Implications of Regulations on Large Generative AI Models in the *Super-Election Year* and the Impact on Disinformation

Vera Schmitt[1,5], Aljoscha Burchardt[5], Jakob Tesch[2], Eva Lopez[3], Salar Mohtaj[1,5]
Konstanze Neumann[4], Tim Polzehl[5] and Sebastian Möller[1,5]
Technische Universität Berlin[1], Ubermetrics GmbH[2], Deutsche Welle[3], delphai GmbH[4];
German Research Center for Artificial Intelligence[5]
{vera.schmitt, sebastian.moeller}@tu-berlin.de, jakob.tesch@ubermetrics-technologies.com,
eva.lopez@dw.com, {aljoscha.burchardt, salar.mohtaj, tim.polzehl}@dfki.de, konstanze@delphai.com

Abstract

With the rise of Large Generative AI Models (LGAIMs), disinformation online has become more concerning than ever before. Within the *super-election year* 2024, the influence of mis- and disinformation can severely influence public opinion. To combat the increasing amount of disinformation online, humans need to be supported by AI-based tools to increase the effectiveness of detecting false content. This paper examines the critical intersection of the AI Act with the deployment of LGAIMs for disinformation detection and the implications from research, deployer, and the user's perspective. The utilization of LGAIMs for disinformation detection falls under the high-risk category defined in the AI Act, leading to several obligations that need to be followed after the enforcement of the AI Act. Among others, the obligations include risk management, transparency, and human oversight which pose the challenge of finding adequate technical interpretations. Furthermore, the paper articulates the necessity for clear guidelines and standards that enable the effective, ethical, and legally compliant use of AI. The paper contributes to the discourse on balancing technological advancement with ethical and legal imperatives, advocating for a collaborative approach to utilizing LGAIMs in safeguarding information integrity and fostering trust in digital ecosystems.

Keywords: AI Act, DSA, LGAIMs, disinformation

1. Introduction

The World Economic Forum's *Global Risks Report 2024* (Forum, 2024) identifies Artificial Intelligence (AI)-generated disinformation as the second most critical risk, potentially causing significant global crises. In the context of 2024, a year witnessing over 70 elections globally, including major elections such as the U.S. presidential election, India's general elections, and the European Parliament elections, there is increasing concern about the profound influence that AI-generated content may have (Iskandar et al., 2023). In recent years, the field of generative AI has seen impressive advancements. Models such as *ChatGPT-4* (OpenAI, 2023) for text generation, *DALL·E 3* (Nguyen et al., 2024), and Sora (Brooks et al., 2024) for visual content creation, and *Whisper* (Radford et al., 2022) for voice cloning have undergone significant improvements. The AI models discussed in this paper are commonly known as *Foundation Models*, *Large Language Models* (LLMs), or *Large Generative AI Models* (LGAIMs) (Hoffmann et al., 2022) — the terminology we have chosen to use here. LGAIMs have advanced to a stage where they are user-friendly and do not demand deep technical know-how for their utilization. LGAIMs have the potential to liberate professionals to concentrate on important tasks, like direct patient care, potentially leading to a more efficient and fairer distribution of resources (Hacker et al., 2023). As a result, AI is becoming more embedded in our everyday experiences, significantly influencing the transformation of the digital environment, especially with its role in the creation of disinformation. The capacity to generate disinformation, create deepfakes, and disseminate hate speech has greatly escalated, posing serious threats to the integrity of information ecosystems (Hacker et al., 2023; Simon et al., 2023; Longoni et al., 2022; Khamsehashari et al., 2023). The *infodemic* experienced during the Covid-19 pandemic (Balakrishnan et al., 2022) and the military conflicts in Ukraine and Israel (Darwish et al., 2023) exemplify the significant influence that LGAIMs can wield in producing disinformation and shaping public opinion (Monsees, 2023; Satariano and Mozur, 2023). In light of these developments, the impending AI Act, aimed at regulating the use and deployment of AI technologies, assumes critical importance. In the disinformation context, LGAIMs can be used for both generating and detecting mis- and disinformation. The AI Act offers a framework to assess the risk of AI systems and defines obligations depending on the risk category in which the AI system falls. However, the risk assessment and the associated obligations are sometimes not straightforward. Especially when transferring the often generally formulated obligations to concrete technical implementations, much freedom is given concerning the concrete design scope and technical interpretation

