Liability of Online Platforms in Defamation Cases

by **Laura Herrerías Castro** *

Abstract: Online platforms provide an unprecedented space for exercising freedom of expression while simultaneously facilitating the immediate and potentially global spread of defamatory content. At the same time, AI plays a dual role as a generator of risks to individuals' fundamental rights and as an indispensable tool for detecting and preventing illegal content. This article explores the risks to the right to honor arising from the use of online platforms and their increasing reliance on AI, with a threefold aim: to establish the standard of conduct of online platforms in defamation cases, to assess the impact of AI developments on their liability regime, and to identify the remedies available to victims when platforms fail to comply with their due diligence obligations.

Keywords: Defamation, Digital Services Act, Artificial Intelligence, Online Platforms, Due Diligence Obligations

Recommended citation: Laura Herrerías Castro, Liability of Online Platforms in Defamation Cases, 16 (2025) JIPITEC 252 para 1.

A. Introduction

1 Social networks provide users with an unprecedented space to exercise their right to freedom of expression[2]. The US Supreme Court has referred to social media as "the vast democratic forums of the Internet"[3]. However, factors such as the immediate dissemination of content, its accessibility and interactivity, the lack of editorial control, and the permissibility of anonymity increase the risk of users infringing fundamental rights, especially the right to honor.

2 Article 8 of the European Convention on Human Rights (hereinafter ECHR) and Article 7 of the Charter of Fundamental Rights of the European Union (hereinafter CFEU) protect reputation as part of the right to respect for private life. The ECHR does not define "defamation", but under the European Court of Human Rights (hereinafter ECtHR) case law, it is generally a civil wrong committed by an individual against another or others that harms a person's reputation or good name[4].

* Laura Herrerías Castro is a Postdoctoral Researcher at the Faculty of Law of Pompeu Fabra University (Barcelona). This work was supported by the research project "Contractual and non-contractual liability of online platforms", funded by the Spanish Ministry of Economy, Industry and Competitiveness, the European Fund of Regional Development and the Spanish State Research Agency (PID2021-126354OB-I00). The author would like to thank Prof. Antoni Rubí Puig and Prof. Sonia Ramos González for his valuable comments on earlier versions of this article.

2 *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 December 2012) para 54.

3 *Packingham v North Carolina* 137 US 1730, 1735 (2017). As laid down in *Moody v NetChoice LLC* 603 US 707 (2024), 'Social-media platforms … structure how we relate to family and friends, as well as to businesses, civic organizations, and governments. The novel services they offer make our lives better and make them worse – create unparalleled opportunities and unprecedented dangers'.

4 T McGonagle, *Freedom of Expression and Defamation: A Study of the Case Law of the European Court of Human Rights* (Council of Europe 2016) 14 <https://rm.coe.int/16806ac95b> accessed 23 June 2025. Similarly, the US *Restatement (Second) of Torts* § 559 provides that 'A communication is defamatory if

3 The EU has neither harmonized substantive law on defamation[5] nor the conflict-of-law rules in that field[6]. Consequently, each court applies the law designated as applicable under its national conflict rules. In Spain, Article 7.7 of the Organic Act 1/1982, of 5 May, on the civil protection of the right to honor, to privacy and to one's own image, defines defamation as "the imputation of facts or the manifestation of value judgements by acts or expressions that infringe in any way the dignity of another person, damaging his fame or impinging on his self-esteem". In a similar vein, Article 29 of the French Law on the Freedom of the Press, of 29 July 1881, defines defamation as "any allegation or imputation of a fact which is prejudicial to the honor or reputation of the person or entity to which the fact is attributed".

4 In *Delfi v Estonia*, the ECtHR noted that defamatory and other types of clearly unlawful speech "can be disseminated like never before, worldwide, in a matter of seconds, and sometimes remain persistently available online"[7]. Facebook's Transparency Report[8] shows that defamation is the main reason for notices submitted in accordance with Article 16 of 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 (hereinafter DSA)[9]. On Google Maps, 99.3% of reported illegal content is for defamation[10], probably due to user reviews and opinions[11].

5 From a regulatory perspective, Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (hereinafter, ECD)[12] established a safe harbor for hosting service providers[13], which has been maintained in the DSA. While some of the principles that inspired the ECD have remained unchanged, the fight against illegal content has led to the adoption of a myriad of sector-specific and horizontal regulatory solutions, increasing platform responsibility. In turn, AI plays a dual role as a generator of new risks and as a tool for detecting and preventing illegal content. To address some of these risks, the EU legislator adopted Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (hereinafter, AIA)[14].

6 This paper explores the risks to the right to honour posed by online platforms and their increasing reliance on IA. It seeks to answer the following research questions: What is the required standard of conduct that online platforms must observe to avoid incurring liability for hosting defamatory content, what influence do developments in AI have on this liability regime, and what remedies are available to victims for platform's infringement of their due diligence obligations.

7 The paper is structured as follows: Section B examines the platform's liability regime under the DSA. Section C explores online platforms' reactive and proactive duties regarding defamatory user-generated content. Finally, Section D analyses the remedies for non-compliance with the due diligence obligations of the DSA.

---

it tends so to harm the reputation of another as to lower him in the estimation of the community or to deter third persons from associating or dealing with him'.

5 For a comparative analysis on the protection of the right to honor, see H Koziol, A Warzilek. *The protection of personality rights against invasions by mass media*. New York: Springer, 2005.

6 Art 1.2(g) Regulation (EC) 864/2007 of the European Parliament and of the Council of 11 July 2007 on the law applicable to non-contractual obligations (Rome II) [2007] OJ L199/40.

7 *Delfi v Estonia* App no 64569/09 (ECtHR [GC], 16 June 2015) para 110. See also *Editorial Board of Pravoye Delo and Shtekel v Ukraine* App no 33014/05 (ECtHR, 5 May 2011) para 63.

8 Regulation (EU) 2022/2065 *Digital Services Act* Transparency Report for Facebook (25 April 2025) <https://transparency.meta.com/reports/regulatory-transparency-reports/> accessed 23 June 2025.

9 OJ L 277, 27 October 2022, 1–102.

10 *EU Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report* (28 February 2025) <https://transparencyreport.google.com/report-downloads?hl=en> accessed 23 June 2025.

11 The commercial reputational interests of a company could not be equated with the reputation of an individual concerning his or her social status. Whereas the latter might have repercussions on one's dignity, interests of commercial reputation are devoid of that moral dimension. However, States enjoy a margin of appreciation as to the means they provide under domestic law to enable a company to challenge the truth, and limit the damage, of allegations which risk harming its reputation. See P Hirvelä, S Heikkilä *Right to respect for private and family life, home and correspondence*. Cambridge (UK): Instersentia. 2022, 69.

12 OJ L 178, 17 July 2000, 1–16.

13 This framework was inspired by Section 5 of the German *Teledienstegesetz* of 1997 and, especially, by Section 512 of the US Digital Millennium Copyright Act of 1998. See M Peguera Poch The DMCA Safe Harbors and Their European Counterparts: A Comparative Analysis of Some Common Problems. *Columbia Journal of Law & the Arts.* 2009, 32(4), 481.

