

# A Multi-Level Strategy for Deepfake Content Moderation under EU Regulation

## Research Paper

Luca Deck<sup>1,2,3</sup>, Max-Paul Förster<sup>1</sup>, Raimund Weidlich<sup>1</sup>, and Niklas Kühl<sup>1,2,3</sup>

<sup>1</sup> University of Bayreuth, Germany  
luca.deck@uni-bayreuth.de

<sup>2</sup> Fraunhofer FIT, Bayreuth, Germany

<sup>3</sup> FIM Research Center, Bayreuth, Germany

**Abstract.** The growing availability and use of deepfake technologies increases risks for democratic societies, e.g., for political communication on online platforms. The EU has responded with transparency obligations for providers and deployers of Artificial Intelligence (AI) systems and online platforms. This includes marking deepfakes during generation and labeling deepfakes when they are shared. However, the lack of industry and enforcement standards poses an ongoing challenge. Through a multivocal literature review, we summarize methods for marking, detecting, and labeling deepfakes and assess their effectiveness under EU regulation. Our results indicate that individual methods fail to meet regulatory and practical requirements. Therefore, we propose a multi-level strategy combining the strengths of existing methods. To account for the masses of content on online platforms, our multi-level strategy provides scalability and practicality via a simple scoring mechanism. At the same time, it is agnostic to types of deepfake technology and allows for context-specific risk weighting.

**Keywords:** Deepfakes, EU Regulation, Online Platforms, Content Moderation, Political Communication

## 1 Introduction

Due to rapid progress in the development and availability of artificial intelligence (AI) technologies, deepfakes have found their way into modern society. Deepfakes are AI-generated media content pretending to resemble real persons, objects or events (Art. 3(60) AI Act) and cover a wide variety of applications ranging from artistic use to disinformation. The US elections in 2024, but also current geopolitical conflicts, show that deepfakes are used to manipulate and divide society (Byman et al. 2023). In Germany, a deepfake video with former chancellor Olaf Scholz announcing a ban of the Alternative für Deutschland (AfD) has caused huge media attention (Center for Political Beauty 2023, netzpolitik.org 2023). In its global risk report 2025, the World Economic Forum (WEF) categorized risks of misinformation and disinformation in combination with deepfake technologies as the most relevant risks for the coming years (Foster et al. 2025). To mitigate these risks, the European Union (EU) has passed regulations requiring providers and deployers of AI systems as well as online platforms to make deepfakes transparent. However, enforcing these regulations remains difficult as practitioners and

enforcement agencies face significant challenges (Demková & De Gregorio 2025). First, deepfake technology is rapidly evolving, making reliable detection technically difficult and costly—particularly with huge amounts of content. Second, handling deepfakes is subject to intricate tradeoffs between economic and societal risks, such as efficiency and freedom of speech. At the same time, practical guidance is still scarce. In this paper, we combine expertise from law and IS to foster efficient and effective content moderation of deepfakes under EU regulation. Having an interdisciplinary perspective is essential. IS research provides insights into the technical and economic feasibility of methods to mitigate risks of deepfakes, while legal scholarship addresses accountability and governance, ensuring a proportionate approach. This unique setup allows for a twofold contribution: *First*, we analyze to which extent existing methods are suitable to mitigate societal risks of deepfakes specified under EU regulation. *Second*, we propose a novel multi-level strategy for detecting deepfakes that is both practical and flexible enough to be employed for content moderation on online platforms. The remainder of this work is organized as follows. We start by summarizing societal risks connected to deepfakes and introducing relevant EU regulation on deepfakes in Section 2. Based on a multivocal literature review described in Section 3, Section 4 summarizes existing methods for implementing the requirements of EU regulation. Section 5 synthesizes the strengths and weaknesses of the identified methods into a multi-level strategy that can be adopted by online platforms and enforcement agencies. We also discuss context-specific tradeoffs and challenges in Section 6 and conclude with limitations and prospects of our multi-level strategy in Section 7.

## 2 Background

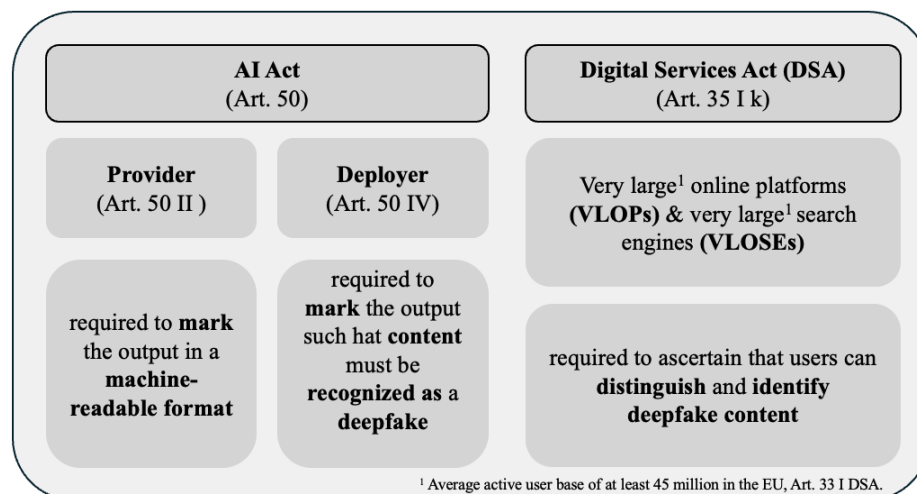
Societal risks connected to deepfakes have prompted the EU to pass several regulatory acts to face and mitigate dangers of deepfakes, including the Artificial Intelligence Act (AI Act) and the Digital Services Act (DSA). Here, we summarize societal risks of deepfakes and introduce relevant EU regulation to mitigate these risks.