LEGAL2024 @LREC-COLING-2024, pages 28–38
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

of the obligations defined. From the researcher's perspective, the AI Act presents both opportunities and constraints. It offers a structured environment for ethical AI research, emphasizing the need for responsible innovation and the importance of addressing AI's societal impacts. The AI Act offers so-called *Sandboxes*, allowing the fostering of research and innovation. From the deployer's perspective, the Act is a double-edged sword. On the one hand, it offers a much-needed framework for ethical and transparent AI deployment, ensuring that AI technologies are used responsibly and transparently. Deployers must navigate the complex landscape of compliance, grappling with the challenges of integrating ethical considerations into their AI systems without hindering innovation. On the other hand, the Act represents a significant compliance challenge, with stringent regulations potentially hindering the pace of AI development and deployment. From the user's perspective, the AI Act is beneficial in ensuring user rights and safety in the digital age. It promises to safeguard users from the risks associated with AI-generated disinformation, deepfakes, and other forms of digital manipulation. By setting clear standards for transparency, accountability, and reliability, the Act aims to foster trust in AI technologies, enabling users to benefit from AI advancements while being protected from their potential harms. However, the effectiveness of the AI Act depends heavily on its effectiveness in enforcement. When considering the shortcomings of the General Data Protection Directive (GDPR), the enforcement was and is still the major hurdle (see (Schmitt et al., 2023)), where there is no control of GDPR compliance on a technical level as long as there is no complaint from a user. If the enforcement of the AI Act repeats similar mistakes, it will remain ineffective and offer insufficient protection to users. Moreover, the roles of *deployers* and *providers* specified within the AI Act and carrying specific responsibilities have raised discussions and concerns. The definition and responsibilities are vaguely defined, which leaves room for interpretation and may lead to differences on a national level when enforcing the AI Act. Overall, the AI Act and its implications are multifaceted, emphasizing the importance of a balanced approach to AI regulation that considers the perspectives of all stakeholders involved. Thus, within this research, the implications of the AI Act for the use case of mis-and disinformation detection are analyzed from different perspectives: (1) research, (2) provider, (3) deployer, and (4) user perspective. This contribution aims to facilitate the understanding of the AI Act's implications for the stakeholders involved.

2. Background

Different regulations and obligations must be considered when using LGAIMs in the EU. The European Council and Parliament have reached a provisional consensus on the proposed AI Act, establishing uniform regulations for artificial intelligence. This includes Article 13, known as "Transparency and Provision of Information to Users," within the EU AI Act[1]. In this context, the requirements for adequate transparency of AI systems are specified, ensuring that both providers and users can reasonably comprehend the functioning and recommendations of the AI system. Therefore, adherence to transparency obligations is mandatory when utilizing AI systems for disinformation detection within the EU. Furthermore, the voluntary Code of Practice on Disinformation[2] has been crafted collaboratively by various stakeholders from industry, legal, and research sectors to establish a unified approach for addressing disinformation online on an international scale. The AI Act, along with the Code of Practice on Disinformation, mandates transparent and detailed system architecture for AI applications tasked with disinformation detection. The considerable data demands for developing LGAIMs typically mean that creators must depend on publicly accessible internet data for training, a source that is rarely ideal in terms of data quality (Luccioni and Viviano, 2021). Consequently, the output produced by these models can be biased, discriminatory, or detrimental (Nadeem et al., 2020). To prevent or at least lessen this problem, model developers should employ appropriate curation methods (Bai et al., 2022). Although the absence of transparency from most LGAIMs makes it impossible to confirm assertions about handling harmful content, it appears that most LGAIMs depended, or still depend, on human intervention to train an automated content moderation system, aiming to inhibit the generation of abusive content (Frey and Osborne, 2023; Helberger and Diakopoulos, 2023a). However, even if the detection of abusive content were automated and flawless, it would only address part of the issue. The persistent risk is the generation of disinformation, which can be challenging to identify (Goldstein et al., 2023). Nevertheless, LGAIMs can not only be utilized to generate harmful and potentially fake content but also to detect disinformation. Several endeavors are made to fight mis-and disinformation by developing advanced AI models for facilitating its detection. Within the media landscape, AI is progressively employed to perform content verification tasks to detect disinforma-

[1] *Laying down Harmonised Rules on Artificial Intelligence* (AI Act), 15.01.2024.

[2] *Strengthened Code of Practice on Disinformation*, 15.01.2024.

tion. Several research and development projects, such as AI4Media, vera.ai, and news-polygraph face similar questions in the interdisciplinary consortium including partners from research, industry and media what implications existing regulations enforce on the outcome of the projects. Hereby, the question arises of how LGAIMs can be used for this specific use case by complying with the new obligations outlined in the AI Act and Digital Services Act (DSA) in using AI systems for combating disinformation. Given the considerable challenges AI-generated disinformation poses, the legal landscape is evolving to address these complex issues. As AI technologies become increasingly capable of generating persuasive and realistic disinformation, the need for a robust legal framework to mitigate the risks and protect public discourse becomes paramount. Similarly, for the use of LGAIMs, the legal regulations become more pronounced and need to be considered when using AI-based tools for dis- and misinformation detection.