14 OJ L 1689, 12 July 2024.

B. The Safe Harbor for Online Platforms under the DSA

8 The safe harbors for mere conduit, caching and hosting services regulated in Articles 12-14 ECD rested mainly on three factors[15]: the impossibility or excessive cost of monitoring user-generated content, the inequity of imposing liability on mere passive intermediaries, and the prevention of the chilling effects that the risk of liability could have on freedom of expression[16].

9 Nowadays, content moderation is not an ancillary aspect of what online platforms do; it is rather essential and definitional. As Gillespie claims: "Not only can platforms not survive without moderation, they are not platforms without it"[17]. The current best industry practice is to use automatic tools to narrow down the set of contentious content for vetting by human experts (human-in-command principle)[18]. For example, according to TikTok's Community Guidelines: "Content first goes through an automated review process. If content is identified as a potential violation, it will be automatically removed, or flagged for additional review by our moderators"[19]. Nevertheless, as Gillespie points out: "The overwhelming majority of what is being automatically identified are copies of content that have already been reviewed by a human moderator. Stats like these are deliberately misleading, implying that machine learning (ML) techniques are accurately spotting new instances of abhorrent content, not just variants of old ones"[20].

10 There are no neutral platforms, not only because they all moderate content but also because their main and sometimes only source of funding is advertising. For example, in 2024, Meta Platforms, Inc. obtained 98.9% of its net profit from targeted advertising[21]. Recommender systems aim to maximize platform revenue by displaying content tailored to users' interests[22]. Although illegal content may harm a platform's credibility and reputation[23], it often boosts user engagement, increases ad exposure, and ultimately drives more clicks on advertising links[24].

11 Despite the above, the DSA preserves the knowledge-and-take-down principle (Articles 4-6 DSA) as well as the no general monitoring obligation (Article 8 DSA), as both have allowed many novel services to emerge and scale up across the internal market[25]. Besides, some form of conditional immunity is still necessary to prevent collateral censorship[26]. Otherwise, platform operators would have strong incentives to over-censor, limit access or deny users' speech[27]. As Wilman highlights[28]: "The knowledge-based liability model thus aims to strike a middle-way. It avoids the negative consequences of stricter forms of liability that would impact not only the service providers themselves, but also their users[29]. At the same time, it does not completely preclude the possibility for aggrieved parties to have recourse to the service provider concerned where their rights are at stake"[30].

I. Knowledge-and-Take-Down

12 Pursuant to Article 6.1 DSA, hosting service providers

15 *Pursuant to art. 89(2) DSA*: 'References to Articles 12 to 15 of Directive 2000/31/EC shall be construed as references to Articles 4, 5, 6 and 8 of this Regulation, respectively'.

16 L Edwards, 'With Great Power Comes Great Responsibility?: The Rise of Platform Liability' in L Edwards (ed), *Law, Policy and the Internet* (Hart Publishing 2019) 257.

17 T Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media (Yale University Press 2018) 21.

18 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms* COM (2017) 555 final 14.

19 TikTok Community Guidelines <https://www.tiktok.com/community-guidelines/es> accessed 25 June 2025.

20 T Gillespie, 'Content moderation, AI, and the question of scale' (2020) 7(2) *Big Data & Society* 3.

21 Meta, *Meta Reports Fourth Quarter and Full Year 2024 Results* <https://investor.atmeta.com/investor-news/press-release-details/2025/Meta-Reports-Fourth-Quarter-and-Full-Year-2024-Results/> accessed 25 June 2025.

22 Personalized content leads to echo chambers and filter bubbles, see E Pariser *The filter bubble. What the Internet is hiding from you*. London: The Penguin Press, 2011, 9-10.

23 M C Buiten, A de Streel and M Peitz, 'Rethinking liability rules for online hosting platforms' (2020) 28(2) *International Journal of Law and Information Technology* 150.

24 R Griffin, 'The Sanitised Platform' (2022) 13(1) *JIPITEC* 42.

25 Recital 16 DSA.

26 M Husovec, 'Rising above liability: The Digital Services Act as a blueprint for the second generation of global internet rules' (2023) 38(3) *Berkeley Technology Law Journal* 110. See also J Grimmelmann and P Zhang, 'An economic model of online intermediary liability' (2023) 38(3) *Berkeley Technology Law Journal* 1039.

27 JM Balkin, 'Old-school/new-school speech regulation' (2014) 127 *Harvard Law Review* 2309.

28 F Wilman, 'The EU's system of knowledge-based liability for hosting service providers in respect of illegal user content – between the e-Commerce Directive and the Digital Services Act' (2021) 12 *JIPITEC* 323.

29 Cf art 47 *Cybersecurity Law of the People's Republic of China* <https://digichina.stanford.edu/work/translation-cybersecurity-law-of-the-peoples-republic-of-china-effective-june-1-2017/> accessed 25 June 2025.

30 Cf *s 230(c) US Communication Decency Act 1996.*

are exempt from liability for users' content as long as they lack actual knowledge or awareness of the illegality, and, upon obtaining such knowledge or awareness, act expeditiously to remove or restrict access to it. For an online platform to qualify for safe harbor protection, it must also provide its services neutrally, by a merely technical and automatic processing of the information provided by users.

1. Actual Knowledge v. Red Flag Knowledge

13 Under Section 512 (c)(1) of the US Digital Millennium Copyright Act of 1998 (hereinafter DMCA), the difference between actual and red flag knowledge is not between specific and generalized knowledge[31], but instead, between a subjective and an objective standard. The actual knowledge provision turns on whether the provider actually or "subjectively" knew of a specific infringement, while the red flag provision turns on whether the provider was aware of facts that would have made the specific infringement objectively obvious to a reasonable person[32].

14 Similarly, the European Court of Justice (hereinafter CJEU) interprets red flag knowledge as being aware, in one way or another, of facts or circumstances on the basis of which a diligent economic operator should have identified the illegality in question[33]. The situations covered include those in which the platform operator finds out illegal content as the result of an own-initiative investigation, as well as situations in which the operator is notified of the existence of such content by public authorities, trusted flaggers[34], or users. Under Article 16.3 DSA notices shall be considered to give rise to actual knowledge or awareness in respect of the specific item of information concerned "where they allow a diligent provider of hosting services to identify the illegality of the relevant activity or information without a detailed legal examination"[35].

15 Knowledge must be human, i.e. it is not sufficient that an algorithm detects potentially illegal content[36]. In the case of legal entities, the question arises as to when a content moderator's knowledge of illegality can be attributed to the platform operator. As stated by Hofmann, it can be assumed that the platform operator has knowledge or awareness of the illegality when it entrusts its employees with the autonomous management of content[37].

16 Finally, when platforms host manifestly illegal content, the rights and interests of others and society may entitle States to impose liability on online intermediaries without contravening Article 10 ECHR if they fail to take measures to remove it without delay, even without previous notification[38]. Content is considered manifestly illegal where it is evident to a layperson, without any substantive analysis, that is illegal[39]. This would be the case for war crimes, crimes against humanity, incitement to or apology of violence, certain acts of terrorism or child abuse content[40], but not for defamation[41].

31 Recital 22 DSA also establishes that knowledge must be content-specific: "Such actual knowledge or awareness cannot be considered to be obtained solely on the ground that that provider is aware, in a general sense, of the fact that its service is also used to store illegal content".

