell. Prop. Media & Ent. LJ 31, 1084.