2.1. AI Act

Before delving into the legal implications, an introduction to the AI Act and its foundational concepts is provided. The EU is actively pursuing a broad regulatory effort, the AI Act, designed to create a thorough regulatory framework for AI governance. The European Parliament has endorsed new regulations that focus on enhancing transparency and risk management in creating AI systems across the EU, prioritizing a human-centric and ethical approach. The AI Act encompasses AI applications within both the public and private sectors, targeting systems either sold in the EU market or impacting EU citizens. Its central aim is to provide AI developers, deployers, and users with detailed guidance by outlining requirements and obligations for various AI system applications. Hereby, the adoption of a risk-based approach has been driven by thorough consultations with essential stakeholders, notably the High-Level Expert Group on AI. The risk-based approach balances recognizing AI's inherent benefits and potentials against acknowledging possible dangers and risks from novel AI applications and systems. The regulation adopts an inclusive definition of AI in *Article 3*, covering general AI systems influencing decision-making and opinions by providing content, predictions, recommendations, or decisions. This definition covers a variety of methodologies, including machine learning techniques (such as supervised, unsupervised, reinforcement, and deep learning), logic- and knowledge-based approaches (including inductive logic programming, knowledge representation, and deductive engines), as well as statistical methods like Bayesian estimation and search optimization. Within the framework of the AI Act, a risk-based classification outlines four distinct categories of risks concerning AI systems, with particular emphasis on delineating between *high-risk* and *limited risk* categories. (1) **Unacceptable risk**: This category includes AI systems that pose clear threats to the safety, fundamental rights, and well-being of individuals. Examples encompass state-run social scoring mechanisms and unsafe voice-activated toys explicitly banned from the European market. (2) **High risk:** AI systems necessary to sectors important to human health and safety, such as infrastructure, education, safety components, law enforcement, and public administration, are classified here. Compliance with stringent requirements, as specified in *Chapters 2 and 3* of the AI Act (eu, 2021), is mandatory before these systems can be introduced to the EU market. These requirements cover using high-quality data sets, risk management systems, transparency, accuracy, security and robustness measures, user guidance, human oversight, and conformity evaluations. (3) **Limited risk:** This classification applies to AI applications that necessitate transparency to ensure user interactions with AI are intelligible. It primarily mandates that users be adequately informed when they are interacting with AI systems or AI-generated content, including audio and video manipulations (e.g., deepfakes). (4) **Minimal risk:** AI systems that are supposed to pose a minor risk to humans, such as those used in video games, email spam filters, and certain consumer applications, fall under this category. For these, the Act defines no additional specific regulatory obligations. In light of technological advancements, regulatory bodies have incorporated a provision mandating the continuous evaluation of AI systems' risk classifications. The EU is instructed to consider the "intended purpose of the AI system" during the risk classification process of AI technologies (eu, 2021). This provision underscores the critical issue stemming from the potential of AI systems to bypass or dodge the Act's protective measures. This problem is attributed to the complex interplay among the developers and deployers providing AI systems and the distinct purpose(s) these systems are designed to fulfill (Gutierrez et al., 2022).

2.2. DSA

When discussing regulatory frameworks concerning mis-and disinformation detection, it is also important to consider the DSA. Like almost all new technologies, generative models can be employed for positive uses (such as creating birthday cards) or negative ones (such as starting a *shitstorm* on social media platforms) (Brundage et al., 2018). Specifically, the developers of ChatGPT foresaw the possibility of misuse and trained an in-house AI moderator to detect harmful content, albeit with contentious assistance from contractors in Kenya

(Perrigo, 2023). Nonetheless, individuals determined to use ChatGPT and similar LGAIMs, such as Mixtral and Llama 2, to create deceptive or harmful content will discover methods to elicit such responses. Prompt engineering is evolving into a sophisticated technique for extracting any type of content from LGAIMs, and detecting disinformation becomes more and more challenging despite ongoing industry initiatives to enhance the transparency of models and sources (Deiseroth et al., 2023). In response to the rising challenge of fake news and hateful content, the EU has recently implemented the DSA. However, when the DSA was crafted, LGAIMs were not the center of public discourse. Therefore, the DSA aimed to address illegal content on social networks, which was predominantly generated by human users or the occasional automated X (Twitter) account, rather than tackling the challenges posed by LGAIMs. The DSA appears to be outdated as soon as it was implemented due to two significant limitations in its scope. Firstly, it is applicable only to what is termed intermediary services (as per Articles 2(1) and (2) of the DSA). Article 3(g) of the DSA categorizes these as "mere conduits" (like Internet service providers), "caching," or "hosting" services (such as social media platforms, also referred to in Recital 28 of the DSA). However, it is arguable that LGAIMs do not fit into any of these categories. They differ distinctly from mere conduit or caching services that facilitate internet connections. On the other hand, hosting services are described as entities that store information provided by and at the request of a user (Article 3(g)(iii) DSA). In contrast to traditional social media setups, in the context of LGAIMs, it is the AI model, not the user, that generates the content (Hacker et al., 2023). Therefore, the scope of the DSA mechanisms remains applicable only to the sharing of content generated by LGAIMs on conventional social networks. Mis- and disinformation can also be disseminated effectively and broadly through direct personal communication. Despite the EU legislator's decision to leave closed groups outside the DSA's ambit, this decision necessitates reconsideration in light of the accessibility of LGAIM-generated outputs, which amplify the associated risks. Even the strictest enforcement of DSA regulations, possibly in conjunction with the General Data Protection Regulation (GDPR) mandates for data deletion (Articles 17(2) and 19 GDPR), is insufficient to reverse the damage or often prevent the ongoing spread of problematic content. Despite commendable attempts through the DSA to tackle the spread of disinformation and hate speech, the current EU legislation is inadequate in fully addressing the negative implications of LGAIMs. Thus, a selective expansion of the DSA to LGAIMs is necessary to make them useful for disinformation detection.

