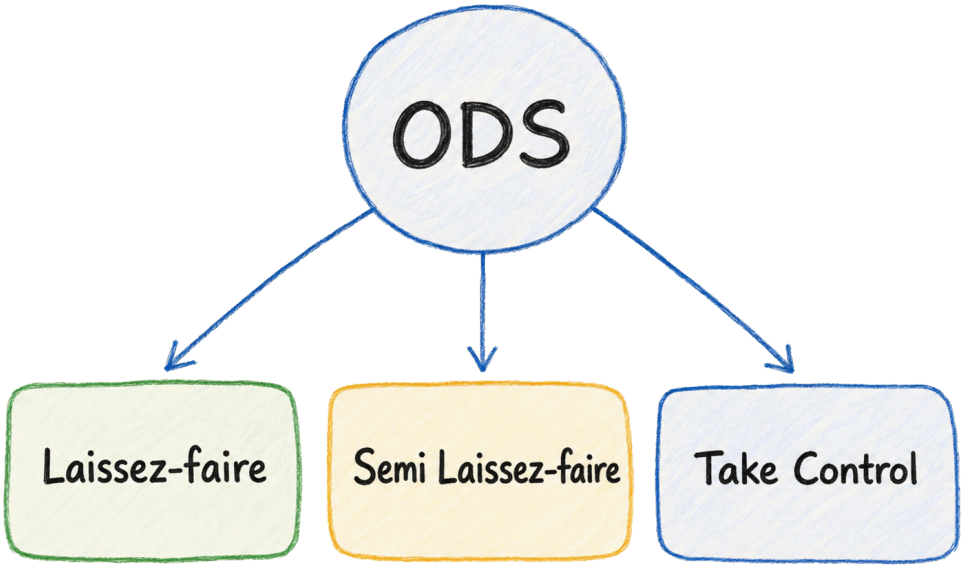
Catch me if you can:

How to handle deepfakes for ODS bodies under art. 21 DSA.



Information Law and Policy Lab

Universiteit van Amsterdam

Table of Contents

Table of Contents **0**
1. Introduction and research questions **2**
2. Methods **4**
2.1 Interviews 4
3. Legal framework: ODS-bodies under the DSA **5**
3.1 Introduction 5
3.2 Definitions of AI-generated content and deepfakes under the AI-Act 5
3.3. Deepfake obligations under the AI-Act 6
3.4 The DSA Framework for ODS Bodies 7
3.5 Interplay DSA and AI-Act 8
3.6 Platform Terms & Conditions 9
3.7 Practical examples 11
3.8 Intermediary Concluding remarks on the legal framework 12
4. Shifting the burden of proof in ODS cases. **13**
5. Dealing with deepfakes, different approaches **16**
5.1 Laissez-faire Approach 16
5.2 Semi-laissez-faire Approach 21
5.3 Take-control Approach 24
6. Conclusion **27**
7. Bibliography **29**
7.1 Legislature: 29
7.2 Institutional Documents: 30
7.3 Academic resources: 31

About the report

This policy report was written for User Rights, an out of court dispute settlement body. The report was written by Information Law students at the University of Amsterdam, as part of the Information Law and Policy Lab. This policy report does not reflect the position of User Rights.

Authors:

Sophie van den Boogaard

Sterre Fleer

Emil de Jong

Matteo Tonicello

Supervisor:

João Pedro Quintais

Published: 4 September 2026, Amsterdam, the Netherlands.

1. Introduction and research questions

Under Article 21 of the Digital Services Act, ODS bodies review whether platforms' moderation decisions were justified. This includes cases in which users' content was removed or labelled and users request that their content be restored or that labels be removed. It also includes cases where users report content they believe violates platform rules or legal provisions and should be removed or potentially labelled.

In many cases, the question of whether content is AI-generated is essential for the assessment by ODS bodies. Yet ODS bodies often struggle to carry out this assessment. For example, when a user reports content as violating platform policies, such as social media company policies, and claims the violation is due to the reported content being AI-generated, ODS bodies face difficulty assessing whether that claim is true.

Absent a means to determine whether content is AI-generated, ODS bodies may rely on procedural mechanisms, such as shifting the burden of proof onto platforms. Instead of the complainant (the platform user) proving that the content is AI-generated, the ODS body would now require the platform to prove that it is not. Another approach is to make "cautious" decisions. If it cannot be determined with certainty whether content is AI-generated, ODS bodies can collect indicators suggesting that it might be. They then share these indicators with the platform and ask it to reassess the content in light of them.

This means that where there is suspicion that content is AI-generated and it is relevant to the case, ODS bodies may request additional reviews by platforms to determine whether the content is, in fact, AI-generated. In practice, this results in the matter being referred back to the platforms, with ODS bodies relying on the platforms' own assessments and actions. This can be viewed as unsatisfactory and arguably counter to the DSA's purpose of transparency.

The detection of deepfakes can also be relevant in other areas of the DSA, such as systemic risk assessments, or, once entered into application, to the assessment of compliance with the deepfake labelling obligations under the AI Act.[1]

One possible avenue to address the issues above is the use of deepfake detection tools. The development of these is still in its early stages, but addressing the promises and perils is important for gaining a clear understanding. Another is to ensure that solid processes are in place to address

[1] Article 34 of the DSA and Article 9 AI Act.

remaining uncertainties. Underlying these issues is the question of what the appropriate normative framework should be and how the burden of proof should be allocated in ODS decisions involving deepfakes.

Research Questions:

1. What is the current framework for ODS bodies to review deepfakes as laid down in EU legal framework AIA (/DSA) and platform T&C?
2. How do ODS bodies allocate and assess the burden of proof in disputes concerning platform moderation decisions relating to alleged deepfake content and what role do technological tools play in this?
3. To what extent can the transparency obligations applicable to ODS bodies under Article 24(5) DSA contribute to greater consistency, predictability, and coherence in deepfake-related moderation decisions?

2. Methods

2.1 Interviews

We conducted a series of semi-structured interviews with relevant stakeholders to explore potential responses by out-of-court dispute settlement (ODS) bodies to challenges arising from AI-generated content. Semi-structured interviews were selected due to their methodological flexibility and their capacity to generate nuanced insights into institutional practices, technical constraints, and normative considerations. The interview sample included four categories of stakeholders.

First, representatives of ODS bodies were consulted to examine existing approaches to the detection and assessment of AI-generated content in dispute settlement procedures (CEAH, Platform Control, ADR Center and ADROIT).

Second, we have tried to conduct interviews with representatives of technical solution providers, including DeepTrust, in order to assess the effectiveness, reliability, and limitations of content verification and deepfake detection systems in the context of dispute resolution. No interviews have so far been attained.

Third, we have tried to obtain interviews with platform representatives to obtain insight into internal moderation practices, detection tools, and review procedures. Their participation is particularly relevant given both their operational expertise and their financial responsibility for ODS proceedings under the Digital Services Act framework. However, none of the platform representatives have replied.

Fourth, we contacted platforms. The response to this request was limited to a written document drafted by Pinterest. This response only referred to relevant public sources concerning their policy.

Finally, representatives from the Directorate General CNECT, from the European Commission were interviewed to provide the institutional perspective on the design and implementation of art. 21 DSA. The interview provided in depth background information.

3. Legal framework: ODS bodies under the DSA

3.1 Introduction

Under Article 21 of the DSA, ODS bodies review whether platforms' content moderation decisions were justified. In many cases, this requires ODS bodies to assess whether specific content is AI-generated. This assessment can be difficult in practice.