### 2.1 Societal Risks of Deepfakes

Actors who instrumentalize deepfakes to influence the individual and public opinion pose a substantial threat to society (Karaboga et al. 2024). The deceptive nature of deepfakes makes it particularly difficult for individuals—especially with low digital literacy—and the public to recognize this content, resulting in unprecedented and hardly foreseeable risks for society. This is particularly evident in the context of digital structural changes in public communication and information strategy (Karaboga et al. 2024). Nowadays, public debates are not only shaped by trusted news agencies who publish curated news content, but also by private individuals with vast reach via online platforms (Eisenegger et al. 2021, Karaboga 2023). The gatekeeper function of the “traditional” media no longer exists in its historic form (Flamme 2024, Eisenegger et al. 2019). Individuals can now be reached directly without any controlling body (Habermas 2022, Karaboga et al. 2024). Although this facilitates deliberation in the opinion-forming process, it also leads to uncontrolled distribution of content and a fading of necessary rules for public

debate (Habermas 2022). This provides the potential for both foreign state and non-state actors or political parties to strategically manipulate the public opinion (Kleemann 2023, Karaboga et al. 2024). These potential threats to society and democracy are a main driver for the EU regulation (Recital 1 AI Act).

## 2.2 EU Regulation on Deepfakes



**Figure 1.** Overview of EU regulation to mitigate dangers of deepfakes.

The AI Act aims to regulate AI systems in the EU and their providers and deployers through a risk-based approach (Von Welser 2024; Art. 2 AI Act). Starting from August 2026, it obliges providers to disclose that the content has been generated or was manipulated by AI-systems (Von Welser 2024; Art. 50(2), 50(4) AI Act). Art. 50 of the AI Act divides the transparency obligation for deepfakes between the use cases of providers (Art. 50(2) AI Act) and deployers (Art. 50(4) AI Act). A provider of AI-systems who generates deepfakes is required to *mark*<sup>1</sup> the output in a machine-readable format (Art. 50(2) AI Act). Additionally, deployers of AI systems who generate deepfakes are required to *mark* the output such that content must be recognized as a deepfake. One of the reasons for this requirement is to facilitate compliance with the transparency obligations under the DSA (Merkle 2024; AI Act Recitals 136). The DSA focuses on the regulation of online intermediary services, such as social media platforms and search engines, that offer their services to users in the EU (Art. 1 DSA). In the context of deepfake regulation, the most relevant transparency obligations only apply to the highest level (Art. 35(1)(k) DSA), i.e., very large online platforms (VLOPs) and very large search engines (VLOSEs). VLOPs and VLOSEs need to *label* content which contains deepfakes (Art. 35(1)(k) DSA).

<sup>1</sup> To avoid ambiguity, we refer to information inherent to the content as “mark”, whereas we refer to information added by online platforms as “label”.

### 2.3 Research Gap

In light of recent technological and legal advances, growing attention is being directed towards interpretations of EU regulation for “ethical AI” (Siau & Wang 2020), sparking debates on specific requirements like transparency (Holst et al. 2024, Hacker & Passoth 2022) and fairness (Wachter et al. 2021, Deck et al. 2024). Prior work has also started to address the practical implications of deepfake regulation. Karaboga et al. (2024) provide an extensive impact assessment of deepfakes and derive practical recommendations to a range of stakeholders, including online platforms, regulators, and society. This work particularly highlights implications for political communication and presents an essential basis for our work. However, it focuses primarily on the Swiss judicial area (Karaboga et al. 2024). On the other hand, a range of studies (e.g., Molavi Vasse 2024, Molavi Vasse & Udoh 2024) respond to EU regulation and propose specific techniques such as watermarking to meet regulatory requirements. However, the majority of the state-of-the-art marking and detection methods rely on a single point of assessment, with none of them presenting a one-size-fits-all solution (see Section 4.2). This calls for a combination of methods into more robust strategies. For example, Gandhi et al. (2024) integrate both graphical and auditory elements into a multimodal detection approach. Moreover, multimodal approaches are also already used in traditional misinformation detection (Zeng et al. 2024).

## 3 Method

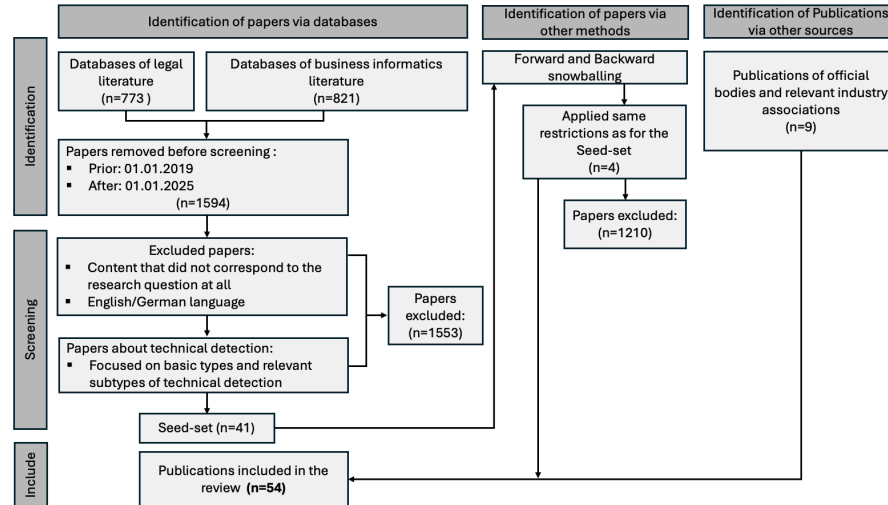
To fill in the gap outlined in Section 2, we surveyed existing methods for marking and detecting deepfakes and analyzed whether they are suitable to mitigate risks of deepfakes under EU regulation. To this end, we conducted a multivocal literature review (Garousi et al. 2019). This method accounts for the rapid recent developments surrounding technical and legal aspects of deepfakes by including grey literature (Schöpfel & Farace 2010) such as preprints and official guidelines. Our literature review strictly followed the central research question: “*How can online platforms effectively and efficiently handle deepfakes to mitigate the societal risks specified in EU regulation?*” To identify relevant literature, we defined a search string consisting of four key elements: regulation, detection, deepfake, and political communication. For each element, we accounted for several variations, synonyms, and related terms. This led to the following search string<sup>2</sup>:

(Act OR Legislation OR Law OR Regulation OR Compliance OR Standards OR Detection) **AND** (deepfake OR deep fake OR synthetic media OR manipulated content OR disinformation OR AI OR Artificial Intelligence OR Machine Learning OR political OR political communication OR misinformation OR Election campaign OR Electoral Integrity OR Democracy)

To combine both the relevant literature on EU regulation as well as computer science and IS, we applied the search string to the databases Scopus, arXiv, Beck, and Juris. The application of the search strings across all databases yielded a total of 1.594 results (dataset-1) published between 01.01.2019 to 01.01.2025. To narrow down the relevant

<sup>2</sup> For the German Beck and Juris databases, we used a translated and slightly shortened search string.

papers, we employed a filtering procedure, which is illustrated in Figure 2 following the PRISMA standard for literature reviews (Page et al. 2021). We first excluded papers that were not available in German or English and papers whose titles and abstracts obviously did not correspond to the research question. Afterwards, we switched to more in-depth screening of abstracts and full texts to select the papers with explicit contributions to our research question. This process reduced a total of 41 papers, which served as a seed set for the further process.



**Figure 2.** PRISMA flowchart describing the source selection procedure.

To identify relevant papers beyond the literature included in the databases and the search strings, we relied on proven guidelines for iterative backward and forward snowballing (Wohlin et al. 2022). Using the literature management tool *Citationchaser* (Haddaway et al. 2022) for the seed set and the same filter criteria as before, we obtained four additional papers. We further augmented the academic literature with publications from official bodies of the EU and the German federal authorities, as well as literature from relevant industry associations, yielding nine additional papers. Following the guidelines of Garousi et al. (2019), we evaluated the gray literature based on authorship, methodology, institutional background, objectivity, and outlet type.

## 4 Results

The final set of 54 papers is organized into three categories, as depicted in Table 1. *Regulatory Guidance* comprises papers that explicitly engage with regulatory requirements, while *Methods* include methods to handle deepfakes, divided into marking, detection, and trusted methods. *Limitations and Risks* includes papers that address technical and practical limitations with regard to EU regulation, and the risks connected to these limitations. The following section summarizes regulatory requirements under the AI Act and the DSA and analyzes the suitability of existing methods for effective and efficient fulfillment of these requirements.

**Table 1.** Overview of the results of the multivocal literature review.

Category	References
<b>Regulatory Guidance</b>	Von Welser (2024), Klos & Taylan (2024), Merkle (2024), Raue & Heesen (2022), European Commission (2024b), German Bundestag, Research Services (2024), European Commission (2024a), Hennemann (2024), Bitkom (2024), Parliament (2023), Karaboga (2023), Gersdorf & Paal (2024), Wybitul (2024), Buchheim & Schrenk (2024), Taeger & Pohle (2024)
<b>Marking Methods</b>	Molavi Vasse (2024), Molavi Vasse & Udoh (2024), Bellaaj & Ouni (2019), Nadimpalli & Rattani (2024), Becker (2024), Boenisch (2021), European Commission (2024a), Dobber et al. (2023)
<b>Technical Detection</b>	Mirsky & Lee (2021), Agrawal et al. (2019), Kaur et al. (2024), Mittal et al. (2020), Marra et al. (2018), Qi & Luo (2022), Le et al. (2023), Baier et al. (2023), Gambín et al. (2024), European Commission (2024a), Zhou & Zafarani (2020)
<b>Trusted Detection</b>	Jaroucheh et al. (2020), Groh et al. (2021), European Commission (2024a), Becker (2024), Jarrahi & Safari (2021), Sänglerlaub (2020), Boenisch (2021), Allen et al. (2021)
<b>Risks and Limitations</b>	European Commission (2024a), Parliament (2023), Gersdorf & Paal (2024), net-politik.org (2023), Dobber et al. (2023), Karaboga et al. (2024), Wybitul (2024), Foster et al. (2025), Buchheim & Schrenk (2024), Gambín et al. (2024), US Department of State (2022), Brömmelhörster (2024), Byman et al. (2023), Chesney & Citron (2019), Eisenegger et al. (2021, 2019), European Parliamentary Research Service (2021), Statista (2023), European Economic and Social Committee (2024), Habermas (2022), <i>Stern</i> (2025), Center for Countering Digital Hate (2020), Klee-mann (2023), Simon et al. (2023), Zhou & Zafarani (2020)

#### 4.1 Regulatory Guidance

The recitals of the AI Act propose a range of technical methods to mark deepfakes, such as metadata, frequency component, cryptographic, and statistical watermarking (AI Act Recital 133). In selecting a technical method, providers should consider the limitations of each method (Art. 50(2) AI Act). Furthermore, the deployer must mark the output in a way that discloses the artificially generated origin (Art. 50(4) AI Act). This can be accomplished, e.g., using banners, overlays, popups, or auditory cues (Bitkom 2024).

The implementation of the labeling obligations under Art. 35 (1)(k) of the DSA presents a more significant challenge for the obligated parties than the implementation of the transparency obligations under Art. 50 of the AI Act. This is because VLOPs cannot trace the origin of content from third parties (users) with sufficient certainty. For example, users outside of the EU have access to technology that is not subject to the transparency obligations under the AI Act, and even within the EU, there is no guarantee that users comply with these transparency obligations. Before labeling deepfakes, VLOPs must therefore implement detection frameworks that can distinguish real content from deepfakes. Consequently, a robust detection framework for deepfakes is a fundamental prerequisite for the successful implementation of the transparency obligations under Art. 35(1)(k) of the DSA.