3. Risk Assessment of Disinformation Detection

As LGAIMs become more advanced and are also applied for disinformation detection, the risk categorization of LGAIMs needs to be clarified. Given the sensitive nature of disinformation, which frequently entails determinations regarding the flagging, removal, or blocking of information, there exists a potential for infringement upon freedom of expression. Consequently, the deployment and subsequent actions derived from AI systems' classification or prediction outcomes can be classified under the *high-risk* or *limited risk* category, depending on the concrete usage scenario. First, within the AI Act, LGAIMs are defined as General-Purpose AI Systems (GPAIS) designed by the provider to execute universally applicable tasks such as image and speech recognition, generating audio and video, detecting patterns, answering questions, translating, among others; a general-purpose AI system is capable of being utilized across multiple contexts and incorporated into various other AI systems (Art. 3(1b) AI Act). The late inclusion of LGAIMs in the AI Act was a key point of the debate for the final version of the AI Act and was mostly motivated by the emergence and wide adoption of ChatGPT (Hacker et al., 2023). Conceptually, the term *generality* might pertain to various aspects such as their capabilities (like language processing versus visual comprehension or their integration in multimodal models), the range of application areas (such as educational or economic domains), the wide array of tasks they can perform (like summarization versus text completion), or the flexibility in the types of outputs they can generate (such as producing images in black and white or in full color) (Gutierrez et al., 2022). General Purpose AI Systems (GPAIS) fall under high-risk obligations (such as Articles 8 to 15 of the AI Act) if they can be employed as high-risk systems or as parts of such systems (as per Article 4b(1)(1) and 4b(2) of the AI Act). Thus, unless it can be technically guaranteed that misuse is prevented, LGAIMs will generally be classified as high-risk systems under the suggested regulation. Second, even if we would not use GPAIS for disinformation detection, one can easily argue that these systems, through content moderation, impact fundamental rights, in particular freedom of expression and information. As defined in Recital 28a of the AI Act, this is a strong argument for classifying them as high-risk. Third, Annex III lists application areas where systems are classified as high-risk per se. This includes, according to Article 8 (aa):

> *AI systems intended to be used for influencing the outcome of an election or referendum or the voting behavior of natural persons in the exercise of their vote in elections or referenda*.

It can, in our view, easily be argued that the detection and potential deletion of disinformation can influence the outcome of elections (in a positive way, we hope). If now AI systems in the domain of disinformation fall under high-risk, this necessitates their compliance with high-risk obligations, specifically data governance, the creation of an extensive risk management system, transparency obligations, and human oversight as specified by Chapter 2 of the AI Act.

For example, Article 10 on data governance demands that:

> (3) *Training, validation, and testing datasets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose*.

This means that only such GPAIS can be used where the respective data has been documented, which is currently not the case for most commercial models. Moreover, when applied to such open domains as misinformation detection, it is by no means clear what the demand "free of errors and complete" could mean and how this can be proven.

Another obligation of high-risk AI systems in Article 9 on risk management demands that:

> (2) *The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular, systematic review and updating*.

This means that parallel to the disinformation detection process, a monitoring process needs to be installed and maintained to ensure, e.g., the:

> (a) *identification and analysis of the known and reasonably foreseeable risks that the high-risk AI system can pose to the health, safety, or fundamental rights when the high-risk AI system is used in accordance with its intended purpose*.

As discussed above, in the target domain, this could mean constantly checking if the system does not hinder the freedom of expression. As the requirement mentions the entire lifecycle, it might even include the phase of research (that is usually excluded, see below), which would mean that research and development projects would need additional resources and probably also interdisciplinary cooperation to ensure that at any time, there is a well-defined "intended purpose" and a process to come up with "reasonably foreseeable risks".

4. Implications of Regulatory Frameworks

In the following, the implications of mainly the AI Act and the DSA will be analyzed in more detail for the misinformation use case. Within the news-polygraph, several partners from research, industry, and media are concerned with different challenges concerning legal obligations. Thus, the legal implications will be described from three different perspectives, namely the (1) research perspective, (2) the (provider and) deployer perspective, and (3) the user perspective. Notably, the observations below are not to be understood as a legal exegesis but as an attempt to assess the consequences of the AI Act in the domain of disinformation.