Before examining the measures that ODS bodies may use to determine whether content is AI-generated, this brief first considers the definitions of deepfakes under the AI Act, the DSA, and the terms and conditions of platforms such as Instagram, TikTok, LinkedIn, Facebook, and Pinterest. The brief then explores the scope of ODS bodies' discretion under the DSA in determining whether content is AI-generated. It concludes by analysing practical examples, including enforcement decisions, the AI Act Guidelines.

3.2 Definitions of AI-generated content and deepfakes under the AI Act

The AI Act establishes the first comprehensive legal framework for artificial intelligence in the EU, focusing on the risks associated with AI systems and the measures required to mitigate those risks.[2] The framework seeks to promote safety, protect fundamental rights, and support innovation in AI across the EU. For ODS bodies assessing whether specific content is AI-generated or constitutes a deepfake, it is therefore important to understand the relevant legal definitions in the AI Act and how they apply in practice.

AI-generated content refers to content that is either generated or modified with the help of AI, which includes images, audio or videos.[3] While the AI Act does not have a legal definition of this concept, it does refer to this concept in its legal provisions. The AI Act regulates certain categories of AI-generated content, with a focus on transparency obligations for providers and deployers of AI systems.

Article 3 (60) of the AI Act does give a definition of the term ‘’deepfake”. This concerns ‘’ *AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful”*.

[2] European Commission comments on AI Act 2026
[3] European Parliament 2023

The term ‘’existing persons’’ may create interpretative difficulties. A strict literal interpretation could lead to the conclusion that characters generated entirely by AI (such as virtual influencers or fictional politicians) would not qualify as ‘deepfakes’ because they are not based on an ‘existing’ person.[4] This term could also be interpreted in a more teleological way. The teleological approach focuses on the objectives of the AI Act, namely protecting individuals from deception and misinformation. From this perspective, even fully synthetic images that resemble a human archetype could be classified as deepfakes, ensuring that they fall within the scope of the stricter transparency requirements.[5]

It’s important to realise the difference between AI-generated content and deepfakes. Deepfakes are always considered AI-generated content, but AI-generated content is not always a deepfake. Deepfakes are images, videos, and auto created and manipulated by AI to look and sound realistic.[6]

The AI Act’s definition of deepfakes does not play a decisive role in assessments of ODS bodies. Rather, ODS bodies focus on whether platforms have correctly applied their terms and conditions when moderating content.[7] As a result, platforms’ terms and conditions are of greater practical significance in determining whether content should remain online or be removed.[8]

3.3. Deepfake obligations under the AI Act

In addition to defining relevant concepts, the AI Act establishes several obligations concerning deepfakes. These obligations are designed to address risks associated with AI-generated or manipulated content and to mitigate their potential harms.

As noted above, a deepfake refers to AI-generated or manipulated image, audio, or video content that resembles existing persons, objects, places, entities, or events and would falsely appear to a person to be authentic or truthful. Such content can create significant risks, including deception, manipulation, and harm to individual rights. To address these risks, Article 50 of the AI Act establishes transparency obligations for providers and deployers of certain AI systems.

Under Article 50, providers and deployers may be required to disclose that content has been AI-generated or manipulated when it qualifies as a deepfake. This obligation seeks to ensure that users of online platforms are informed about the nature of the content they encounter and are less likely to

[4] Labuz 2025, p. 2.
[5] Labuz 2025
[6] AI safety hub 2025
[7] Pending final implementation of the AIA, this is subject to change.
[8] ADR Center interview

be deceived. The disclosure must be clear and distinguishable, enabling users to recognise AI-generated or manipulated content.

The implementation of Article 50 AI Act is further supported by the ongoing Code of Practice on the marking and labelling of AI-generated content. The Code is intended to assist providers and deployers of generative AI systems in complying with their transparency obligations under Article 50(2) and Article 50(4) AI Act. Its drafting is organised around two working groups: one focused on providers' obligations to mark AI-generated or manipulated outputs in a machine-readable and detectable format, and another focused on deployers' obligations to label deepfakes and certain AI-generated or manipulated text on matters of public interest. If approved by the Commission, the final Code will operate as a voluntary tool for demonstrating compliance, while Commission guidelines are expected to further clarify the scope of the legal obligations.

In practice, deepfakes may be difficult to detect because they often involve the manipulation of real persons, events, or content. They may also be unlabelled. This makes it difficult for both platforms and ODS bodies to assess whether specific content should be classified as a deepfake.

3.4 The DSA Framework for ODS Bodies

Under the DSA, ODS bodies provide users with an additional mechanism to challenge content moderation decisions made by online platforms.[9] Online platforms may make decisions on user generated content by restricting their accounts or their content. They have the authority or obligation when certain content breaches their terms and conditions or when it can be classified as illegal content. This also applies to deepfakes. Platforms have an obligation to provide an inclusive and safe environment for their users. When certain content is considered a deepfake, it may be harmful to people. However, sometimes these moderation decisions are unjustified. Users can contest these moderation decisions in three ways: the internal complaint-handling system of platforms, judicial redress, or out-of-court settlement bodies.[10]

Article 20 of the DSA requires online platforms to provide their users with an internal complaint-handling system through which they can contest certain content moderation decisions. Users may file complaints against decisions concerning the removal or restriction of content or accounts, and restrictions on monetisation. This system must be easily accessible and free of charge. Platforms are also required to assess these complaints in a timely, non-discriminatory way. Users must

[9] European Commission 2026
[10] Art 17(3) f DSA.

also be informed of the reasoning for a certain outcome of a complaint and the possibility to go to an out-of-court dispute settlement.

This brings us to Article 21 of the DSA, where ODS bodies play an important role. When faced with a moderation decision or outcome of a complaint by a platform, users are entitled to select an ODS Bodies to resolve any disputes relating to those decisions. These ODS bodies need to provide an accessible and independent alternative. Their role is to assess whether the moderation decisions given by the platforms comply with the DSA and the T&Cs. This way they can contribute to accountability and transparency in content moderation decisions made by an online platform. However, decisions by ODS bodies are not legally binding on platforms. While the ODS bodies play an important role in the context of online platforms and under the DSA, it can be difficult to assess whether certain content is AI-generated and can be specifically classified as a deepfake.

Article 21(3)(b) DSA requires ODS bodies to possess the necessary expertise in relation to the issues arising in one or more specific areas of illegal content, or in relation to the application and enforcement of the terms and conditions of one or more types of online platform, enabling the body to contribute effectively to the resolution of a dispute. In the context of deepfakes, this expertise may also depend on the technical tools available to ODS bodies for detecting or verifying whether content has been AI-generated.

As noted above, Article 21 of the DSA provides only limited guidance on the exact role of ODS bodies and the scope of their authority in detecting deepfakes. In principle, this leaves ODS bodies with considerable discretion to define the contours of their own role and approach. As a result, ODS bodies have different approaches available for cases involving deepfakes. The brief further explores technical tools that could help ODS bodies in assessing whether content is considered a deepfake and the role of these tools in disputes concerning suspected deepfakes.

3.5 Interplay DSA and AI Act

Detecting deepfakes can be challenging in practice. Both the DSA and the AI Act provide relevant regulatory frameworks for addressing this challenge, but they do so from different perspectives. The AI Act focuses on the risks, mitigation measures, and transparency obligations associated with AI-generated or manipulated content, including deepfakes. The DSA, by contrast, regulates how online platforms moderate content, structure their T&Cs, and provide redress mechanisms for disputes concerning content moderation decisions.