32 *Viacom Intern., Inc. v YouTube* (2nd Cir. 2012) 676 F.3d 19. In *Capitol Records, LLC v Vimeo, LLC* (2016) 826 F.3d 78 the US Court of Appeals 2nd Cir. concluded that a copyright owner's showing that a video posted by a user on the service provider's site includes substantially all of a recording of recognisable copyrighted music, and that an employee of the service provider saw at least some part of the user's material, was insufficient to sustain the copyright owner's burden of proving that the service provider had red flag knowledge of the infringement. The US Copyright Office argues that such a narrow interpretation of red flag knowledge minimizes an online platform's duty to act upon information of infringement and, in doing so, protects activities that Congress did not intend to protect. See US Copyright Office. Section 512 of title 17: a report of the register of copyrights (2020) 123 <https://www.copyright.gov/policy/section512/section-512-full-report.pdf> accessed 23 June 2025.

33 *Case C-324/09 L'Oréal and Others v eBay International AG* [2011] ECR I-6011, paras 120–122. See also P Valcke, A Kuczerawy and P-J Ombelet, 'Did the Romans Get It Right? What Delfi, Google, eBay, and UPC TeleKabel Wien Have in Common' in M Taddeo and L Floridi (eds), *The Responsibilities of Online Service Providers* (Springer International Publishing 2017) 101.

34 Art 22 DSA.

35 Article 16.3 establishes an irrebuttable presumption of knowledge. See F Raue Article. 16. Notice and action mechanisms. In B Hofmann/F Raue (dirs.). *Digital Services Act: Article-by-article commentary*. Baden-Baden: Nomos. 2024, 337.

36 P Van Eecke, 'Online service providers and liability: A plea for a balanced approach' (2011) 48 *Common Market Law Review* 1475.

37 B Hofmann, 'Article 6. Hosting' in B Hofmann and F Raue (eds), *Digital Services Act: Article-by-Article Commentary*(Nomos 2024) 170.

38 *Delfi* (n 7) para 159.

39 Recital 63 DSA.

40 G Frosio and C Geiger, 'Taking fundamental rights seriously in the Digital Services Act's platform liability regime' (2023) 29 *European Law Journal* 64.

41 *Magyar Tartalomszolgáltatók Egyesülete* (MTE) *and Index.hu Zrt v Hungary* no 22947/13 (ECtHR, 2 February 2016) para 64.

2. Expeditious Reaction

17 The DSA does not include any time limit for removing or disabling access to illegal content[42]. Regarding the treatment of notifications, Recital 52 DSA merely states that: "Providers of hosting services should act upon notices in a timely manner, in particular by taking into account the type of illegal content being notified and the urgency of taking action"; and Recital 89 DSA that: "Other types of illegal content may require longer or shorter timelines for processing of notices, which will depend on the facts, circumstances and types of illegal content at hand".

18 Facebook's Transparency Report shows that the average time needed to take action on reported content is 13.2 hours, while Instagram's is 18.4 hours[43]. Both reports warn that more complex decisions, such as defamation or harassment, may require more time or additional guidance from specialised staff.

19 In conclusion, the expeditious reaction of platforms should be assessed on a case-by-case basis depending on factors such as[44]: the type of illegality, the volume of hosted content, the number, accuracy and source of notifications, as well as the availability of content moderation mechanisms.

3. Neutrality Test

20 Recital 18 DSA sets forth that: "The exemptions from liability established in this Regulation should not apply where, instead of confining itself to providing the services neutrally by a merely technical and automatic processing of the information provided by the recipient of the service, the provider of intermediary services plays an active role of such a kind as to give it knowledge of, or control over, that information"[45].

21 As Peguera Poch points out the test consisting of whether the activity gives the provider knowledge of or control over the hosted information seems ill-suited because platforms usually have some basic form of control over the information they host. Furthermore, it is at odds with the fact that under the hosting safe harbor, a provider is only supposed to lose protection when it obtains knowledge regarding the illegal nature of specific content and fails to expeditiously remove or block access to it[46].

22 The neutrality test should be interpreted narrowly so that a platform cannot benefit from the safe harbor if it knowingly participates or collaborates in the dissemination of illegal content or if it has editorial control over it[47]. Nonetheless, when editorial control is fully automated or AI is used to validate content before publication, platforms should be considered "neutral" in terms of Recital 18 DSA, as the activity consists of a "purely technical and automatic processing of information".

II. No General Monitoring Obligation

23 Article 8 DSA, in very similar terms to Article 15.1 ECD, reads as follows: "No general obligation to monitor the information which providers of intermediary services transmit or store, nor actively to seek facts or circumstances indicating illegal activity shall be imposed on those providers"[48]. The prohibition of general monitoring does not affect the possibility for judicial or administrative authorities to require the service provider to terminate or prevent specific infringements[49], even where platform operators

42 During the parliamentary debate on the DSA proposal, the Committee on legal affairs (Rapporteur: Geoffrey Didier) proposed to add to Article 6.1 the following paragraph (amendment 111): "1a. Without prejudice to specific deadlines, set out in Union law or within administrative or legal orders, providers of hosting services shall, upon obtaining actual knowledge or awareness, remove or disable access to illegal content as soon as possible and in any event: (a) within 30 minutes where the illegal content pertains to the broadcast of a live sports or entertainment event; (b) within 24 hours where the illegal content can seriously harm public policy, public security or public health or seriously harm consumers' health or safety; (c) within 72 hours in all other cases where the illegal content does not seriously harm public policy, public security, public health or consumers' health or safety" <https://www.europarl.europa.eu/doceo/document/A-9-2021-0356_EN.html> accessed 23 June 2025.

43 None of the reports detail the reaction time according to the type of illegal content, nor do they include information on standard deviation.

44 J Riordan, *The Liability of Internet Intermediaries* (Oxford University Press 2016) 408.

45 See also *Joined Cases C-236/08, C-237/08 and C-238/08 Google France* (CJEU 23 March 2010) paras 116–119; *Case C-324/09 L'Oréal and others* (n 33) paras 115–116; *Joined Cases C-682/18 and C-683/18 YouTube and Cyando* (CJEU 22 July 2021) paras 107–109.

46 M Peguera Poch, 'The Platform Neutrality Conundrum and the Digital Services Act' (2022) 53 *International Review of Intellectual Property and Competition Law* 683.

47 Art 6(2) and recitals 18 and 20 DSA.

48 For a thorough analysis of the CJEU's interpretation of general monitoring prohibition see T H Oruç, The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in light of Recent Developments: Is it still necessary to maintain it? *JIPITEC*, 2022, 13(3), 179-190.

49 Art 9 DSA.

meet the conditions set out in Article 6.1 DSA[50].

24 The distinction between general and specific obligations has been developed in the case law of the CJEU. In *L'Oréal and others*, the CJEU resolved whether it is possible to issue an injunction requiring a website operator to prevent future infringements of intellectual property rights. The CJEU responded that injunctions cannot consist of active monitoring of all users' data, but accepted injunctions to prevent further infringements by the same user in respect of the same trademarks[51]. In *Tommy Hilfiger Licensing and others,* the CJEU insisted on the idea that: "The intermediary may be forced to take measures which contribute to avoiding new infringements of the same nature by the same market-trader from taking place"[52]. Thus, an injunction that meets this double identity requirement - same subject and same object - does not entail a general monitoring obligation.