4.1. Research Perspective

The impact of the AI Act on research remains very limited. AI systems and models developed and used solely for scientific research and development purposes are explicitly excluded from the scope of the Act. This exemption acknowledges the distinct nature of research activities from commercial or operational AI applications (Haataja and Bryson, 2022). Moreover, the Act clarifies that AI systems used in the context of product-oriented research, testing, and development activities are not subject to its requirements prior to being placed on the market or put into service. This exclusion aims to encourage exploratory research and innovation without imposing premature regulatory burdens. However, researchers still have to consider ethical principles such as human agency, technical robustness, privacy, transparency, diversity, societal well-being, and accountability. However, they are non-binding and serve as a foundational guide for responsible AI development. Researchers are encouraged to consider these ethical principles in their work, aligning research practices with values that promote trustworthiness and human-centric AI (Helberger and Diakopoulos, 2023b). For research and development activities, this approach underscores the importance of assessing and mitigating potential risks associated with AI systems at an early stage. As research does not take place in the void, researchers developing AI systems that may (later) be classified as high-risk are encouraged to incorporate risk assessment and management practices into their development processes. Within research and development projects such as news-polygraph, it is meaningful and rational to consider the risk assessment and compliance with the respective obligations from an early point in time to design potential resulting products in accordance with the AI Act obligations. Additionally, purely research-based LLMs and LGAIMs are exempt from most regulations and can be developed in so-called *regulatory sandboxes*. However, the AI Act aims to foster ethical considerations and prioritize transparency and the mitigation of biases also in *regulatory sandboxes* (Helberger and Diakopoulos, 2023b). Thus, researchers remain merely unaffected by the AI Act as long as the LGAIMs and LLMs are not put into application and commercial use. Moreover, Researchers and stakeholders are encouraged to engage in activities that promote AI literacy, ensuring that AI technologies are accessible and understandable to a broader audience. This initiative aims to build public trust in AI technologies and foster an informed dialogue about AI's role in society.

4.2. Deployer Perspective

The deployer perspective from a commercial viewpoint is distinct from research. Using fine-tuned LGAIMs as a deployer within the EU for use cases, such as disinformation detection, has three main consequences resulting from the fact that these LGAIMs fall under the high-risk category (Hacker et al., 2023). First, deployers will only be able to use LGAIMs from providers who themselves adhere to the obligations for high-risk applications demanded by the AI-Act if they do not develop the LGAIMs themselves. The specific obligations for providers include comprehensive management for quality assurance and system performance and, as a pre-condition, an assessment of conformity and CE-Marking. While the legislation intends the step towards conformity and establishing standards, the distinction between provider and deployer may raise questions in practice, particularly in the case of Open-Source LLMs. Second, deployers must follow a comprehensive list of obligations. This includes the establishment of a risk management system, transparency obligations, need to be in place, indicating that developers and deployers need to build up an in-depth understanding of potentially risky outputs of LGAIMs and their intended use cases. The question of which training data can be lawfully used cannot be answered by the AI Act alone, as additional regulations such as the Data Act and the GDPR need to be considered in deciding the lawful handling of data. This is especially difficult when using LGAIMs from providers where only limited information about the training data is available. The obligations for high-risk AI systems demand representative, complete, and error-free datasets on which the AI systems are trained on. These criteria are very hard to meet, as no concrete metrics or measures are provided in guiding the assessment of datasets. The assumption that AI systems operate accurately, family and without bias when the aforementioned conditions of the dataset are met is misleading, as also model biases can occur not inherent in the training data. Moreover, the requirements of representative and bias-free datasets can be contradicting. When a representative sample is drawn, e.g., about the social media posts of nurses, there might be a clear gender bias towards female nurses. In such cases, it remains very opaque if representative or bias-free datasets are more important. One of the central obligations following form the high-risk categorization is AI systems' transparency and human oversight. Hereby, emerging research in the realm of eXplainable AI (XAI) has demonstrated its capacity to clarify the opaque *black box* aspects of AI algorithms, enhancing the comprehensibility of AI-driven classifications or outputs (Longo et al., 2023; Speith and Langer, 2023). XAI features not only facilitate the understanding of LGAIMs outputs but also the human oversight of such systems for the disinformation detection use case. However, the concrete interpretation of what meaningful explanations allow for transparency and effective human oversight in specific use cases heavily depends on the background of the users (Schmitt et al., 2024). For companies developing high-risk AI systems, it remains challenging to adopt the obligations to concrete technical measures and metrics as only very limited guidance is given. Additionally, providers of GPAIs with a dual-use character, for example, in the domain of media intelligence, are specifically affected by the regulations as their applications could also be classified as high-risk applications. Here, one business application is the detection of company-related dis- or misinformation with the use of LGAIM-based applications. Such applications may, among others, assist in detecting AI-generated content or tracking the diffusion of disinformation. As the detection of company-based disinformation is a subset, disinformation detection providers may attempt to limit the scope of their applications towards company-related disinformation detection in order to circumvent the obligations for high-risk applications. Nevertheless, companies failing to comply with the obligations outlined for the respective risk category may result in high fines, which can reach up to 35 million € or 7% of the company's worldwide annual turnover. Thus, companies, also within research and development projects, need to undergo the risk assessment of applications developed in such frameworks from an early stage to be aligned with the obligations outlined in the AI Act when products are put on the market.