## 4.2 Marking and Detection Methods

*Marking Methods.* As part of the implementation of the AI Act, the marking methods mentioned above are of particular relevance. Metadata watermarking embeds the mark directly in the metadata (Molavi Vasse 2024). This method has the advantage of not changing the content itself (e.g., in terms of quality; Molavi Vasse 2024). However, this method is vulnerable to manipulation, in particular through the automatic removal of the metadata (Molavi Vasse & Udoh 2024). A complex but more robust method is frequency component watermarking, which includes a watermark in the frequency domain of a digital signal (Bellaaj & Ouni 2019, Molavi Vasse & Udoh 2024). Similarly, cryptographic watermarking is robust to manipulation by integrating a signature into the content itself (Molavi Vasse 2024). However, it cannot be applied to text content (Molavi Vasse & Udoh 2024). Statistical watermarking embeds a statistical pattern into the content, e.g., by changing pixel values (Molavi Vasse & Udoh 2024). This watermark is not detectable by the human eye and constitutes a robust method for labeling content (Molavi Vasse 2024). Notably, none of the watermarking methods can withstand all forms of manipulation attacks (Molavi Vasse & Udoh 2024). Moreover, the methods need to be constantly adapted to the evolving landscape of deepfake technologies and manipulation attacks to ensure long-term robustness.

*Technical Detection Methods.* As previously outlined, deepfake detection methods are practically necessary to fulfill the obligations under Art. 35(1)(k) of the DSA. The large number of deepfake detection frameworks can be divided into different fundamental types with homogeneous strengths and weaknesses.

One example of *artefact-based methods* focuses on behavior. It uses monitoring of facial expressions and gestures to detect anomalies (Mirsky & Lee 2021). Although this method requires large amounts of training data, it has the potential to be an effective detection method for popular individuals such as politicians (Agrawal et al. 2019). However, the applicability of such approaches is limited (Kaur et al. 2024). First, they require the presence of specific artefacts (Mittal et al. 2020). Second, the restriction to a specific artefact or group of artefacts makes them easy to anticipate for detection bypasses (Kaur et al. 2024, Le et al. 2023).

Unlike artefact-specific methods, *undirected methods* do not rely on one pre-defined set of artefacts. Instead, the algorithm autonomously selects distinctive criteria, either via classification or anomaly detection (Mirsky & Lee 2021). Classification-based methods rely on supervised deep learning models and labeled images to differentiate between real content and deepfakes. Although they can perform significantly better than artefact-specific methods (Marra et al. 2018), they also show limitations. First, the training of such models requires a large and representative amount of labeled deepfakes, which is difficult to obtain (Kaur et al. 2024, Qi & Luo 2022). This makes them prone to concept drift (Baier et al. 2023, Kaur et al. 2024), i.e., loss of performance due to changes in the environment (e.g., new deepfake technologies). Further, deep learning models are blackboxes, which makes it difficult to retrace and challenge the reasoning for classification decisions (Kaur et al. 2024). Another undirected approach is anomaly detection, which requires real images to detect differences from deepfakes (Mirsky & Lee 2021). However, due to the increasing amount of artificially generated or manipulated

digital content, training data (i.e., real images) must be carefully curated to avoid mislabeled data (Molavi Vasse 2024).

*Trusted Methods.* In contrast to fully technical detection methods, trusted methods use collective intelligence or human knowledge to detect deepfakes. Jaroucheh et al. (2020) differentiates between “expert-based fact checking” and “crowd-sourced fact checking”. Expert-based fact checking focuses not only on superficial patterns, but can also account for context (Jaroucheh et al. 2020). For example, deepfakes might be recognized due to untrustworthy statements made in the deepfake itself. However, independent and accepted experts are scarce (Jaroucheh et al. 2020) and not sufficiently efficient for large amounts of content in a short period of time (Sängerlaub 2020), limiting its applicability for VLOPs. Crowd-sourced fact checking relies on collective intelligence of many individuals rather than expert individuals (Jaroucheh et al. 2020). For example, collective signing has a selected group of verifiers independently reviewing content (Jaroucheh et al. 2020). The advantages of these methods are scalability and diversity of expertise and knowledge (Jaroucheh et al. 2020). Initial evidence suggests that representative crowds can classify disinformation just as accurately as experts (Allen et al. 2021). However, it requires sufficient incentives for the verifiers (Jaroucheh et al. 2020) and poses the risk of external manipulation (Allen et al. 2021). Therefore, VLOPs would face the challenge to establish scalable systems with sufficient availability and performance in real-time (Jaroucheh et al. 2020).

## **5 A Multi-Level Strategy for Deepfake Detection and Labeling**

With an understanding of the range of available methods and their individual strengths and weaknesses at hand, we combine the strengths into a multi-level strategy that can be adopted by online platforms and enforcement agencies to label potential deepfakes in a proportionate manner, to overcome the current systemic challenges of the EU.

### **5.1 Multi-Level Detection**

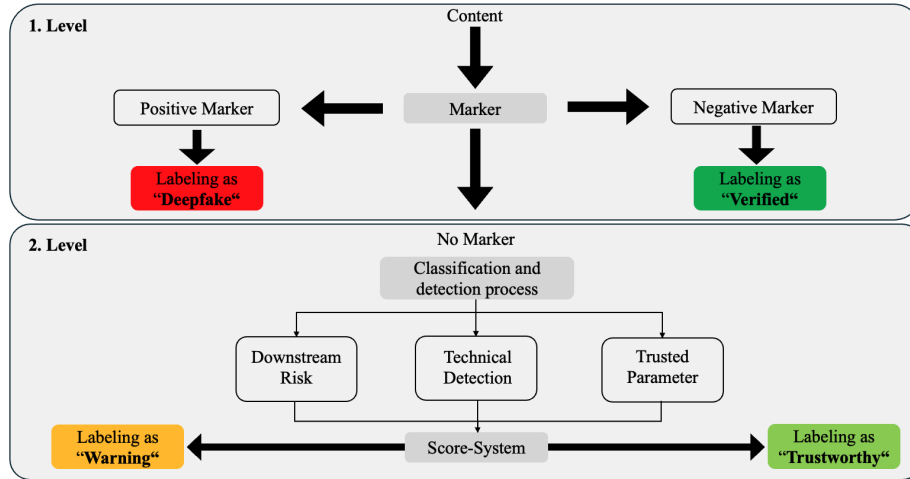
As online platforms cannot reliably expect deepfakes to be marked, two levels are required. As depicted in Figure 3, the first level pre-categorizes content based on markers that may be embedded under the AI Act. The second level employs a more complex procedure to account for content without markers for efficient handling of difficult cases.