25 Subsequently, in *Scarlet Extended* and *SABAM*, the CJEU concluded that an injunction for preventing copyright infringements requiring an online intermediary to install a system for filtering all information stored on its servers, exclusively at its expense and for an unlimited period of time would be contrary to Article 15.1 ECD[53]. The CJEU emphasized that a fair balance must be struck between the fundamental rights protected by the CFEU[54]. A filtering system of this type that seeks to protect intellectual property rights (Article 17.2 CFEU) does not respect the principle of proportionality insofar as it implies, on the one hand, a substantial infringement of the intermediary's freedom to conduct business (Article 16 CFEU); and, on the other hand, it would significantly affect the right to the protection of personal data of users (Article 8 CFEU) and their right to freedom of expression (Article 11 CFEU) due to the risk that the system would not adequately distinguish between lawful and unlawful content[55].

26 Finally, in *Glawischnig-Piesczek* the CJEU stated that it is not contrary to Article 15.1 ECD an injunction ordering a social network to remove information the content of which is identical or equivalent to information which was previously declared to be defamatory, or to block access to that information, irrespective of who the author is. Such an injunction would not entail a disproportionate impact on the right to freedom to conduct a business, as: "The monitoring of and search for information which it requires are limited to information containing the elements specified in the injunction, and its defamatory content of an equivalent nature does not require the host provider to carry out an independent assessment, since the latter has recourse to automated search tools and technologies"[56].

27 The CJEU seems to ignore that human communication is culturally sensitive and that it is highly complex at a technical level to capture the context of a publication. Identical content can have different meanings; for example, swear words or insults can be harmless when addressed to a close person[57]. Despite progress in IA, automatic moderation systems cannot reliably distinguish between defamation and its critique, news coverage or satire[58].

28 As for equivalent content, AG Szpunar warned that: "A reproduction of the information that was characterized as illegal containing a typographical error and a reproduction having slightly altered syntax or punctuation constitutes equivalent information. It is not clear, however, that the equivalence referred to in the second question does not go further than such cases"[59]. AG Szpunar's concerns were ultimately confirmed when the CJEU concluded that: "Injunction[s] must be able to extend to information, the content of which, whilst essentially conveying the same message, is worded slightly differently, because of the words used or their combination, compared with the information whose content was declared to be illegal"[60], provided

50 Art 6(4) and recital 25 DSA.

51 *L'Oréal and others* (n 33) paras 139-141.

52 Case C-494/15 *Tommy Hilfiger Licensing and Others v Delta Center* (CJUE, 7 July 2016) para 34.

53 Case C-360/10 *SABAM* (CJUE, 16 February 2012) para 38; Case C-70/10 *Scarlet Extended* (CJUE, 24 November 2011) para 40.

54 GC Case C-275/06 *Promusicae* (CJUE, 29 January 2008) para 68.

55 *SABAM* (n 53) paras 46–50; *Scarlet Extended* (n 53) paras 48–53. On the compatibility of arts 11, 16 and 17(2) CFEU with injunctions to prevent copyright infringements see also Case C-314/12 *UPC Telekabel Wien* (CJEU, 27 March 2014) and Case C-484/14 *Mc Fadden* (CJEU, 15 September 2016).

56 Case C-18/18 *Glawischnig-Piesczek* (CJEU, 3 October 2019) para 46.

57 T Dias Oliva, D M Antonialli, A Gomes, Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 2021, 25, 706 ("In-group/out-group status may help create contextual conditions that predispose particular experiences of language. A word that is experienced as a slur when hurled by an outsider can be experienced as a joke when used by an in-group member. LGBTQ people reclaim slurs by using them within the community. A word that might normally convey malice here conveys solidarity").

58 D Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling' (2020) 69(6) *GRUR International* 618; J Daskal and K Klonick, 'When a Politician Is Called a "Lousy Traitor," Should Facebook Censor It?' (*The New York Times* 27 June 2019) <https://www.nytimes.com/2019/06/27/opinion/facebook-censorship-speech-law.html> accessed 23 June 2025.

59 *Glawischnig-Piesczek* (n 56) Opinion of AG Szpunar para 67.

60 *Glawischnig-Piesczek* (n 56) para 41.

that injunctions contain specific elements, such as the name of the victim, the circumstances in which the infringement was determined and equivalent content to that which was declared to be illegal, "so that the hosting provider concerned is not required to carry out an independent assessment of that content"[61].

29 The problem, again, is that detecting equivalent content requires considering the actual meaning of the publication at issue, and in most cases, it is impossible to do without human oversight[62]. As AG Saugmandsgaard Øe points out: "Although intermediary providers are technically well placed to combat the presence of certain illegal information disseminated through their services, they cannot be expected to make 'independent assessments' of the lawfulness of the information in question. Those intermediary providers do not generally have the necessary expertise and, above all, the necessary independence to do so – particularly when they face the threat of heavy liability. They cannot therefore be turned into judges of online legality, who are responsible for coming to decisions on legally complex questions"[63].

30 The CJEU suggests using algorithmic content moderation systems to avoid making an independent assessment of the lawfulness. Nevertheless, as discussed in the next section, such mechanisms are prone to false positives and false negatives.

C. The Standard of Conduct of Online Platforms in Defamation Cases

I. Required Standard of Conduct

31 The required standard of conduct is that of a reasonable (legal) person in the circumstances of the case[64]. Assessing the defendant's conduct involves considering legal provisions, as well as the custom or best practices of the relevant economic sector, which may be reflected in the corresponding codes of conduct[65]. In the absence of the above, negligence should be established by balancing the expected risk, on one hand, and the cost of precautionary measures, on the other[66].

32 Under Article 4:102 (1) PETL, the required standard of conduct depends, among other factors, on the nature and value of the protected interest involved, the foreseeability of the damage, as well as the availability and the costs of precautionary or alternative methods[67]. Likewise, the US Restatement Third, Torts: Liability for Physical and Emotional Harm § 3 states that the principal factors to consider in ascertaining whether a person's conduct lacks reasonable care are: the foreseeable likelihood that the person's conduct will result in harm, the foreseeable severity of any harm that may ensue, and the burden of precautions to eliminate or reduce the risk of harm.

33 This approach has its origins in the reasoning of Judge Learned Hand in *United States v Carroll Towing Co*[68]. Following Hand's liability formula for negligence, a potential injurer is negligent if and only if B < PL, where B is the burden of taking precautions, P is the probability of loss, and L is the gravity of loss. The balancing approach rests on and expresses a simple idea: conduct is negligent if its disadvantages outweigh its advantages. In other words, the actor's conduct is negligent if the magnitude of the risk outweighs the burden of risk prevention[69].

1. Risk of Harm

34 The foreseeability of disseminating illegal content depends on the type and popularity of platforms.

61 ibid para 45. See also Case C-401/19 *Poland v Parliament and Council* (CJEU 26 April 2022) para 90 'The providers of those services cannot be required to prevent the uploading and making available to the public of content which, in order to be found unlawful, would require an independent assessment of the content by them in the light of the information provided by the rightholders and of any exceptions and limitations to copyright'.

62 E Rosati, 'Material, personal and geographic scope of online intermediaries' removal obligations beyond *Glawischnig-Piesczek* (C-18/18) and defamation' (2019) 41(11) *European Intellectual Property Review* 676.