This interplay is particularly important for ODS bodies. On the one hand, they may need to assess whether the content at issue is AI-generated or qualifies as a deepfake. On the other hand, they must also consider the obligations that platforms have when moderating such content, including how those obligations are reflected in their terms and conditions.

Under the AI Act, providers and deployers of AI systems may be required to disclose that certain content has been AI-generated or manipulated. These transparency obligations aim to reduce the risk that users are deceived or misled. However, such obligations do not, by themselves, determine whether a platform was justified in removing, demoting, restricting, or otherwise moderating specific content. That question is primarily governed by the DSA, including its rules on content moderation, transparency, internal complaint-handling, and out-of-court dispute settlement.

For ODS bodies assessing whether a content moderation decision concerning deepfakes was justified, both regulatory frameworks are therefore relevant. However, the legal framework alone is not sufficient. In practice, platforms often base their moderation decisions concerning deepfakes and AI-generated content on their own T&Cs. ODS bodies must therefore consider not only the DSA and the AI Act, but also the platform-specific rules that formed the basis for the contested moderation decision

3.6 Platform Terms and Conditions

In practice, the terms and conditions of platforms are sometimes more important than the regulatory frameworks that have been mentioned before. ODS Bodies review content moderation decisions made by platforms. In practice, it appears that platforms often base these decisions on their own terms and conditions. It is important to see in what way the definition of deepfakes in the terms and conditions differs from the definition given in the AI Act and how this could influence ODS Bodies as they handle both (EU) law, domestic criminal law and platform definitions.

3.6.1 META

Meta does not explicitly define the concept of ‘’deepfakes’’. However, it falls within the scope of manipulated media. Under the broader category of misinformation, deepfakes can be classified as manipulated media when the following applies: ‘’when the content is a photorealistic image or video, or realistic-sounding audio, that was digitally created or altered and creates a particularly high risk of materially deceiving the public on a matter of public importance’’.[11] Meta also has a requirement that

[11] https://transparency.meta.com/policies/community-standards/misinformation/

people disclose whenever they post organic content with photorealistic video or realistic-sounding audio that was digitally created or altered. They will receive penalties if they fail to do so.[12]

3.6.2 TikTok

The community guidelines of TikTok provide several definitions relating to AI-generated content, such as deepfakes. First, TikTok defines *AI-generated content (AIGC)* as content, including images, video, or audio, that is created or modified by artificial intelligence (AI) technology or machine-learning processes. This content may include images of real people, and may show highly realistic-appearing scenes or use a particular artistic style, such as a painting, cartoons, or anime.[13] TikTok further defines *significantly edited content* as content that shows people doing something they did not do, saying something they did not say, or altering their appearance in a way that makes them difficult to recognize or identify. This includes applying certain face filters, or an animation of an individual.

Additionally, TikTok identifies types of manipulated content that could mislead users. TikTok defines *Misleading AIGC or edited media* as audio or visual content that has been edited, including by combining different clips together, to change the composition, sequencing, or timing in a way that alters the meaning of the content and could mislead viewers about the truth of real world events. Lastly, it defines *Realistic-appearing scenes or people* as content that uses are scenes, using images, video, or audio, that would lead someone to believe that the person shown is real or the event took place in the real world, such as a scene that is shown in the style or quality of a photograph or video.

3.6.3 LinkedIn

The terms and conditions of LinkedIn does not give a precise definition of the term deepfake. However, deepfakes may fall within LinkedIn's broader category of false and misleading content, for which the platform does provide specific guidance. LinkedIn removes specific claims that are presented as facts and are demonstrably false, clearly misleading and likely to be harmful. Further, they also remove content that has been synthetically created or manipulated to misrepresent or distort real-life events without a clear indication of the falsified or altered nature of the material. Content that is inaccurate or clearly misleading but unlikely to be harmful will not be distributed beyond the author's network. Posting inaccurate or misleading content can be a violation of LinkedIn's guidelines for the professional community.

[12] https://transparency.meta.com/policies/community-standards/misinformation/
[13] https://www.tiktok.com/safety/en/policies-and-engagement/integrity-authenticity?cgversion=2024H1update

3.6.4 **Pinterest**

Pinterest does not allow antagonistic, explicit, false or misleading, harmful, hateful, or violent content or behaviour. They may remove, limit or block the distribution of such content and the accounts, individuals, groups and domains that create or spread it based on how much harm it poses. These policies apply to everyone and to all types of content, including synthetically generated content. While Pinterest does not define deepfakes, it may fall within the category of misleading content.

3.6.5 Comparative analysis

A comparison of the T&Cs of Meta, TikTok, LinkedIn and Pinterest shows that the platforms take different approaches in defining deepfakes and AI-generated content. While TikTok provides comprehensive and clear definitions of what it considers AI-generated content, Meta, LinkedIn and Pinterest rely on broader categories of false and misleading content, under which deepfakes may fall, when they involve realistic representations that mislead users.

3.7 Practical examples

The following section examines practical examples to demonstrate how the legal framework may operate in practice.

3.7.1 EU Commission: enforcement decisions

Under Article 20 of the DSA, we have seen that users can contest certain content moderation decisions through an internal complaint-handling system. Platforms have the obligation to make this mechanism effective and also give users the ability to notify illegal content. However, in 2025 the EU Commission preliminarily found Meta, both Instagram and Facebook, in breach of its obligations to provide users simple mechanisms to notify illegal content, as well as to allow them to effectively challenge content moderation decisions.[14] The internal complaint-handling system does not appear to allow users to provide explanations or supporting evidence to substantiate their appeal.

This also makes it difficult for ODS bodies to assess whether specific content is considered a deepfake. Further difficulty arises due to the rather sparse Statement of Reasons given by the platforms. When users can not provide any evidence supporting their appeal, ODS bodies have to look at different approaches in assessing deepfakes.

3.7.2 Guidelines of the AI Act

Although the guidelines of the AI Act are not binding, it can provide a useful insight into the transparency obligations of the AI Act, as mentioned before. The Draft Guidelines explain that

[14] https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503

deepfakes are content that closely resembles real persons, objects, places, entities, or events, without necessarily being identical. The content must be realistically conceivable and capable of existing in reality, while also creating a false impression of authenticity or truthfulness, in a way that it may be misleading.[15]

Recently the EU commission published its Code of Practice on Transparency of AI-Generated Content. In a world where law equals compliance, AI system providers and deployers ascertain the deepfakes are marked and labeled.[16] However, user uploaded content does not fall within the responsibility of the deployer. Therefore platforms could still host unlabeled deepfakes. Furthermore, if platforms were to actively label uploaded content as deepfakes, there will be a margin of error concerning the labeling. This entails that disputes over the presence of deepfake content will still arise, even under the new AI Act and the Code of Practice on Transparency of AI-Generated Content.

3.8 Intermediary Concluding remarks on the legal framework

In conclusion, the legal framework of deepfake consists of EU regulations, such as the definitions provided in the AI Act and the obligations provided by the DSA. The definition of deepfakes that the AI Act provides, is not of great importance to ODS bodies. ODS bodies assess whether platforms comply with their own T&Cs, which makes the definitions given by these T&Cs far more important.