The first level checks markers such as watermarks (see Section 4.2). Positive markers mark the content as a deepfake and will be increasingly common due to the transparency obligations in the AI Act. As a counterpart, negative markers can authenticate content that has not been modified or created by artificial intelligence (Nadimpalli & Rattani 2024, Becker 2024). Although negative markers are currently not widespread, it can be anticipated that they will play a more significant role in the future, e.g., due to new regulations and industry standards. Crucially, markers—in particular negative markers—require a verified certification chain to prevent potential tampering (Boenisch 2021).

In the absence of markers, e.g., when content originates from outside the EU or is intentionally deceptive (Kleemann 2023, Chesney & Citron 2019), deepfakes need to



be detected. To combine the individual strengths of existing methods, we propose a multimodal approach that combines methods from both detection categories identified in Section 4.2, i.e., technical detection and trusted detection. While technical detection



**Figure 3.** Multi-level strategy using markers and classification mechanisms to label potential deepfake content.

methods are highly scalable and sensitive to minor details, trusted detection methods can account for the context of the content. Also, the detection process accounts for potential downstream risks, which have to be classified based on the content and its context, such as origin and potential reach. This is particularly relevant when outcomes of individual detection methods differ or a clear result cannot be achieved quickly enough. The systemic risks described in Art. 34(1)(c) of the DSA could be used to classify downstream risk, in particular threats to our democratic society, such as political communication and catastrophic events. Similar to deepfake detection, the downstream risk of specific content can be classified using technical approaches or human expertise. The exact selection and configuration of methods, as well as the classification of downstream risk, depend on the contextual requirements. Still, we shortly sketch how our multi-level strategy might be instantiated for a VLOP. As technical methods, undirected methods are broadly applicable and highly scalable to detect a majority of deepfakes with high certainty. Trusted methods become particularly relevant when technical methods fail or the content embodies high downstream risk (Gambín et al. 2024). Similar to Google’s reCAPTCHA (Vinay 2014), VLOPs could create collective signing systems for paid, voluntary, or mandatory content moderation. Additionally, they could install experts for intricate cases where average people struggle (Groh et al. 2021). With regard to downstream risk, content on VLOPs could be classified into harmless categories, such as animals, and potentially harmful categories, such as political communication. In addition, selected content providers, such as media outlets following certain journalistic standards, could be verified (European Commission 2024a, Becker 2024, Jarrahi & Safari 2021), to add more context during the detection procedure.

## 5.2 Scoring and Labeling

Score (S) = Technical Detection + Trusted Detection + Downstream Risk			
Labeling		Score-System Output	
		First Level	Second Level
Labeling as “Deepfake”	The content is a confirmed deepfake.	Positive Marker	n.a.
Labeling as “Warning”	It is highly probable that the content is a deepfake of a dangerous <sup>1</sup> nature.	No Marker	S = 2 or 3
Labeling as “Trustworthy”	It is highly probable that the content is <u>not</u> a deepfake of a dangerous <sup>1</sup> nature.	No Marker	S = 0 or 1
Labeling as “Verified”	The content is confirmed real.	Negative Marker	n.a.

<sup>1</sup> A deepfake that poses a threat to democratic society or the rule of law, e.g. AI-Act Recital 1.

**Figure 4.** Exemplary scoring and labeling system based on our multi-level strategy.

As the ultimate goal is not detection but labeling of deepfakes, our multi-level strategy also guides decisions on how to assign labels. In order to make a final decision on the classification of the content, a score system should combine three aspects: technical detection, trusted detection, and downstream risk. The exemplary score system in Figure 3 assigns binary values (0 if untrustworthy and 1 if trustworthy) and weights all three values equally. These values are combined into an aggregate score that enables justifiable decisions about the classification and labeling of content. For more fine-granular classification, the scoring can be refined, e.g., by modifying the weights based on criteria such as technical reliability or expertise. The multimodal detection approach results in four potential labeling levels as outlined in Figure 4. Positive markers are directly labeled as “Deepfakes”. Contrarily, negative markers warrant a “Verified” label (European Commission 2024a), (Dobber et al. 2023). The rest of the labels depend on the result of the score system from Figure 3. Such transparent scoring and labeling could lead to more trust in the content moderation process and the content itself (Dobber et al. 2023).

## 6 Discussion

By combining the merits of multiple approaches, our multi-level strategy could be more robust against the manifold types and risks of deepfakes. Importantly, however, the strategy must be instantiated based on context-specific requirements regarding economic practicality and legal compliance. Here, we shed light on three key challenges and embed them in a broader discourse.

## 6.1 Misclassification

A high amount of mislabeled content is expected to lead to a loss of trust in the classification system (Karaboga et al. 2024). Therefore, it is paramount to minimize misclassifications by means of technical state-of-the-art and a combination of methods. Still, no deepfake detection system will ever guarantee perfect accuracy, leading to intricate tradeoffs between types of misclassifications. In this context, actual deepfakes classified as verified or trustworthy are regarded as *false negatives*, while real content classified as deepfake or untrustworthy presents a *false positive*. Notably, the scoring system depicted in Figure 4 allows online platforms to trade-off false positives with false negatives by setting a more or less sensitive threshold and weighting of the three values. It does, however, not specify how to set this threshold. This decision depends on context-specific priorities and needs to consider the respective costs of false positives and false negatives. False negatives (deepfakes labeled as verified content) entail the very risks to our democratic society that the AI Act aims to mitigate, i.e., the spread of misinformation or even intentional disinformation campaigns. At the same time, singular false positives without high downstream risk or low reach are less costly, making the potential costs difficult to estimate. On the other hand, costs of false positives (real content labeled as deepfake) entail unwarranted loss of trust in the specific content and reputational damage for the source of the content (Dobber et al. 2023). False positives may also severely limit freedom of speech or artistic freedom, posing another risk to our democratic society (Kozyreva et al. 2023). Future research must examine how to determine these tradeoffs. Costs of misclassifications might be further decreased by establishing fallback strategies that, e.g., allow for labeling or banning deepfakes even after they have been published on an online platform. Rare cases of critical downstream risk and high uncertainty might also warrant expert assessments, which may include human-AI collaboration combining technical methods with human expertise.