4.3. User Perspective

From a media organization's perspective, the spread of mis- and disinformation is a significant challenge, and it is expected to become even more so in the coming years, particularly when dealing with synthetic and altered media content. However, LGAIMs are not only potential sources of spreading mis- and disinformation but can also assist journalists in uncovering such content. As reported in a white paper by the EU-funded project AI4media, AI technologies are regarded as highly valuable by most fact-checking and verification specialists (AI4, 2022). The debate over whether LGAIMs in the media should be classified as high-risk and subjected to the strictest regulatory measures is closely linked with ongoing discussions about the influence of algorithm-driven platforms and, more broadly, the effects of AI utilization in media on fundamental rights like freedom of speech and privacy rights. As fundamental rights might be affected when using LGAIMs to detect disinformation and harmful content, LGAIMs can be categorized as high-risk AI systems and need to follow the respective obligations (Helberger and Diakopoulos, 2023b). Therefore, among others, effective human oversight, transparency obligations, and a risk management system need to be ensured when applying LGAIMs for mis- and disinformation detection. When using LGAIMs for disinformation detection, the next regulatory framework relevant to their application from a user's perspective is the Digital Services Act (DSA). The user perspective is relevant to consider to gain a more in-depth understanding of the implications of the AI Act on advanced transparency and human oversight of AI systems used for mis-and disinformation detection. While some use cases for applying LGAIMs in the journalistic verification process may be obvious, others may appear less relevant at first glance. Overall, LGAIMs and AI systems can be used differently in the journalistic context, which is also partially covered by tools developed within the news-polygraph research and development project.

LGAIMs-driven tools are widely accepted in the field of Human Language Technologies. These tools, such as plainX[3] allow for the transcription and translation of video and audio content. While these technologies are not primarily designed to detect mis- and disinformation, they are undoubtedly useful for journalists to learn what content in a foreign language is about and whether it is the same as it claims to be about. Videos that intentionally mistranslate the original foreign language speech through incorrect voiceovers or subtitles are often used for entertainment purposes. There are tools available, such as the Caption Generator, that enable users to create content with fictional subtitles for popular videos, such as 'Dimitri Reacts'[4]. However, there are also many examples of critical videos with fake subtitles. For example, Full Fact reported on several videos addressing the ongoing conflict in Israel and the Gaza Strip. One social media video suggests that a Palestinian woman said in Arabic, 'We are prisoners of Hamas,' which is a deliberately incorrect translation[5]. Other videos wrongfully claim to show North Korean leader Kim Jong Un making a speech about the Israel-Gaza conflict or Putin and Erdogan warning America over its support for Israel[6]. These examples show that even AI tools used for translation can result in harmful outputs when AI system predictions are wrong. Wrong translations can result in misinterpretation of the meaning. This can lead to the blocking of such information or printing it as truthful content when no expert-level language knowledge is available to prove the AI-generated translations. However, such AI systems apply rather to the category of *limited risks* and need to comply with minor transparency obligations. Additionally, different AI tools are already used to support journalists in their fact-checking tasks. The InVid WeVerify[7] Chrome plugin provides an advanced forensic toolbox for image verification suspected of being manipulated. Moreover, approaches such as Retrieval-Augmented Generation (RAG) can be used to integrate external knowledge sources for knowledge enrichment for certain fact-checking tasks. For example, the Database of Known Fakes (DBFK)[8] provides a useful integration of external knowledge in multiple languages relevant for checking context information about a specific claim or entity. When using such tools, transparency is highly important to journalists as they need to understand the reasoning behind a specific AI model output for content verification. Thus, independent of the risk category, journalists require sufficient and meaningful transparency to rely on the AI model output. As most of the AI tools applied in the fact-checking and content-verification process apply to the high-risk category, they must integrate transparency measures and effective human oversight. Previous research has

[3] https://www.plainx.com/, last accessed 03.04.2024.

[4] https://www.captiongenerator.com/make-a-dimitri-finds-out-video, last accessed 03.04.2024.

[5] https://fullfact.org/online/fake-subtitles-video-palestinian-woman/, last accessed 03.04.2024.

[6] https://fullfact.org/online/fake-kim-jong-un-north-korea-israel-gaza/, last accessed 03.04.2024.