[15] https://www.twobirds.com/en/insights/2026/uk/when-marketing-meets-genai-update--the-eu-ai-acts-draft-deep-fake-guidelines-and-what-they-mean-for

[16] Code of Practice on Transparency of AI-Generated Content, https://ec.europa.eu/newsroom/dae/redirection/document/129555

4. Shifting the burden of proof in ODS cases.

A potential solution to the issue of recognising deepfakes for ODS bodies is shifting the burden of proof to the platforms. Users will not have a strong position to prove the use of deepfakes. They likely lack technical and financial capacity to assess whether a deepfake is used. Users will not often be confronted with the need to prove the presence of deepfakes. On the other hand, platforms are repeat players in assessing the use of deepfakes. Further, they will have financial and technical means to assess the presence of a deepfake. Therefore, platforms have a better position to prove the presence of deepfakes than users. Additionally, platforms will be able to retrieve the content that is removed by them for further review. Users might have a backup of the content that they posted, but do not necessarily have such a backup. This complicates their position to assess the content on presence of deepfakes.

Depending on the procedural posture of a case the burden of proof is assigned to the user or the platform. There are two procedural postures of cases at ODS bodies:

1. A user reports content for lacking an AI-generated content label.
2. A platform removes or restricts content on the basis that it is a deepfake.
    a. The presence of a deepfake can be the reason content is deleted
    b. The presence of a deepfake can be a factor in deciding illegality or infringement of terms and conditions.

In the first procedural posture the user makes a complaint to the platform. If the platform does not remove the content based on the complaint the user can appeal this decision at an ODS body. E.g. cases where the user's image is used for deep fake material and initiative is taken by the platform to remove the content.

In the second procedural posture the platform the content of the user is removed and the platform has to give a statement of reasons pursuant article 17 DSA. This statement of reason often is lacking (See section 5.1.2). The user can launch an appeal at an ODS body. E.g. cases where the user posts an real image, but the platform removes the content stating it is a deepfake.
In both procedural postures the user has the initiative of launching the procedure. This is reflected in the Rules of Procedure of User Rights article III §3 (4).

User Rights could be inspired by national procedural law regarding evidence. The principle of burden of proof in the Netherlands, France, Italy and Germany is that the one who claims something has to prove it (See table 1, page 14). Therefore in the first procedural posture the burden of proof is on the user, because they claim that certain content is a deep fake. In the second procedural posture the

burden of proof is on the platform. This is affirmed by art. 17 DSA, where the platform is obliged to give a statement of reasons.

Jurisdiction	Law	Text	English Translation
Netherlands	Artikel 150 Wetboek van Burgerlijke Rechtsvordering	De partij die zich beroept op rechtsgevolgen van door haar gestelde feiten of rechten, draagt de bewijslast van die feiten of rechten, tenzij uit enige bijzondere regel of uit de eisen van redelijkheid en billijkheid een andere verdeling van de bewijslast voortvloeit.[17]	The party who relies on the legal consequences of facts or rights asserted by that party bears the burden of proving those facts or rights, unless a different allocation of the burden of proof follows from a specific rule or from the requirements of reasonableness and fairness.
France	Article 9 Code de procédure civile	Il incombe à chaque partie de prouver conformément à la loi les faits nécessaires au succès de sa prétention.[18]	It is for each party to prove, in accordance with the law, the facts necessary for the success of its claim.
Italy	Articolo 2697 del Codice Civile	Chi vuol far valere un diritto in giudizio deve provare i fatti che ne costituiscono il fondamento. Chi eccepisce l'inefficacia di tali fatti ovvero eccepisce che il diritto si è modificato o estinto deve provare i fatti su cui l'eccezione si fonda.[19]	A person who seeks to assert a right in court must prove the facts that constitute the basis of that right. A person who alleges that those facts are ineffective, or that the right has been modified or extinguished, must prove the facts on which that objection is based.
German	Rosenbergsche Formel	This is partly applied in art. DSA 17 for cases with the second procedural posture, but cases with the first procedural posture find no such provision in formal law.	

Table 1

In the Rules of Procedure of User Rights article III §3 (5), the parties (the user and the platform) are both allowed to submit evidence. The rules of procedure do not specify who is responsible for proof of certain facts.

"The parties are given the right to be heard and the opportunity to comment. They may submit facts and assessments insofar as these are relevant."

Rules of Procedure of User Rights article III §3 (5)

[17] Article 1:150 Dutch Civil Code of Procedure (DCCP) (*Wetboek van Burgerlijke Rechtvoering*)
[18] Article 9 French Code of procedure (*Code de procédure legislative)*
[19] Article 2697 Italian civil code (*Codice Civile*)

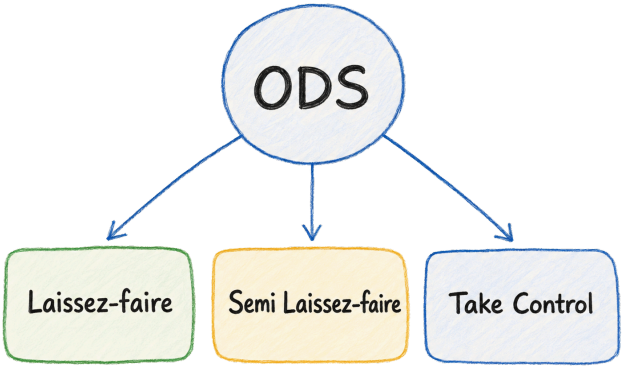
Following the logic from the national civil codes the dynamic of burden of proof in the two procedural postures can be distributed between the parties. In the first procedural posture where the user makes a complaint, the burden of proof is on the user. For the second procedural posture where the platform removes content, the burden of proof is on the platform.

Considering the position of the parties to gather evidence about the presence of deepfakes it can be suggested platforms are capable of bearing the burden of proof. Nonetheless, users might not be able to bear the burden of proof. In cases where users have to bear the burden of proof ODS bodies could shift the burden of proof to the platform. This could be based on their superior position to bear the burden of proof. The first procedural posture could be reframed as a flagging by the user, rather than the start of a procedure. If the user merely flags the content, then the decision by the platform that content is deepfake is the initial claim. Then the burden of proof that the content is not deepfake material is on the platform. Although this is an attractive approach to shift the burden of proof to the platforms there are some caveats to the fairness of this solution. The platform has no active duty to monitor the content on their website, pursuant Article 8 DSA. And systematically shifting the burden of proof to platforms may place them at a structural disadvantage, particularly in technically ambiguous or evidentially uncertain cases. Therefore shifting the burden of proof in these cases can be considered.

The technical solutions will have a margin of error (See 5.3.3). The party bearing the burden of proof will be disadvantaged due to this margin of uncertainty. Unprovable cases will be decided against the party bearing the burden of proof. In difficult-to-determine cases, proving whether content constitutes a deepfake may be excessively costly, technically infeasible, or unreliable. Hence, a de facto presumption against platforms could therefore emerge, whereby disputed content remains online whenever platforms cannot conclusively demonstrate manipulation. Such an outcome risks undermining the integrity of ODS review by predetermining outcomes through procedural allocation rather than substantive assessment of the facts.
Because the shifting of the burden of proof can systematically disadvantage platforms and undermine the position of ODS bodies, different approaches to dealing with assessment of deepfakes can be taken. These approaches are explained in the next chapter.