63 *Poland* (n 61) Opinion of AG Saugmandsgaard Øe para 197. See also A/HRC/38/35 para 17 'Complex questions of fact and law should generally be adjudicated by public institutions, not private actors whose current processes may be inconsistent with due process standards and whose motives are principally economic'.

64 Art 4:102(1) *Principles of European Tort Law* (PETL); Art VI – 3:102 *Draft Common Frame of Reference*.

65 There is currently no code of conduct for combating online defamation. On disinformation see 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 (OJ C, C/2024/3014, 26.4.2024).

66 C van Dam, *European Tort Law* (2nd edn, Oxford University Press 2013) 236.

67 P Widmer, 'Art 4:102 Required standard of conduct' in European Group on Tort Law, *Principles of European Tort Law: Text and Commentary* (Springer 2007) 75–79.

68 159 F2d 169 (2nd Cir 1947).

69 American Law Institute, *Restatement Third, Torts: Liability for Physical and Emotional Harm* (2010) comment e 30.

As outlined in the introduction to this paper, defamation is one of the main reasons for users' complaints under Article 16 DSA. The permissibility of anonymity is also relevant, as it facilitates wrong by eliminating accountability[70]. As Citron notes: 'Online, bigots can aggregate their efforts even when they have insufficient numbers in any one location to form a conventional hate group. They can disaggregate their offline identities from their online presence, escaping social opprobrium and legal liability for destructive acts. Both of these qualities are crucial to the growth of anonymous online mobs'[71].

35 Platform operators are aware that users sometimes post illegal content. Nonetheless, foreseeability cannot be assessed in abstract terms; it must be evaluated in relation to whether the party who caused the harm could have reasonably foreseen the specific outcome of their conduct. Platform operators should be held liable only when they have actual knowledge or become aware of a specific illegal content, as it is not foreseeable in advance that one of their millions of users would commit a specific infringement.

36 In terms of the severity of harm, Article 2.102 (2) PETL sets forth that: 'Life, bodily or mental integrity, human dignity and liberty enjoy the most extensive protection'. The right to honour derives from human dignity, aiming to preserve both the feeling that a person has of their qualities (subjective honour) and reputation (objective honour)[72].

37 Online defamation usually causes non-pecuniary losses. In general, these losses are recoverable only when the infringement of the protected interest causes substantial harm to the victim's emotional well-being[73]. Nonetheless, Article 9.3 of the Spanish Organic Act 1/1982 establishes an irrebuttable presumption of non-pecuniary damages. This presumption is justified both on the grounds of the difficulty of the proof as well as in the fact that the specific nature of the protected interests that have been infringed permits the reasonable presumption that a non-pecuniary loss has taken place[74].

70 *McIntyre v Ohio Elections Com'n*, 514 US 334 (1995) 1537.

71 DK Citron, 'Cyber Civil Rights' (2009) 89 *Boston University Law Review* 64.

72 A De Cupis, *I diritti della personalità* (2nd edn Giuffrè Editore 1982) 251–252.

73 C von Bar, *The Common European Law of Torts*, vol II (Clarendon Press 2000) 20.

74 M Martín-Casals and J Solé Feliu, 'The protection of personality rights against invasions by mass media in Spain' in H Koziol and A Warzilek (eds), *The Protection of Personality Rights Against Invasions by Mass Media* (Springer 2005) 329.

2. Benefits of the Conduct

38 The benefits of the conduct should be assessed by considering both the interests of platforms and users. On the one hand, online platforms enjoy the right to freedom to conduct a business as provided for in Article 16 CFEU. This right encompasses the freedom for any platform to use, within the limits of liability for its own acts, the economic, technical and financial resources available to it. Additionally, platforms benefit from the freedom to impart information as guaranteed by Articles 11 CFEU and 10 ECHR.

39 To resolve the question of whether the domestic courts' decisions holding an online intermediary liable for defamatory comments were in breach of its freedom of expression, the ECtHR identified the following aspects as relevant for its analysis: a) the context and content of comments, b) the measures taken by the intermediary to prevent or remove the comments, c) the liability of the actual authors of the comments as an alternative to the intermediary's liability, d) the prior conduct of the injured party, e) the consequences of the domestic proceedings for the intermediary, and f) the consequences of the comments for the injured party[75].

40 Based on these criteria, in *Delfi v Estonia* the ECtHR held that it had been justified to order a news portal to pay damages (approximately 320€) for anonymous comments posted on its site, given its failure to take measures to remove clearly illegal comments, which amounted to hate speech or incitements to violence, without delay[76]. In contrast, in *MTE and Index.hu Zrt v Hungary*, the ECtHR found that strict liability of news portals for defamatory comments was incompatible with Article 10 ECHR. It held that there was no reason to state that, accompanied by effective procedures allowing for rapid response, the notice-and-take-down system had not functioned as an appropriate tool for protecting commercial reputation[77].

41 On the other hand, users have the right to freedom of expression, which applies not only to information or ideas that are favorably received but also to those that offend, shock or disturb[78]. For Article 8 ECHR to come into play, the attack on a person's reputation must attain a certain level of seriousness, in a manner causing prejudice to personal enjoyment

75 *Delfi* (n 7) para 142; *MTE* (n 41) paras 72–88; *Høiness v Norway* no 43624/14 (ECtHR, 19 March 2019) para 67; *Jezior v Poland* no 31955/11 (ECtHR, 4 June 2020) para 53; *Sanchez v France* no 45581/15 (ECtHR [GC], 15 May 2023) para 167.

76 *Delfi* (n 7) para 159.

77 *MTE* (n 41) para 91.

78 *Handyside v United Kingdom* no 5493/72 (ECtHR, 7 December 1976) para 49.

of the right to respect for private life[79]. Despite millions of users posting content online every day[80], many of those comments are likely to be too trivial for them to cause any significant damage to another person's reputation[81]. In this sense, in *MTE and Index.hu Zrt v Hungary* the ECtHR indicated that: 'Without losing sight of the effects of defamation on the Internet, especially given the ease, scope and speed of the dissemination of information (...) regard must be had to the specificities of the style of communication on certain Internet portals. For the Court, the expressions used in the comments, albeit belonging to a low register of style, are common in communication on many Internet portals - a consideration that reduces the impact that can be attributed to those expressions'[82].

3. Cost of Precautionary Measures

42 Article 16 DSA addresses one of the gaps in the ECD by obligating hosting service providers to establish notice and take-down mechanisms[83]. These mechanisms must be easy to access and user-friendly and must allow for the submission of notices exclusively by electronic means. Additionally, they should facilitate the submission of notices that are sufficiently precise and adequately substantiated[84].

43 In defamation cases, before notification the costs of detection and removal usually exceed the expected harm due to the limited information that the platform has about the truthfulness of the information. In contrast, the receipt of a notification that complies with the requirements mentioned in Article 16.2 DSA considerably reduces the burden on the platform operator. Therefore, the standard of conduct expected from platforms depends, to a large extent, on the diligence previously exercised by the victim[85].

44 Platforms generally have teams of reviewers and algorithmic content moderation systems to fight against illegal content. For example, according to the Facebook and Instagram Transparency Reports, Meta has a team of 212 moderators that review content in English[86]. Even if platforms have sufficient content reviewers who understand the language, cultural, political and social context of the publications, incorrect decisions cannot be ruled out, as non-lawyers must decide in just a few seconds whether a message is likely to harm a person's dignity in a given country.