## 6.2 Specification of Labeling Obligations

A major question arises regarding the scope and limits of the labeling obligation. Art. 35(1)(k) of the DSA and Art. 50(4) in connection with Recital 133 of the AI Act are both framed very vaguely and constantly refer to the possible state-of-the-art. From the perspective of online platforms, however, clarification is needed on whether verification of real content falls under the obligations of the regulations. Art. 50(4) of the AI Act only provides labeling obligations for providers and deployers of AI systems that generate deepfakes. On the other hand, Art. 35(1)(k) of the DSA demands “*reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks*”. In addition, the EU Commission’s guidelines already point toward cases such as scientific research where sources should be actively labeled as trustworthy (European Commission 2024a, Dobber et al. 2023). One could therefore argue that in some cases, only multi-level labeling systems constitute “*effective mitigation measures*” under Art. 35(1)(k) of the DSA, particularly if one is aware of the regulation goals, which also aim to ensure a “*safe, predictable and trusted online environment*” (Art. 1(1) DSA). As discussed before, trust in the content on VLOPs can only be effectively established through the classification and labeling of verified and trustworthy content in combination with a

warning against deepfakes and untrustworthy content. Against this backdrop, even if explicit obligations are currently debatable, we expect multi-level labeling systems to become more relevant for types of content posing high potential downstream risk, such as political communication.

### **6.3 Enforcement of the Labeling Obligations**

Lastly, for the transparency obligations under EU regulation to become effective, they will require proactive enforcement. In the context of VLOPs, the EU Commission is specifically responsible for the enforcement (Art. 56(2), 56(3), 49(4) DSA), which is to be carried out by an independent auditing organization (Art. 37(1)(a), 37(3) DSA). Random sample tests may be a feasible option to evaluate the performance of content moderation systems employed by VLOPs. Nonetheless, effective enforcement requires auditing authorities to have an understanding of practical solutions themselves. Ultimately, we encourage them to collaborate on context-specific applications of our multi-level strategy and to test different configurations of detection methods and score systems.

## **7 Conclusion**

As the origin of digital content can not be reliably controlled, and deepfake detection is a moving target, online platforms and enforcement agencies alike are struggling to put the legal requirements of DSA and AI Act into practice. Combining the strengths of existing detection methods, we propose a multi-level strategy for online platforms to efficiently label content and effectively mitigate societal risks under EU regulation. While our proposal could be more holistic and robust than existing approaches, future work must evaluate whether our proposal holds in real application settings. The multi-level strategy has not been evaluated by external experts or tested and is merely a blueprint that does not specify concrete methods, risk weightings or implications for action. We also leave many related questions for future work such as proportionality with regard to freedom of expression and other fundamental rights from the EU charter. We perceive our multi-level strategy to be a building block in the larger endeavor to make democratic society resilient to the risks of deepfakes. A fully holistic solution will—among others—require a better understanding of the societal risks of deepfakes on online platforms (Al-Khazraji et al. 2023), substantial efforts to improve media literacy of users (European Economic and Social Committee 2024), and more interdisciplinary collaboration for practical solutions.

## **8 Acknowledgements**

Work on this paper has been supported by the project „For the Greater Good? Deepfakes in Criminal Prosecution (FoGG)“ funded by the Bavarian Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities, under project code KON-024-008.

## References

- Agrawal, S., Farid, H., Gu, Y., He, M., Nagano, K. & Li, H. (2019), 'Protecting world leaders against deep fakes', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* p. 38–45.
- Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I. & Mishkhal, I. A. (2023), 'Impact of deepfake technology on social media: Detection, misinformation and societal implications', *The Eurasia Proceedings of Science Technology Engineering and Mathematics* **23**, 429–441.
- Allen, J. N., Arechar, A. A., Pennycook, G. & Rand, D. G. (2021), 'Scaling up fact-checking using the wisdom of crowds', *Science Advances* **7**(36).
- Baier, L., Schlör, T., Schöffel, J. & Köhl, N. (2023), 'Detecting concept drift with neural network model uncertainty', *Proceedings of the 56th Hawaii International Conference on System Sciences. - Honolulu, HI* p. 836–844.
- Becker, M. (2024), 'Generative KI und Deepfakes in der KI-VO für eine Positivkennzeichnung authentischer Inhalte', *Computer und Recht* **40**(6), 353–366.
- Bellaaj, M. & Ouni, K. (2019), 'Watermarking technique for multimedia documents in the frequency domain. digital image and video watermarking and steganography [working title]', *IntechOpen*.
- Bitkom (2024), 'Umsetzungsleitfaden zur KI-Verordnung', <https://www.bitkom.org/sites/main/files/2024-10/241028-bitkom-umsetzungsleitfaden-ki.pdf>. Accessed: 10.01.2025.
- Boenisch, F. (2021), 'A systematic review on model watermarking for neural networks', *Frontiers in Big Data* **4**.
- Brömmelhörster, L. (2024), 'Potentielle Gefahren von Deepfakes', *Kriminalistik* (4), 248–256.
- Buchheim, J. & Schrenk, M. (2024), 'Geldbußen wegen Verstößen gegen die KI-Verordnung', *Neue Zeitschrift für Verwaltungsrecht* **43**(1).
- Byman, D. L., Gao, C., Meserole, C. & Subrahmanian, V. (2023), Deepfakes and international conflicts, Technical report, Brookings Institution.
- Center for Countering Digital Hate (2020), 'MALGORITHM — How Instagram's algorithm publishes misinformation and hate to millions during a pandemic', <https://counterhate.com/wp-content/uploads/2022/05/210309-Malgorithm-Report.pdf>. Accessed: 30.01.2025.
- Center for Political Beauty (2023), 'Der Bundeskanzler greift durch: Die AfD muss verboten werden!', <https://politicalbeauty.de/afd-verbot.html>. Accessed: 18.06.2025.
- Chesney, R. & Citron, D. K. (2019), 'Deep fakes: A looming challenge for privacy, democracy, and national security', *California Law Review* **107**, 1754–1820.
- Deck, L., Müller, J.-L., Braun, C., Zipperling, D. & Köhl, N. (2024), Implications of the AI act for non-discrimination law and algorithmic fairness, in 'European Workshop on Algorithmic Fairness (EWAf'24)'.
- Demková, S. & De Gregorio, G. (2025), 'The looming enforcement crisis in European digital policy: A rule-of-law centered path forward', *Verfassungsblog*.