5. Dealing with deepfakes, different approaches

Three approaches for ODS bodies to deal with cases concerning deepfakes can be distinguished. (1) Laissez-faire, (2) Semi-Laissez-faire and (3) Take-control. The first approach strictly looks at the evidence presented by the parties (user and platform). The second approach looks at the evidence presented by the parties (user and platform), but does allow limited analysis by the ODS body for affirmation or in extreme cases. E.g. the user submitted evidence discharging the illegality, and the platform only provided a limited statement of reasons, but from the context it is likely the content was illegal. The third approach leaves the realm of laissez-faire. In the third approach we suggest that the ODS body takes control over the situation by acquiring the expertise to conduct the assessment of deepfakes in-house and produce their own evidence.



5.1 Laissez-faire Approach

5.1.1 The laissez faire approach and consequences for burden of proof and evidence

Under the laissez-faire approach, the ODS body assumes a purely passive role in assessing deepfake content. The ODS body neither investigates independently nor supplements the evidence placed before it, it relies entirely on what the user and the platform choose to admit. There is therefore no use of technical tools in the laissez faire approach. The practical consequence is that the quality of the outcome is entirely determined by the information asymmetry between the two parties. Users, who are generally the victims of deepfake content, must raise grounds on which they consider the content unlawful and supply the content itself.[20] However, they are the party with the least resources to do so. As noted in the interviews with ODS bodies, complainants often forget to take sufficient evidence (i.e. they provide a link, but the content is removed and they have no screenshots). In principle, the approach places a corrective burden on the platform that it should provide additional information as to why the content was (or was not) taken down. However, in practice we have seen that attempts (from ODS bodies) to seek further clarification have not really helped.

[20] Dit komt ook uit ADR interview maar quote is hier niet echt gepast

The interviews illustrate why a purely laissez-faire approach is problematic in practice, as platforms are often unwilling to share detailed reasoning behind AI-generated content moderation. ADR Center explains that the efficient functioning of ODS bodies depends heavily on platforms sharing the content and the underlying assessment that led to the moderation decision. Without that information, the ODS body cannot meaningfully review the case and may be forced to issue a default decision *against* the platform. However, platforms may be limited in disclosing their internal assessment methods, as doing so might reveal trade secrets or sensitive enforcement methods. Forcing platforms to share these poses a risk of user workarounds. ADROIT recognises this, but argues that the platform should still provide a minimum explanation sufficient for users and ODS bodies to understand why a decision was taken.

5.1.2 Statement of reasons and the DSA transparency database as additional tools for ODS bodies

One DSA mechanism that could help ODS bodies gain additional insight into platforms' deepfake moderation practices is the statement-of-reasons requirement. In Article 17, the DSA states that a *'clear and specific statement of reasons'* must be given by the platform when moderating. Article 24(5) DSA adds a further transparency obligation, requiring providers of online platforms to submit these statements of reasons to the Commission 'without undue delay' for inclusion in a publicly accessible, machine-readable database. These decisions are registered in the DSA transparency database, which is intended to support transparency and public oversight of platform moderation.

An effective repository of statements of reasons could assist ODS bodies in several ways.[21] First, it may provide insight into how platforms assess AI-generated content, including whether automated detection tools were used and what type of reasoning supports the finding that content is AI-generated. Second, to what extent the content is deemed in violation of the platforms' terms and conditions and/or the law. Third, it can help compare the disputed case with the platform's general practice and thus assess the consistency to which AI-generated content is moderated.

This approach therefore examines whether the current mechanism for providing transparency in platform moderation decisions is sufficiently effective to support ODS bodies in disputes concerning deepfake moderation.

The database was created to: 'enable more transparency and scrutiny over the content moderation decisions taken by providers of online platforms, and to better monitor the spread of illegal and

[21] Even if published Statements of Reasons are anonymised, they can serve as 'case law' for subsequent cases, given the circumstances of the moderation decision are adequately described.

harmful content online'.[22] To achieve this goal, the DSA Transparency database requires the registration of a number of 'attributes' of a moderation decision. First, a platform-unique identifier, which is a string that identifies a statement of reasons within the online platforms' systems.[23] Second, the specifications of the content affected by the decisions. The content type, creation date, language, and a language identifier must be registered. Third, the type of restriction, including visibility, monetary, provisioning of the service, and service's account (art. 17(1)(a)-(d) DSA). Fourth, the duration of the restriction. Then, most importantly for deepfake moderation, information on the facts and circumstances relied on in taking the decision and information on the use made of automated means. Finally, the legal or contractual grounds relied on in taking the decision. The DSA transparency database also provides a list of categories to codify the types of illegally/contractual grounds.

AI-generated content can span across these categories. For example, it could fall under the ground of 'negative effects on civic discourse or election', however also under 'cyber violence against women', or 'misleading information about the characteristics of the goods and services'. Not including a distinct classification mechanism for AI-generated content makes it difficult to determine whether a moderation decision is based on illegal/noncompliant AI use. For example, when searching with the search term 'deepfake', no results are shown. When searching with the search term 'Artificial intelligence', a very large quantity of results are shown, however only by TikTok and VSCO. These are the most prevalent statements:

TikTok

> *"We welcome creativity, including when it comes from new digital tools like generative artificial intelligence (AI). But AI and editing can blur the line between fact and fiction. To help keep content on TikTok trustworthy and provide people with important context about what they are viewing, we require creators to label AI-generated or significantly edited content that shows realistic-looking scenes or people. Unlabeled content may be removed, restricted, or labeled by our team, depending on the harm it could cause. Even with labels, some edited or AI-generated content can still be harmful. We don't allow content that's misleading about matters of public importance or harmful to individuals.We proactively enforce our Community Guidelines through a mix of technology and human moderation. We have detected this policy violation using automated measures."*[24]

[22] DSA Transparency Database, European Commission, 2024
[23] https://transparency.dsa.ec.europa.eu/page/documentation
[24] https://transparency.dsa.ec.europa.eu/statement

VSCO

> *"Be honest in your conduct and interactions on VSCO. We do not allow the following content or behaviors: Attempting to deceive, manipulate, or take advantage of others on VSCO; Non-genuine solicitation of services of or from others on VSCO; or Using artificial intelligence to facilitate any of the above"*[25]

The TikTok statement is more AI-use specific as it states that realistic AI-generated content may blur the line between fact and fiction. Still, it remains generic. It does not specify what exact part of the content triggered moderation, whether the content was labeled, or why the imposed restriction was necessary. The VSCO statement is even less revealing, as it covers an even broader set of situations.

A few questions arise from this. Firstly, is this transparency database working properly to ensure its efficacy to achieve its goals, as only VSCO and TikTok platforms specify explicitly in their statement of reasons that moderation has occurred on the basis of AI-use? Second, do these statements of reasons satisfy the *'clear and specific statement of reasons'* requirement of Article 17 DSA?