45 Investing in the development of algorithmic content moderation systems is costly, so it is a viable option only for large platforms[87]. Requiring the same level of diligence from SMEs as from Big Tech companies would stifle market participation and free competition, making it extremely difficult for new businesses to enter or forcing out those that cannot afford the costs. For this reason, the additional obligations imposed under the DSA on providers of online platforms do not apply to providers that qualify as micro or small enterprises[88].

46 Hiring third-party services to carry out content moderation tasks is possible, but many of these services are not designed to detect defamatory content, and the few that can detect text toxicity have a high error rate. For example, Perspective - a free API developed by Jigsaw and Google - automatically evaluates messages and ranks them according to attributes such as toxicity, severe toxicity, identity attack, insult, profanity and threat.

79 *Axel Springer AG v Germany* no 47940/99 (ECtHR [GC], 7 February 2012) para 83.

80 DOMO, 'Data Never Sleeps 12.0' (2024) <https://www.domo.com/learn/infographic/data-never-sleeps-12> accessed 23 June 2025.

81 *Tamiz v United Kingdom* no 3877/14 (ECtHR, 19 September 2017) para 80; Çakmak *v Turkey* no 45016/18 (ECtHR, 7 September 2019) para 50.

82 *MTE* (n 41) para 77.

83 Art 14(3) and art 21(2) Directive 2000/31/EC on Electronic Commerce (ECD).

84 Recital 50 DSA; P Wolters and R Gellert, 'Towards a better notice and action mechanism in the DSA' (2023) 14(3) *JIPITEC* 413–418.

85 M Husovec, 'The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown? Which Is Superior? And Why?' (2018) 42 *Columbia Journal of Law & the Arts* 66.

86 Content moderators by official EU language for Facebook and Instagram combined: Bulgarian (55), Croatian (56), Czech (63), Danish (39), Dutch (154), Estonian (6), Finnish (24), French (630), German (470), Greek (37), Hungarian (44), Irish (0), Italian (427), Latvian (4), Lithuanian (11), Maltese (1), Polish (112), Portuguese (2088), Romanian (74), Slovak (49), Slovenian (8), Spanish (3110), Swedish (78). For languages widely spoken outside the EU (French, English, Spanish and Portuguese) there are additional reviewers for reports from non-EU countries. *Regulation (EU) 2022/2065 Digital Services Act* Transparency Report for Facebook (25 April 2025) <https://transparency.meta.com/reports/regulatory-transparency-reports/> accessed 23 June 2025.

87 Buiten, de Streel and Peitz (n 23) 153 ("Large platforms may be able to save on costs of detection, monitoring and removal because of economies of scale. It may pay off for large hosting platforms to invest in developing or acquiring software tools to identify and filter out illegal content. Large hosting platforms can spread the high fixed costs of such software tools over all instances of illegal material (...). Investments in advanced software tools might not pay off for smaller platforms, forcing them to do more detection and monitoring work manually, at higher costs and often with less precision per instance of illegal material").

88 Arts 19 and 29 DSA.

However, Hosseini *et al*. showed that the system has a high false negative rate, as it is relatively easy to deceive[89]. The developers of Perspective themselves have admitted its fallibility: "Our models are not perfect and will make errors. It will be unable to detect patterns of toxicity it has not seen before, and it may incorrectly detect toxicity in healthy comments that contain patterns similar to previous toxic conversations. Because of this, Perspective is not intended for use cases such as fully automated moderation"[90].

47 In summary, a diligent economic operator should not be required to conduct *ex ante* or proactive control of defamatory content. Such an obligation would not only contradict Article 8 DSA - interpreted in light of Recital 30 DSA -, but also would not respect the principle of proportionality, because platforms do not have the technical and human resources necessary to proactively identify and remove defamatory content with a sufficient level of accuracy.

48 A different question is whether, given the current state of the art, online platforms can be subjected to notice-and-stay-down obligations to prevent the reappearance of previously notified defamatory content.

II. Content Moderation Mechanisms for Preventing Defamatory Content

49 Online platforms generally employ two automated techniques for content moderation: matching systems and predictive systems[91]. The former checks if a piece of content is identical to another previously identified as defamatory, while the latter predicts the likelihood that previously unseen content is defamatory.

1. Matching Systems

a.) Word Filters

50 Word filters compare words or expressions against a database to prevent, block or remove undesired text. Most social networks allow users to personalize blacklists. For instance, Facebook lets users choose up to 1.000 keywords in any language to block from comments on their profiles[92]. Many of these platforms also have their own blacklists. Still, the functioning of these filters is opaque, as no platform provides a definition or examples of what it considers offensive or inappropriate.

51 In *Alone in the Dark*, the German Federal Court of Justice concluded that it was technically and economically reasonable for an online intermediary to use word filters to prevent copyright infringements[93]. The Frankfurt Regional Court reached the same conclusion in a reputation protection case[94]. Nonetheless, word filters have important limitations. Firstly because of the lack of exhaustiveness of all words or combinations of words that may constitute a defamatory comment. Secondly, users can easily circumvent the system by introducing small modifications to the text[95]. Thirdly, filters do not consider context, and therefore generate a high rate of false positives. For example, YouTube deleted several accounts of well-known YouTubers due to a filter error when it interpreted the acronym "CP" as referring to "child pornography" when it meant "combat points" concerning the Pokemon GO video game[96].

b.) Hashing

89 H Hosseini, S Kannan, B Zhang and others, 'Deceiving Google's Perspective API Built for Detecting Toxic Comments' (2017) 2–3 <https://arxiv.org/abs/1702.08138> accessed 23 June 2025.

90 'Perspective FAQs' <https://developers.perspectiveapi.com/s/about-the-api-faqs?language=en_US> accessed 23 June 2025.

91 N Chowdhury, 'Automated Content Moderation: A Primer' (2022) 2 <https://cyber.fsi.stanford.edu/news/automated-content-moderation-primer> accessed 23 June 2025.

92 See <https://www.facebook.com/help/131671940241729> accessed 23 June 2025.

93 BGH 12 July 2012 I ZR 18/11 paras 33–35 'Die Eignung eines Wortfilters mit manueller Nachkontrolle für die Erkennung von Urheberrechtsverletzungen wird nicht dadurch beseitigt, dass er mögliche Verletzungshandlungen nicht vollständig erfassen kann'.

94 Landgericht Frankfurt am Main 14 December 2022 no 2-03 O 325/22 ECLI:DE:LGFFM:2022:1214.2.03O325.22.00 para 3 'Die Kammer kann angesichts dessen nicht erkennen, warum diese Identifizierung der rechtsverletzenden Tweets für die Antragsgegnerin technisch und wirtschaftlich, beispielsweise anhand der von der Antragstellerseite vorgeschlagenen Stichworte, unzumutbar sein sollte'.

95 E Llansó, 'No amount of AI in content moderation will solve filtering's prior restraint problem' (2020) 7(1) *Big Data & Society* 2.

96 T Gerken, 'YouTube backtracks after Pokemon child abuse ban' (*BBC* 18 February 2019) <https://www.bbc.com/news/technology-47278362> accessed 23 June 2025.