- Dobber, T., Kruikemeier, S., Votta, F., Helberger, N. & Goodman, E. P. (2023), 'The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media', *Journal of Information Technology & Politics* **22**(1), 82–97.
- Eisenegger, M., Prinzing, M., Ettinger, P. & Blum, R. (2021), *Digitaler Strukturwandel der Öffentlichkeit*, Nomos.
- Eisenegger, M., Udris, L. & Ettinger, P. (2019), *Öffentlichkeit und Demokratie unter Bedingungen eines radikalen medialen und sozialen Wandels*, Springer VS, p. 31–51.
- European Commission (2024a), 'Commission guidelines for providers of very large online platforms and very large online search engines on the mitigation of systemic risks for electoral processes pursuant to article 35(3) of regulation (eu) 2022/2065', <https://eur-lex.europa.eu/eli/C/2024/3014/oj/eng>. Accessed: 30.01.2025.
- European Commission (2024b), 'Supervision of the designated very large online platforms and search engines under DSA', <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>. Accessed: 22.01.2025.
- European Economic and Social Committee (2024), 'Safeguarding democracy against disinformation', <https://eur-lex.europa.eu/eli/C/2024/4052>. Accessed: 30.01.2025.
- European Parliamentary Research Service (2021), 'Tackling deepfakes in European policy', [https://www.europarl.europa.eu/thinktank/en/document/EPRS\\_STU\(2021\)690039](https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039). Accessed: 30.01.2025.
- Flamme, F. (2024), *Meinungsmacht der Algorithmen*, Nomos.
- Foster, I., Carl, K., Nick, J. & Tuecke, S. J. (2025), The Global Risks Report 2025, World Economic Forum, 2025, Technical report, World Economic Forum.
- Gambín, A. F., Yazidi, A., Vasilakos, A., Haugerud, H. & Djenouri, Y. (2024), 'Deep-fakes: Current and future trends', *Artificial Intelligence Review* **57**(3).
- Gandhi, K., Kulkarni, P., Shah, T., Chaudhari, P., Narvekar, M. & Ghag, K. (2024), 'A multimodal framework for deepfake detection', *Journal of Electrical Systems* **20**(10).
- Garousi, V., Felderer, M. & Mäntylä, M. V. (2019), 'Guidelines for including grey literature and conducting multivocal literature reviews in software engineering', *Information and Software Technology* **106**, 101–121.
- German Bundestag, Research Services (2024), 'Regulierung von Deepfakes', <https://www.bundestag.de/resource/blob/997214/2bc9f3aeda343398c69deb3c32badf3a/WD-7-015-24-pdf.pdf>. Accessed: 22.01.2025.
- Gersdorf, H. & Paal, B. P. (2024), *BeckOK Informations- und Medienrecht*, C.H.Beck.
- Groh, M., Epstein, Z., Firestone, C. & Picard, R. (2021), 'Deepfake detection by human crowds, machines, and machine-informed crowds', *Proceedings of the National Academy of Sciences* **119**(1).
- Habermas, J. (2022), *Ein Neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*, Suhrkamp Verlag.
- Hacker, P. & Passoth, J.-H. (2022), Varieties of AI explanations under the law. from the GDPR to the AIA, and beyond, in A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller & W. Samek, eds, 'xxAI – Beyond Explainable AI', Vol. 13200 of *Springer*