5.1.3 Limits of the DSA Transparency Database

While the Database is described as a 'historical achievement' in platform governance[26], some authors critique its overall effectiveness at achieving transparency. A first apparent issue is the high level of discretion afforded to platforms when submitting their statement of reasons. Many 'attributes' of the statement are labeled as optional, conditionally mandatory or unclear, which permits platforms to be selective about the depth of the information provided.[27]

Next, there exists a ' power and technology gap' between public sector enforcement and VLOPs. This influences the requests from DSA enforcement and makes it difficult to assess compliance with the DSA.[28] Also, as seen by the VSCO and TikTok examples, although platforms are required to provide specific fact and circumstances for their decisions, most of the use automatically generated, generic templates.[29] In this empirical analysis, it was found that for example, Google Shopping used a standard template in 99.5% and TikTok a short standard template in 98.5% of its statements of reasons. Inconsistencies between platforms' transparency reports and the statement of reasons

[25] https://transparency.dsa.ec.europa.eu/statement
[26] Kaushal et al. 2025 p. 1121
[27] Kaushal et al. 2025 p. 1123
[28] Kaushal et al. 2025 p. 1130
[29] Kaushal et al. 2025 p. 1125

included in the DSA transparency database further raise concerns that platforms are underreporting their decisions to be strategically compliant with the DSA rather than being genuinely transparent.[30]

The issues with the current setup of the transparency database can be exacerbated when the contested content is AI-generated. The classification mechanism makes it difficult to identify AI-generated content moderation decisions, as it does not provide a separate 'AI-generated content classification'. Understandably, this cannot be in the same classification categories as content based on its legality/noncompliance, as AI-generated content is not *per se* unallowed. Instead, a separate statement of reasons attribute, similar to the automated moderation one, is desirable as it makes the database more useful for ODS bodies. It can inform assessment of decisions and contribute to consistency between ODS bodies. Furthermore, AI-generated content moderation often relies on automated detection tools. It is required in the transparency database to state the usage of automated means. However, the mere statement of use is insufficient for ODS bodies to derive meaningful information. ODS bodies would benefit from receiving more detailed information on their implementation, for example, what evidence is sufficient to assess content as AI-generated and if there was human oversight. Lastly, the underreporting by the platforms could be addressed through enforcement by the EU Commission. In the Commission's current enforcement activities database, no enforcement action has been taken due to insufficient disclosure under the Statement of Reasons requirement of Aricle 17 DSA.[31] A more active enforcement could benefit the efficacy of DSA Transparency Database to inform ODS Bodies on the mechanisms used by platforms in deepfake moderation.

5.1.4 Subconclusion

The problem of the laissez-faire approach is demonstrated by the problems of the transparency base. The laissez faire approach rests on the assumption that the parties (user as well as platform) will supply sufficient evidence for the ODS bodies to make a meaningful decision. The transparency database was intended to support this kind of informed review, by making platform moderation more comprehensible to outsiders. Yet in reality, platforms systematically exploit the discretion afforded to them under the DSA framework, submitting general statements that reveal little about the actual reasoning. This suggests that many Statement of Reasons do not comply with the requirements of Article 17 DSA. Platforms hold the relevant knowledge and face little pressure to share it. Without either structural reform of the database, enforcement of Article 17 DSA, or a more active role for the ODS body itself, the laissez-faire approach risks producing decisions that are arbitrary rather than a genuine assessment of the deepfake content.

[30] Johnson & Roy 2025 p. 13
[31] Supervision of the designated very large online platforms and search engines under DSA | Shaping Eu…

5.1.5 Recommendations for the Commission:

- Include a separate 'AI-generated content' attribute as an additional filtering function in the DSA Transparency Database.
- More active enforcement on the basis platforms' non-compliance with art. 17 DSA.

5.2 Semi-laissez-faire Approach

5.2.1 Definition of the semi-laissez-faire approach

Another approach is the semi-laissez-faire approach. This approach has a predominantly passive role for the ODS. The scope of the reviewed materials is determined by user and platform submissions. If the correct decision cannot be made based on the materials submitted by the parties, ODS bodies could investigate by asking for specific documents from the parties and doing a limited analysis of the materials.

5.2.1.1 Balance of evidence

Currently users are bearing the burden of proof for deepfake content flagged by them. In most cases, the platforms give very limited statements of reasons. The users usually bring forward a lot of evidence to support their claim. If ODS bodies only base themselves on the evidence submitted to them, the user could see their position affirmed more often. The ODS still would have to assess their evidence. So, in cases where the moderation decision by the platform was evidently right, even the evidence submitted by the user will also point in that direction. Still, this approach poses some risks. The user might not have the means or expertise to generate evidence, such as detection software. And the content that is deleted, might not be backed up by the user. The platform could be better suited to provide the evidence.

5.2.1.2 Restrictive perception on the role of ODS bodies

Talking to ODS bodies, the perception of their role was restricted to reviewing the platform's decision. *"[Our role] is not to independently perform some kind of technical forensic analysis of every piece of content. Instead, our role is to review the platform's decision."*[32] This restrictive perception is uncontroversial. Platforms will agree with this approach because the conflict is limited to the parties' submissions. However, such an approach does not always benefit the user. The user will still be confronted with the task of generating evidence, which can be costly and technically difficult. Recital 58 of the DSA, emphasizes the need for the effectiveness of ODS bodies and their responsibility vis à vis users. *"Recipients of the service should be able to easily and effectively contest certain decisions."*[33] The alignment between the restrictive view and recital 58 can be questioned.

5.2.2 Limited investigative actions

One problem with relying solely on evidence submitted by the parties is that only the most serious cases will receive procedural support. In reality the burden of producing adequate evidence falls on the user. Victims of deepfakes often have trouble producing the evidence required to substantiate a

[32] This statement was made by one of the interviewed ODS bodies.
[33] Ibid.

claim.[34] ODS bodies confirm that they frequently provide links, but forget to take screenshots for examples, which becomes problematic when the content is suddenly removed. The platforms however are reluctant to step in and provide reasoning as to why the content is or is not removed. This creates a power imbalance between the user and the platform. Because of the lack of technical capabilities of the ODS bodies themselves, they will only use technical tools to verify when it concerns a human assumption i.e. a case that is clearly illegal. The problem with deepfakes, however, is that we are often unable to tell whether they are fake. Therefore, cases that are not illegal, but can still be harmful to the victim, will systematically fall through the system. This is a group that cannot rely on support from those ODS bodies.

If the case concerns more serious facts, the parties' interest in having the content removed is stronger. If limited evidence is submitted by either of the parties it would be desirable to investigate the submission and request more evidence. Transparency on the use of these investigative powers would be beneficial to the ODS framework. To users it will be clear what is expected from them in terms of evidence. Under the semi-laissez-faire approach, there still is prejudice against less severe cases. People flagging less severe, but still incompliant uses of deepfakes will not receive an effective remedy. They will still bear the full burden of proof without guidance by the ODS body. This systemically disadvantages their position.

5.2.2.1 Requests for evidence

One of the ways ODS bodies could improve their giving effectiveness to all complainants is being upfront about the expectancy on the burden of proof. ODS bodies should give ex ante clarity on the types of cases and the corresponding required evidence. In addition, ODS bodies should publish for which cases they will request additional evidence and for which cases they will not request additional evidence. Being transparent about the required evidence will allow ODS bodies to apply the restrictive perception of their task. Substantiating and clarifying how they handle evidence and the lack thereof will contribute to the effectiveness of their services. Uncertainty about relevant evidence can lead to overburdening the users with the task to provide evidence. If there is ex ante clarity on the necessary evidence this overburdening can be counteracted.

5.2.2.2 The use of technical tools in the semi-laissez-faire approach:

In the semi-laissez-faire approach, the technological tools enter the process. However, the tools have a confirmatory rather than an investigative role. This is usually done by inhouse tools to confirm whether a hunch, that something is a deepfake, is correct. So the tools are applied as a confirmatory measure to ensure the quality of the decision. Not to conduct an analysis of the presence of deepfakes.