[7] https://weverify.eu/

[8] https://shorturl.at/kBDHL, last accessed 03.04.2024.

shown that natural language explanations can be easily perceived by humans but also create false trust in the AI system when the predictions or classifications are wrong (Schmitt et al., 2024). Therefore, the high-risk obligations model's faithfulness and robustness are highly important for sensitive tasks such as content verification. Moreover, meaningful explanations heavily depend on the users' prior knowledge and background. Therefore, explanations need to be incorporated to provide explanations on different levels of abstractions that users with varying degrees of expert knowledge can comprehend. From a user's perspective, the obligations defined in the AI Act are very beneficial if implemented adequately. The transparency measures, explanations given, and modes of collaboration for ensuring human oversight need to be designed carefully to allow for the effective integration of human knowledge and human oversight, especially in domains where human rights might be affected.

When combining the three perspectives for the news-polygraph research and development project, the research institutes involved in the project need to consider the obligations defined for high-risk AI systems to prepare the tools for the deployer partners adequately. The deployer partners have to establish adequate procedures for risk assessment (also continuously), data governance structure, technical documentation, accuracy, robustness, and security measures. In collaboration with the user partners, the research and industry partners must develop sufficient means of transparency and meaningful explanations to allow for meaningful collaboration between journalists and AI systems for an overall improved performance on the content verification task.

5. Critique

When analyzing the different obligations within the AI Act, such as transparency, complete and representative datasets for model training, accountability, and fairness, the concrete implications of specific use cases remain opaque, as does a clear definition of process steps such as "research" versus "entire lifecycle of a high-risk AI system", which are subject to very different regulatory measures. Some guidance is given on the risk assessment of AI systems conducted by providers and deployers by themselves, but there is still room for interpreting the risk categorization depending on the provider's/deployers' needs. Due to the comprehensive list of obligations defined in the high-risk category, it can be assumed that deployers will avoid categorizing their AI systems as high-risk AI systems. Deployers need to be fully aware of the consequences the choice of LGAIMs as a technology for production, for example, in media intelligence applications, may entail even if the contribution of the LGAIM to the overall functionality is limited, e.g., if the LGAIM is only used for a final language checking of other system's output in a hybrid setting. They need to carefully select providers of LGAIM based on an assessment of eligible certification and existing quality assurance practices, as the failure to do so could result in massive fines of up to 7% of the annual turnover. As the AI Act requires a constant exchange between deployers and developers of LGAIMs, deployers should assign clear responsibilities for these tasks to their respective managers. While LGAIM-based applications provide various opportunities for improved services to uncover company-based disinformation, the deployment will come at the cost of adhering to the regulations imposed by the AI Act. Deployers may find themselves in a situation where they want to contribute as part of their Corporate Social Responsibility campaign an ad-hoc report about the spread of disinformation in light of an upcoming election and may choose to produce this report without the use of LGAIMs in order to circumvent the regulations imposed by the AI Act or disregard such reports at all. In light of the early stage of implication, deployers will need to follow the developments around the implementation of the AI Act and the legal interpretation made for weakly specified terms in the AI Act across Europe closely, for example, in court rulings or administrative regulations. Moreover, as described in Section 4.3, LGAIMs can be valuable in identifying dis- and misinformation. Journalists require such tools, and several are already available or in development. However, it is crucial to explain these tools' functionalities, outcomes, and constraints. Journalists often work under time constraints while also striving for high credibility. As a result, journalists need to have a certain level of technical skills and AI literacy to be able to recognize the strengths and limitations of the tools they are using. Additionally, journalists must be able to determine whether the use of LGAIMs-based tools complies with the DSA, particularly when processing sensitive data. This may include leaked data or information containing personal data. For example, if data requires verification, LGAIMs that use the inserted information for training should not be applied.

The implementation of GDPR has revealed that without clear technical guidelines and the absence of monitoring mechanisms at both national and EU levels, the regulation may not achieve its intended effectiveness. Previous research (Schmitt et al., 2023) indicates that while GDPR has enhanced certain practices in personal data management, it falls short of establishing precise technical criteria for detecting non-compliance. Despite platforms, applications, and services declaring GDPR adherence through privacy policies and consent forms, these claims often lack verifiable technical substantiation. Similarly, without verification processes to assess compliance with the AI Act from a technical standpoint, this regulation risks being as ineffectual as GDPR, yielding only marginal improvements in ethical AI system practices.

Overall, the regulations must be interpreted and understood depending on specific use cases in which LGAIMs are calibrated. Therefore, we recommend 1) setting minimum standards for LGAIMs and not classifying all LGAIMs as high-risk AI systems, 2) defining high-risk rules specific for LGAIMs employed and used in high-risk scenarios, and 3) establishing standards of adequate transparency, human oversight, and risk management to comply with the rules outlined in the AI Act.