- eBook Collection*, Springer International Publishing and Imprint Springer, Cham, pp. 343–373.
- Haddaway, N. R., Grainger, M. J. & Gray, C. T. (2022), ‘Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching’, *Research Synthesis Methods* **13**(4), 533–545.
- Hennemann, M. (2024), ‘Der European Media Freedom Act oder: Von Zielen und Mitteln des Medienrechts’, *Zeitschrift für europäisches Privatrecht* **6**(3), 498–504.
- Holst, L., Mayer, V., Lämmermann, L., Urbach, N. & Wendt, D. (2024), ‘The impact of the EU AI act’s transparency requirements on AI innovation’, *Wirtschaftsinformatik 2024 Proceedings* **92**.
- Jaroucheh, Z., Alissa, M., Buchanan, W. J. & Liu, X. (2020), ‘Trustd: Combat fake content using blockchain and collective signature technologies’, *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)* p. 1235–1240.
- Jarrahi, A. & Safari, L. (2021), ‘Fr-detect: A multi-modal framework for early fake news detection on social media using publishers features’, *arXiv*.
- Karaboga, M. (2023), ‘Die regulierung von deepfakes auf eu-ebene: Überblick eines flick-entepichs und einordnung des digital services act- und ki-regulierungsvorschlags’, *Digitale Hate Speech* p. 197–220.
- Karaboga, M., Frei, N., Puppis, M., Vogler, D., Raemy, P., Ebbbers, F., Runge, G., Rauchfleisch, A., de Seta, G., Gurr, G., Friedewald, M. & Rovelli, S. (2024), Deepfakes und manipulierte Realitäten, Technical report, vdf- Hochschulverlag AG an der ETH Zürich.
- Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S. & Xia, F. (2024), ‘Deepfake video detection: Challenges and opportunities’, *Artificial Intelligence Review* **57**(6).
- Kleemann, A. (2023), *Deepfakes - wenn wir unseren Augen und Ohren nicht mehr trauen können*, Stiftung Wissenschaft und Politik - Deutsches Institut für internationale politik und Sicherheit.
- Klos, C. & Taylan, R. (2024), ‘Von der theorie zur praxis: Die eu-ki-verordnung effektive umsetzen’, *Corporate Compliance Zeitschrift* **17**(7-8), 205–211.
- Kozyreva, A., Herzog, S., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M. & Reifler, J. (2023), ‘Resolving content moderation dilemmas between free speech and harmful misinformation’, *PNAS*.
- Le, B., Tariq, S., Abuadbbba, A., Moore, K. & Woo, S. (2023), ‘Why do facial deepfake detectors fail?’, *Association for Computing Machinery* p. 24–28.
- Marra, F., Gragnaniello, D., Cozzolino, D. & Verdoliva, L. (2018), ‘Detection of GAN-generated fake images over social networks’, *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* p. 384–389.
- Merkle, M. L. (2024), ‘Transparenz nach der KI-Verordnung – von der Blackbox zum Open-Book?’, *Recht Digital* pp. 414–420.
- Mirsky, Y. & Lee, W. (2021), ‘The creation and detection of deepfakes’, *ACM Computing Surveys* **54**(1), 1–41.
- Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. & Manocha, D. (2020), ‘Emotions don’t lie: An audio-visual deepfake detection method using affective cues’, *Proceedings of the 28th ACM International Conference on Multimedia* p. 2823–2832.
- Molavi Vasse, R. (2024), ‘Watermarking von KI-generierten Inhalten als regulatorisches Instrument’, *Recht Digital* pp. 406–414.

- Molavi Vasse, R. & Udoh, G. (2024), 'In Transparency We Trust? Evaluating the Effectiveness of Watermarking and Labeling AI-Generated Content', <https://foundation.mozilla.org/en/research/library/in-transparency-we-trust/research-report/>. Accessed: 22.01.2025.
- Nadimpalli, A. V. & Rattani, A. (2024), 'Social media authentication and combating deepfakes using semi-fragile invisible image watermarking', *Digital Threats: Research and Practice* **5**(4), 1–30.
- netzpolitik.org (2023), 'Landgericht Berlin verbietet Kanzler-Deepfake', <https://netzpolitik.org/2024/kunst-aktion-fuer-afd-verbot-landgericht-berlin-verbietet-kanzler-deep-fake/>. Accessed: 18.06.2025.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E. et al. (2021), 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', *BMJ* **372**.
- Parliament, T. E. (2023), 'Artificial intelligence: Threats and opportunities', <https://www.europarl.europa.eu/topics/en/article/20200918STO87404/artificial-intelligence-threats-and-opportunities>. Accessed: 30.01.2025.
- Qi, G.-J. & Luo, J. (2022), 'Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(4), 2168–2187.
- Raue, B. & Heesen, H. (2022), 'Der digital services act', *Neue Juristische Wochenschrift* **75**(46), 3537–3543.
- Schöpfel, J. & Farace, D. (2010), *Grey Literature in Library and Information Studies*, De Gruyter Saur.
- Siau, K. & Wang, W. (2020), 'Artificial intelligence (AI) ethics', *Journal of Database Management* **31**(2), 74–87.
- Simon, F. M., Altay, S. & Mercier, H. (2023), 'Misinformation reloaded? fears about the impact of generative ai on misinformation are overblown', *Harvard Kennedy School Misinformation Review* **4**(5).
- Statista (2023), 'Anzahl der monatlich aktiven Nutzer nach Region weltweit, Q4 2023', <https://de.statista.com/statistik/daten/studie/885734/umfrage/anzahl-der-monatlich-aktiven-nutzer-von-facebook-nach-regionen/>. Accessed: 29.01.2025.
- Stern* (2025), <https://www.stern.de/politik/deutschland/deepfakes-im-wahlkampf--sind-ki-generierte-videos-eine-gefahr--35401984.html>. Accessed: 29.01.2025.
- Sängerlaub, A. (2020), *Fake News und Desinformation*, Nomos, chapter Fakten versus Fakes: was Fact-Checking als Maßnahme gegen Desinformation leisten kann und was nicht.
- Taeger, J. & Pohle, J. (2024), *Computer-Handbook – Informationstechnologie in der Rechts- und Wirtschaftspraxis*, C.H.Beck.
- US Department of State (2022), 'US department press briefing - february 3, 2022', <https://www.state.gov/briefings/department-press-briefing-february-3-2022/>. Accessed: 18.01.2025.



- Vinay, S. (2014), 'Google reCAPTCHA'. Accessed: 14.03.2025.  
**URL:** <https://security.googleblog.com/2014/12/are-you-robot-introducing-no-captcha.html>
- Von Welser, M. (2024), 'Die KI-Verordnung – ein Überblick über das weltweit erste Regelwerk für künstliche Intelligenz', *GRUR-Prax* **16**(15), 485–489.
- Wachter, S., Mittelstadt, B. & Russell, C. (2021), 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI', *Computer Law & Security Review* **41**.
- Wohlin, C., Kalinowski, M., Romero Felizardo, K. & Mendes, E. (2022), 'Successful combination of database search and snowballing for identification of primary studies in systematic literature studies', *Information and Software Technology* **147**, 106908.
- Wybitul, T. (2024), 'Geldbußen wegen Verstößen gegen die KI-Verordnung', *Neue Juristische Wochenschrift* (37), 2641–2712.
- Zeng, X., La Barbera, D., Roitero, K., Zubiaga, A. & Mizzaro, S. (2024), 'Combining large language models and crowdsourcing for hybrid human-AI misinformation detection', *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* p. 2332–2336.
- Zhou, X. & Zafarani, R. (2020), 'A survey of fake news', *ACM Computing Surveys* **53**(5), 1–40.