[34] Birrer & Just 2025

In the take-control approach ODS bodies are advised to lean into the use of technical detection tools and integrate them into their expertise.

This confirmatory function by AI tools is often used in practice. If the ODS body, based on the submissions before it, has a strong suspicion that the content is AI-generated, a limited technical check can confirm that suspicion without requiring the ODS body to conduct a full forensic investigation. However, it is important to stress that the tools used in the semi-laissez-faire approach are typically in-house and of limited analytical depth. They are not designed to produce independent expert findings that could withstand scrutiny, nor are they applied systematically to all cases. The tool is more a quality-control mechanism rather than one that produces evidence.

5.2.3 Sub conclusion

The semi-laissez-faire approach offers a middle ground between the laissez-faire approach and the take-control approach. The confirmatory use of technical tools provides a degree of quality assurance in evident cases. Nevertheless, the approach has structural limitations. The burden of proof remains primarily with the user, while this is precisely the party with the least resources and expertise. The consequence is that only the most serious and evidently illegal cases receive procedural support, while deepfakes that are harmful but not evidently illegal (and where the distinction between real and fake is already difficult to establish) systematically fall through the cracks.

5.2.4 Recommendations if this approach to handling deepfakes is selected:

- Provide information on the expected evidence to users.
- Set up guidelines to determine thresholds for evidence. If the threshold is not met, the investigative actions should be invoked

5.3 Take-control Approach

5.3.1 Definition of the take-control approach

Another approach is one in which the ODS body 'takes control' of the evidence. The ODS body can conduct its own analysis of the presence of deepfakes. By taking this analysis upon themselves, the user will not be required to provide all the evidence. This makes the process for them easier, because they would only have to report their case to the ODS body. To take control over the evidence gathering the ODS body gains a proactive role. To achieve this, the ODS body will need to rely on inhouse experts and/or AI-fueled tools to detect whether something is a deepfake or not. Based on this technical assessment, the decision would be made. ADR develops on this: "it could be like turning into a little criminal laboratory that dissects the content".

An advantage of this approach is that the relationship between the platform and the user becomes more equal. The ODS body makes a decision apart from the (lack of) evidence submitted by both parties. An investigative standard will ensure consistency in decision-making. Additionally, the asymmetry between the user and the platform is reduced. The possibility to investigate and produce their own evidence, ensures ODS bodies are not as structurally dependent on the platforms to produce evidence. A foreseeable consequence is a change in incentive; non-cooperation now carries a risk for platforms that ODS bodies will reach their own adverse findings. Platforms will thus have more reason to engage substantively, instead of winning by default. If the information asymmetry is lessened, the barrier as a victim of deepfakes to go to an ODS body is lowered. The burden of proof is now with the other actors rather than with the victim.

5.3.2 Certification for area of expertise

The responsibility for assessing the presence of deepfakes can be assigned to ODS bodies. ODS bodies could interpret their role restrictively, as mentioned under the semi-laissez-faire approach. This approach is not necessary. ODS bodies can assess evidence based on the DSA. By including the detection of deepfakes in their areas of expertise, ODS bodies can form the legal foundation to conduct their assessment. There is little further clarification of how ODS bodies should behave according to Article 21 DSA. This leeway could be an opportunity to take this space and mark their territory as a serious body. Lowering the barriers for victims to go to an ODS body would be consistent with the objective to provide accessible out-of-court redress as stated in Article 21 DSA.

5.3.3 The use of technical tools in decision making

In the take-control approach technology becomes the primary investigative instrument and this is not without risk. In reality having a proactive role as an ODS body in the detection of deepfakes means

needing in-house expertise and relying on a third party to supply an AI system that is able to detect deepfakes.

The market in detection tools is still in early stages, but the number of companies offering such services is growing.[35] More and more companies develop deepfake detection tools that recognize artificially created or manipulated content that has been generated through the use of AI. There are several methods to detect AI such as, but not limited to; Convolutional Neural Networks (this method detects pixel level anomalies), Optical Flow Analysis (detects motion inconsistencies between frames) and Facial Behaviour Analysis (analyses unnatural facial movements).[36] Nonetheless, the detection of deepfakes is, and will be, the result of a statistical process.[37] It is inherent to these tools that there will be a margin of error in this process.[38] For most commercial tools on the market there exists a significant performance gap between the laboratory testing and the real-world employment. Detection systems learn to recognize the specific pattern of known generation methods and when confronted with a technique they have not encountered before, their outcome pretty much becomes guesses.[39] Deepfake content often passes through multiple platforms (such as messaging services, social media and re-uploads), causing the pixel level to degrade, and making deepfake detection more difficult. This margin of error implicitly gives a normative judgement on how acceptable it is to wrongly classify a video as deepfake. This error consists of two types of mistakes: false positives and false negatives. False positives are cases where the detection software classifies a video as deepfake, that is in fact not a deepfake. False negatives are cases where the detection software does not classify a video as deepfake, that in fact is a deepfake.[40] This problem is inherent to the burden of proof. By deciding on the level of certainty that is required to prove a fact. However, ultimately, the lawmaker has authority to make this decision. The margin of error is a harsh reality confirmed by other ODS bodies, that confirm ‘they never fully trust AI’.

Another risk identified by the ADR center is the sensitivity of the data that is put into the tools. The most capable deepfake detection tools available are third-party products and using them would mean feeding them with; documents about the complainant, facial images and other potentially sensitive images about which the ODS bodies have no control once put into the system. That could undermine the trust that users place in the ODS bodies. The risks would need to be managed with proper safeguards.

[35] PUBLIC 2026
[36] Sunil et al 2025
[37] Sunil et al. 2025
[38] Yiğitalp Tulga 2026
[39] EDPS 2026
[40] Kumar et al. 2025

5.3.3.1 The costs of these tools

As a consequence of the implementation of these technical tools and hiring inhouse experts on them, the costs of the ODS body will increase. Deepfake detection is costly, especially if you want to minimize the margin of error and mitigate as many risks as possible. Therefore the foreseeable consequence would be to increase the prices of the service. The costs of the service must remain reasonable, otherwise it would stand in the way of the accessibility of the ODS body.[41] Pursuant Article 21(5) DSA, these costs can be charged to the platform. However, they can only be charged if the dispute is ruled in favor of the individual or the entity that submits the notice. These costs are thus charged on a case by case basis, and are less reliable as a consistent stream of revenue. The adjudication of ODS bodies have to be independent, so ex ante there is large uncertainty on their revenue.[42] As ODS bodies are not allowed to make profits on their adjudication it is difficult to make large upfront investments without outside funding. This outside funding could be provided by EU Member States, platforms wishing to reinforce the position of ODS bodies, or other benefactors.

5.3.4 Sub conclusion

Advantages of this approach are that the relationship between the platform and the user becomes more equal. The ODS body makes a decision apart from the (lack of) evidence submitted by both parties. And the asymmetry between the user and the platform is reduced. And the ODS bodies are not as structurally dependent on the platforms to produce evidence. The accumulation of all the risks and responsibilities that the ODS bodies have to take into account, put a substantive burden on the ODS bodies. Placing so much trust on a system to investigate and make decisions, without proper safeguarding could hurt the ODS bodies image. Whether these risks and responsibilities weigh up to the benefits is something the ODS bodies have to consider.