52 A hash value is an alphanumeric string that serves to identify an individual digital file as a kind of digital fingerprint[97]. Hashing consists of two phases: the generation and storage of the hash in a database, and the comparison of hashes for matches. For example, to prevent the reappearance of a meme, the algorithm must transform the image (meme.jpg) into a hash (a996be1eb1e210958219e0bb015d5420) and record that value in a database. Identical files have the same hash so if a user tries to post a copy of the meme, the system will prevent it.

53 Hashing can be cryptographic or perceptual. The advantages of cryptographic hashing are that it requires little data storage capacity, does not involve a large investment as open-source solutions are available, it is relatively easy to implement, and can accurately identify exact duplicates of a digital file[98]. However, one of its major disadvantages is its lack of robustness[99], since even the slightest manipulation results in a completely different hash[100].

54 In contrast, perceptual hashing does not attempt to determine whether two files are identical but whether they are sufficiently similar[101]. In the case of images, perceptual hashing extracts a fingerprint based on certain characteristics that resist possible modifications such as compression, color changes, rotation, the addition of text, or any other that does not fundamentally change the underlying content but alters the pixel values[102]. For example, in 2007, YouTube launched Content ID to identify matches of copyright-protected content[103]; in 2009, Microsoft Corporation developed PhotoDNA to prevent the dissemination of known child exploitation material[104]; and in 2017, the Global Internet Forum to Counter Terrorism created a hash-sharing database of terrorist and violent extremist content[105].

55 The judgement of 8 April 2022 of the Frankfurt Regional Court held that Facebook was liable for disseminating a defamatory meme because it did not take reasonable measures to prevent further identical or similar infringements. For the Frankfurt Regional Court, it was decisive that Facebook could have prevented them by using hashing[106]. However, like word filters, hashing is also prone to false positives, given the difficulties of discerning the publication context.

2. Predictive Systems

a.) Data Collection and Classification

56 For a system to predict the probability that content is illegal it must be trained with numerous examples to identify common characteristics. Each piece of training data is called "document", and a compilation of documents is called "corpus"[107].

57 Training data can be obtained through manual searches or from pre-existing databases. For example, the Hate Speech Dataset Catalogue includes open datasets for hate speech, online abuse, and offensive language[108]. Likewise, the Offensive Language Identification Dataset consists of an open corpus

97 *United States v Wellman* 663 F3d 224 (4th Cir 2011). See also E Engstrom and N Feamster, 'The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools' (2017) 12–13 <https://www.engine.is/the-limits-of-filtering> accessed 23 June 2025.

98 European Union Intellectual Property Office, *Automated Content Recognition: Discussion Paper – Phase 1 Existing Technologies and Their Impact on IP* (2020) 8–9.

99 R Gorwa, R Binns and C Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) *Big Data & Society* 4.

100 This can be verified through the following link: <https://www.md5.cz> accessed 23 June 2025.

101 C Shenkman, D Thakur and E Llansó, 'Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis' (Centre for Democracy & Technology 2021) 39 <https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/> accessed 23 June 2025.

102 H Farid, 'An Overview of Perceptual Hashing' (2021) 1(1) *Journal of Online Trust and Safety* 5.

103 'How Content ID works' <https://support.google.com/youtube/answer/2797370?hl=en-GB&sjid=14937822664490622239-EU> accessed 23 June 2025.

104 Microsoft, 'PhotoDNA' <https://www.microsoft.com/en-us/photodna> accessed 23 June 2025.

105 GIFCT, 'Hash-Sharing Database' <https://gifct.org/hsdb/> accessed 23 June 2025.

106 Landgericht Frankfurt am Main 8 April 2022 no 2-03 O 188/21 ECLI:DE:LGFFM:2022:0408.2.03O188.21.00 para 3 'Es ist zwischen den Parteien unstreitig, dass zum Ausgangspost identische Bilder über den Vergleich der Hashwerte automatisiert identifiziert werden können (...) Es ist zwischen den Parteien ebenfalls unstreitig, dass es technische Möglichkeiten gibt, nicht nur fast identische, sondern sogar ähnliche Bilder zu erkennen, indem man Abstriche hinsichtlich des Grads der Übereinstimmung beim Hashwert macht und die so gefundenen Kandidaten mittels PDNA und OCR überprüft'. The judgment of 25 January 2024 of the Frankfurt Court of Appeals (16 U 65/22) upheld the judgment of first instance.

107 A Stefanowitsch, Corpus Linguistics: A Guide to the Methodology (Language Science Press 2020) 22.

108 'Hatespeechdata' <https://hatespeechdata.com/> accessed 23 June 2025. The dataset is maintained by Leon Derczynski, Bertie Vidgen, Hannah Rose Kirk, Pica Johansson, Yi-Ling Chung, Mads Guldborg Kjeldgaard Kongsbak, Laila Sprejer and Philine Zeinert.

of 14200 English language documents on offensive language[109]. Both databases are useful for training a natural language processing (hereinafter NLP) system to detect this type of language. However, a tool developed using datasets in English may not function well when used to moderate speech in other languages[110].

58 Once the initial corpus has been compiled, crowdsourcing services are usually hired to classify or annotate the documents. Each document is often analyzed by 3-5 people and only those that pass a minimum threshold of consensus among the classifiers are incorporated into the final corpus. In order to correctly classify documents, it is essential to establish a clear, simple and consistent definition of what constitutes illegal content[111]. The problem arises from the lack of a universal definition of "defamation". Attempting to simplify it may result in misclassifying messages whose illegality requires a more complex analysis, which can vary from country to country. Additionally, definitions with subjective components pose a risk of introducing bias[112], as well as under-representation of language or expressions used by or against certain groups of people[113]. As Llansó *et al*. note: "If these datasets do not include examples of speech in different languages and from different groups or communities, the resulting tools will not be equipped to parse these groups' communication"[114].

59 To address these challenges, the Council of Europe recommends evaluating and testing algorithmic systems with sufficiently diverse and representative sample populations, without drawing on or discriminating against any particular demographic group[115]. Regarding high-risk AI systems, Article 10.3 AIA states that: "Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose", and Article 10.4 that: "Data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk AI system is intended to be used". Nevertheless, AI systems used for content moderation are, in principle, not high-risk systems and therefore do not fall under Chapter III of the AIA.

b.) Pre-Processing of Data

60 After collection and classification, the next step is pre-processing data through NLP. Pre-processing involves preparing and reducing all data to facilitate its computational representation. This process includes various methods such as tokenization[116], removing stop words[117], punctuation marks or special characters, and transforming symbols into words.

61 Next, it is necessary to extract a numerical representation of the corpus using vector representation models, such as Bag of Words (hereinafter BoW). BoW involves creating a vocabulary of all the words in the corpus and then generating a matrix where each row represents a document, and each column represents a word from the corpus. The matrix values indicate the frequency or importance of that word in the corresponding

109 'OLID' <https://sites.google.com/site/offensevalsharedtask/olid> accessed 23 June 2025. The dataset was created by Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra and Ritesh Kumar; M Zampieri, S Malmasi, P Nakov and others, 'Predicting the Type and Target of Offensive Posts in Social Media' in *Proceedings of NAACL-HLT 2019* 1415–1420.

110 A Marsoof, A Luco and H Tan, 'Content-filtering AI systems – limitations, challenges and regulatory approaches' (2023) 32(1) *Information & Communications Technology Law* 78.