6. Conclusion

In conclusion, the deployment and utilization of GLAIMs for disinformation detection within the complex landscape of the forthcoming AI Act and DSA offer both significant opportunities and difficult challenges. The paper has examined the multifaceted implications of the AI Act, highlighting the nuanced obligations these frameworks impose on research, deployer, and user perspectives in the context of mis- and disinformation detection. Central to the discourse is recognizing LGAIMs as potentially high-risk systems when applied to disinformation detection, necessitating rigorous compliance with a longer list of obligations such as risk management, (training data) transparency, and human oversight. This designation underscores the critical need for deployers and developers to ensure that LGAIMs are not only effective in detecting and mitigating disinformation but also aligned with ethical standards and legal requirements aimed at safeguarding public discourse and protecting fundamental rights. Moreover, we highlight the challenges and ambiguities in interpreting the AI Act's provisions, offering clear standards and guidelines that facilitate the responsible use of LGAIMs in combating disinformation. Time will tell to what extent the issues we consider will remain in the implementation of the AI Act. In summary, this paper provides a targeted analysis of the legal and ethical landscape surrounding the use of LGAIMs for disinformation detection, offering insights into the complexities of navigating regulatory frameworks. It underscores the imperative for a collaborative effort among stakeholders to ensure that the deployment of LGAIMs is both effective in countering disinformation and compliant with evolving legal standards, thereby contributing to the integrity and resilience of information ecosystems in the digital age.

7. Acknowledgements

The research was conducted in the scope of the BMBF project news-polygraph (03RU2U151x).

8. Bibliographical References

2021. Proposal regulation: laying down harmonised rules artificial intelligence.

2022. Use case 1: Deepfake detection - white paper. Technical report, AI4Media. Accessed: 2024-04-04.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Vimala Balakrishnan, Ng Wei Zhen, Soo Mun Chong, Gan Joo Han, and Tan Jiat Lee. 2022. Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, page 103144.

Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Omar Darwish, Yahya Tashtoush, Majdi Maabreh, Rana Al-essa, Ruba Aln'uman, Ammar Alqublan, Munther Abualkibash, and Mahmoud Elkhodr. 2023. Identifying fake news in the russian-ukrainian conflict using machine learning. In *Advanced Information Networking and Applications: Proceedings of the 37th International Conference on Advanced Information Networking and Applications (AINA-2023), Volume 3*, pages 546–557. Springer.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2023. Atman: Understanding transformer predictions through memory efficient attention manipulation. *arXiv preprint arXiv:2301.08110*.

World Economic Forum. 2024. Global risks 2024: At a turning point.

Carl Benedikt Frey and Michael Osborne. 2023. Generative ai and the future of work: A reappraisal. *Brown Journal of World Affairs*, pages 1–12.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Carlos Ignacio Gutierrez, Anthony Aguirre, Risto Uuk, Claire C Boine, and Matija Franklin. 2022. A proposal for a definition of general purpose artificial intelligence systems. *Available at SSRN 4238951*.

Meeri Haataja and Joanna J Bryson. 2022. Reflections on the eu's ai act and how we could make it even better. *TechREG™ Chronicle*, (March 2022).

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123.

Natali Helberger and Nicholas Diakopoulos. 2023a. Chatgpt and the ai act. *Internet Policy Review*, 12(1).

Natali Helberger and Nicholas Diakopoulos. 2023b. The european ai act and how it matters for research into ai in media and journalism. *Digital Journalism*, 11(9):1751–1760.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Dudi Iskandar, Indah Surywati, and Geri Suratino. 2023. Public communication model in combating hoaxes and fake news in ahead of the 2024 general election. *International Journal of Environmental, Sustainability, and Social Science*, 4(5):1505–1518.

Razieh Khamsehashari, Vera Schmitt, Tim Polzehl, Salar Mohtaj, and Sebastian Moeller. 2023. How risky is multimodal fake news detection? a review of cross-modal learning approaches under eu ai act constrains. In *Proc. 2023 ISCA Symposium on Security and Privacy in Speech Communication*, pages 47–51.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. 2023. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *arXiv preprint arXiv:2310.19775*.

Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 97–106.

Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*.

Linda Monsees. 2023. Information disorder, fake news and the future of democracy. *Globalizations*, 20(1):153–168.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-4 technical report.

B Perrigo. 2023. The 2 dollar per hour workers who made chatgpt safer. `https://time.com/6247678/openai-chatgpt-kenya-workers/`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Adam Satariano and Paul Mozur. 2023. The people onscreen are fake. the disinformation is real. *International New York Times*, pages NA–NA.

Vera Schmitt, James Nicholson, and Sebastian Möller. 2023. Is your surveillance camera app watching you? a privacy analysis. In *Science and Information Conference*, pages 1375–1393. Springer.

Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Moeller. 2024. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Felix M Simon, Sacha Altay, and Hugo Mercier. 2023. Misinformation reloaded? fears about the impact of generative ai on misinformation are overblown. *Harvard Kennedy School Misinformation Review*, 4(5).

Timo Speith and Markus Langer. 2023. A new perspective on evaluation methods for explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 325–331. IEEE.