5.3.5 Recommendations take-control approach

- Get an inhouse AI deepfake detection expert and system and train the staff how to use this system and the risks of the system.
- Set up mechanisms that deal with the margin of error.
- Make sure the system is GDPR compliant and the sensitive data of the complainants is stored locally.

[41] Article 21(3)(e) DSA
[42] Pursuant to Article 21(3)(a) DSA

6. Conclusion

Deepfakes present a difficult challenge for ODS bodies under Article 21 DSA. Their assessment often requires ODS bodies to determine whether content is AI-generated or manipulated, while available legal, procedural and technical frameworks provide insufficient guidance. The AI Act provides a definition of deepfakes and introduces transparency obligations, while the DSA regulated platform moderation, statements of reasons and access to out-of-court dispute settlement. In practice, platforms mostly rely on their own definitions in their own terms and conditions, which do not always align with the AI Act definition. This creates uncertainty for users, platforms and ODS bodies.

A central difficulty is the allocation of the burden of proof. Users often lack technical capacity, financial means and access to evidence needed to prove that content is a deepfake. In contrast, platforms have more technical expertise, access to moderation systems and control over removed content. This makes it tempting to shift the burden of proof towards platforms. However, this causes systematic imbalance in outcomes in technically uncertain cases. The issue should therefore not be resolved by shifting the burden of proof to the platforms consistently.

This report has identified three alternative approaches for ODS bodies to deal with deepfakes. First, a fully laissez-faire approach, where ODS bodies rely only on evidence submitted by the user and the platform, without conducting their own technical investigation and relying on deepfake detection tools. Users may have limited capacity to produce evidence, while platforms often choose to underreport the reasons for their moderation decision. Although the DSA statement of reasons requirement and the DSA transparency database could help address this imbalance, their current usefulness remains limited because platform decisions are often generic and the database lacks an AI-generated content filtering option. Therefore, the laissez-faire approach may be appropriate only in straightforward cases where submitted evidence is sufficient, but cannot serve as a complete solution for deepfake moderation disputes.

Second, a semi-laissez-faire approach, which maintains a predominantly passive role for ODS bodies. This approach relies primarily on the submissions of users and platforms, while allowing for limited investigative action where the parties' submissions are insufficient. Technology enters the process but in a confirmatory rather than investigative role. Only where the circumstances already suggest AI involvement, ODS bodies can use their own trained AI to recognize if it was another AI-generated image. The purpose is "to confirm the assumption," not to generate a finding from scratch. While this approach offers a pragmatic middle ground between full passivity and active control, it remains

constrained by information asymmetries and the limited evidentiary capacity of users, falling short in cases where neither parties can provide adequate evidence.

Third, a take-control-approach. Here, the ODS body takes an active role in gathering evidence and the technology becomes the primary investigative instrument. This means investing in AI-fueled technology and inhouse expertise to detect deepfakes. There are some considerable risks and downsides to this approach. Such as, the inherent margin of error and risks of losing data-security. Mechanisms to manage those risks should be put in place, before the ODS body would become reliant on a machine for the outcome of its decision. However, it could considerably change the power structures. Now the outcome of the decision is for a large part based on the availability of evidence and the willingness of the platform to further cooperate.

The specifics of the role of the ODS body under the DSA remain uncertain. The ODS bodies can actively think and consider this uncertainty. They can decide to take a more passive role, they let all parties submit their evidence and judge based on that to the best of their abilities. However, due to the uncertainty surrounding the role of the ODS body, they also have the opportunity to push back against the platform and produce their own evidence. This would take a considerable amount of investing (as well as in money as in time). All approaches have both their opportunities as well as risks, what to do is a decision the ODS bodies must make themselves based upon their own experiences and expertise and national regulation.

7. Bibliography

7.1 Legislature:

Art 9 Code de procédure civile (French Code of Civil Procedure) (in force since 1 January 1976) https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006410102 accessed 11 June 2026.

Art 150 Wetboek van Burgerlijke Rechtsvordering (Dutch Code of Civil Procedure) (version valid 1 January 2026) https://wetten.overheid.nl/jci1.3:c:BWBR0001827&boek=Eerste&titeldeel=Tweede&afdeling=Negende&paragraaf=1&artikel=150&z=2026-01-01&g=2026-01-01 accessed 11 June 2026.

Art 2697 Codice Civile (Italian Civil Code) (Regio Decreto 16 marzo 1942, n. 262) https://www.codice-civile-online.it/codice-civile/articolo-2697-del-codice-civile accessed 11 June 2026.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277/1.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts [2024] OJ L 2024/1689.

7.2 Institutional Documents:

DSA Transparency Database, European Commission 2024

DSA Transparency Database : Questions and Answers. (2024). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/faqs/dsa-transparency-database-questions-and-answers

EDPS 2026

European Data Protection Supervisor. Deepfake detection. 2026. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en

European Commission 2026

European Commission, 'Out-of-Court Dispute Settlement Bodies under the Digital Services Act (DSA)' (Shaping Europe's Digital Future, last updated 19 May 2026) https://digital-strategy.ec.europa.eu/en/policies/dsa-out-court-dispute-settlement accessed 11 June 2026.

European Commission comments on AI Act 2026

European Commission AI Act. (2026, 10 juni). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Parliament 2023

EU AI Act: first regulation on artificial intelligence | Topics | European Parliament. (2023, 6 augustus). Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

7.3 Academic resources:

AI Safety Hub 2025

AI Safety Hub, 'Deepfakes & AI-Generated Content' (AI Safety Hub, 2025) https://aisafetyhub.au/guides/deepfakes-ai-generated-content accessed 11 June 2026.

Johnson & Roy 2025

Johnson, A., & Roy, P., (2025). EU Should Improve Transparency in the Digital Services Act. In Information Technology & Innovation Foundation (p. 2). https://www2.itif.org/2025-dsa-transparency.pdf

Kaushal et al. 2024

Kaushal, R, van de Kerkhof, J, Goanta, C, Spanakis, G & Iamnitchi, A 2024, Automated Transparency : A Legal and Empirical Analysis of the Digital Services Act Transparency Database. in FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, pp. 1121 - 1132. https:// doi.org/10.1145/3630106.3658960

Kumar et al. 2025

Ashish Kumar, Divya Singh, Rachna Jain, Deepak Kumar Jain, Chenquan Gan, Xudong Zhao, Advances in DeepFake detection algorithms: Exploring fusion techniques in single and multi-modal approach, Information Fusion, Volume 118, 2025, 102993, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2025.102993.

Labuz 2025

Łabuz, M. (2025), A Teleological Interpretation of the Definition of DeepFakes in the EU Artificial Intelligence Act—A Purpose-Based Approach to Potential Problems With the Word “Existing”. Policy & Internet, 17: e435. https://doi.org/10.1002/poi3.435

PUBLIC 2026

Deepfake detection technology. (2026). Commissioned by Department for Science, Innovation and Technology (DSIT) GOV.UK. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology

Sunil et al. 2025

Reshma Sunil, Parita Mer, Anjali Diwan, Rajesh Mahadeva, Anuj Sharma, Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation, Heliyon, Volume 11, Issue 3, 2025, e42273, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2025.e42273.

Yiğitalp Tulga 2026

Yiğitalp Tulga, A. (2026). Evaluating web-based deepfake detection tools: Risks, limits, and practical solutions for information security. EDPACS, 71(5), 93–104. https://doi.org/10.1080/07366981.2025.2564783