111 M Barral Martínez, 'Platform regulation, content moderation, and AI-based filtering tools: some reflections from the European Union' (2023) 14(1) *JIPITEC* 216.

112 A Balayn, J Yang and Z Szlavik, 'Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature' (2021) 4(3) *ACM Transactions on Social Computing* 26; R Binns, M Veale and M van Kleek, 'Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation' in GL Ciampaglia, A Mashhadi and T Yasseri (eds), *Social Informatics: 9th International Conference, SocInfo 2017 (Part II)* (Springer 2017) 411–12.

113 A Díaz and L Hecht-Felella, 'Double Standards in Social Media Content Moderation' (2021) 11 <https://www.brennancenter.org/es/node/9225> accessed 23 June 2025; N Duarte, E Llansó and A Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' (Center for Democracy & Technology 2017) 16 <https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/> accessed 23 June 2025.

114 E Llansó, J van Hoboken, P Leerssen *et al*, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' (Transatlantic Working Group 2020) 8 <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 3 March 2025.

115 Committee of Ministers, 'Recommendation CM/Rec(2020)1 to member States on the human rights impacts of algorithmic systems' (adopted 8 April 2020 at the 1373rd meeting of the Ministers' Deputies) C.3.2.

116 A token is a contiguous sequence of characters with a semantic meaning. See Aggarwal, C.C. *Machine Learning for Text*. New York: Springer, 2022, 21.

117 Common prepositions, conjunctions, pronouns, and articles are considered stop words.

document[118]. This allows the system to detect the words that appear most frequently in defamatory messages and predict new ones. Nonetheless, a significant limitation of BoW is that it does not consider the order of words or their relationship to the rest of the document. Words that may be offensive in isolation can have a harmless meaning when combined with other words in the document.

62 Using neural networks to represent words as numerical vectors in a multidimensional space is also possible. Methods such as word embedding capture the semantic information of words, so similar terms, related terms or terms with the same connotation are placed close together in the vector space[119]. Word embedding also has its limitations, as it reproduces implicit biases in the corpus[120]. Bolukbasi *et al.* demonstrated that word embedding reproduces gender stereotypes by associating professions such as architect, economist, philosopher, computer programmer, pilot or captain with men, and housewife, nurse, receptionist, librarian, hairdresser or nanny with women[121]. Likewise, Caliskan *et al.* found that the concepts most associated with men include areas such as technology, engineering, religion and sports; while the concepts most associated with women include areas such as beauty, cooking, fashion and luxury[122].

c.) System Training

63 Algorithmic content moderation systems are generally trained through supervised learning, meaning that the system learns from labelled data to generalize or infer their common characteristics to classify new content. To train a system to predict defamatory content, the nodes of the artificial neural network can consider factors such as whether a message contains insults, comes from an anonymous user, or is directed at a specific person. Each parameter is assigned a weight based on its importance, with a higher weight indicating a more significant influence on the final result. When a node receives a value through its input connections, it multiplies it by the weight associated with each variable. If the result exceeds a certain threshold, the information passes through the output connections until it reaches the final layer, where the system provides its ultimate prediction.

64 Nonetheless, artificial neural networks do not offer explanations for their outputs, a phenomenon known as the black box problem[123]. As Burrell points out: "When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension. Machine optimizations based on training data do not naturally accord with human semantic explanations"[124]. Understanding how the algorithm interacts with the learning environment to get the final prediction, even when the input variables are known, is extremely complex[125]. Moreover, decision-making is obscured by a code typically protected by intellectual property rights[126].

d.) Evaluation

65 The final step before implementing a prediction system is its evaluation. Classification errors can lead to false positives, i.e. identifying content that is lawful as unlawful[127], and false negatives, i.e., identifying as lawful content that is unlawful. False positives lead to over-blocking, while false negatives lead to under-blocking[128].

118 J Eisenstein, Introduction to Natural Language Processing (MIT Press 2019) 13–16.

119 D Jurafsky and JH Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd edn 2023) 105–108.

120 H Gonen and Y Goldberg, 'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them' (2019) <arXiv:1903.03862> accessed 23 June 2025; A Caliskan, JJ Bryson and A Narayanan, 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases' (2017) 356(6334) *Science* 183–186.

121 T Bolukbasi, KW Chang, J Zou *et al*, 'Man is to computer programmer as woman is to homemaker? Debiasing word embeddings' (2016) <arXiv:1607.06520> accessed 23 June 2025.

122 A Caliskan, PP Ajay, T Charlesworth *et al*, 'Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics' (2022) <arXiv:2206.03390> accessed 23 June 2025.

123 D Castelvecchi, 'The Black Box of AI' (2016) 538 *Nature* <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731> accessed 23 June 2025.

124 J Burrell, 'How the machine "thinks": Understanding opacity in machine learning algorithms' (2016) 3(1) *Big Data & Society* 10.

125 C Rudin and J Radin, 'Why are we using Black Box Models in AI when we don't need to? A lesson from an Explainable AI competition' (2019) 1(2) *Harvard Data Science Review* 3.

126 M Maggiolino, 'EU Trade Secrets Law and Algorithmic Transparency' (Bocconi Legal Studies Research Paper No 3363178, 2019) 6–9.

127 F Reda, 'When filters fail: These cases show we can't trust algorithms to clean up the internet' <https://felixreda.eu/2017/09/when-filters-fail/> accessed 23 June 2025.

128 Four types of measures are usually used to assess the performance of a system: accuracy, precision, recall and specificity. See G Sartor, A Loreggia, The impact of algorithms for online content filtering or moderation.

66 Machine learning systems are based on probabilistic methods, so errors cannot be avoided[129]. The error rate is higher when the unlawfulness depends on language nuances and social and cultural particularities. As Bender and Koller claim: "In contrast to some current hype, meaning cannot be learned from form alone[130]. This means that even large language models (...) do not learn meaning; they learn some reflection of meaning into the linguistic form"[131]. Nowadays, the highest accuracy rates of automatic offensive language detection systems do not exceed 80%[132]. Even the most advanced large language models, such as GPT-4o, have limitations in terms of context understanding[133].

67 Principle 1 of the Santa Clara Principles on Transparency and Accountability in Content Moderation recommends that companies use automatic content moderation systems only when there is sufficiently high confidence in the quality and accuracy of those processes[134]. In a similar vein, Recital 26 DSA states that online intermediaries should take reasonable measures to ensure that, where automated tools are used to detect, identify and act against illegal content, the relevant technology is sufficiently reliable to limit the rate of errors to the maximum extent possible.

68 In summary, AI systems cannot achieve accurate outcomes when content decisions require a high degree of contextual understanding[135]. As observed in the European Parliament resolution of 20 October 2020 on the Digital Services Act and fundamental rights issues posed: "Current automated tools are not capable of critical analysis and of adequately grasping the importance of context for specific pieces of content, which could lead to unnecessary takedowns and harm the freedom of expression and the access to diverse information, including on political views, thus resulting in censorship"[136]. Similarly, in *Poland v Parliament and Council*, the CJEU stressed that: "A filtering system which might not distinguish adequately between unlawful content and lawful content (...) would be incompatible with the right to freedom of expression and information, guaranteed in Article 11 of the Charter, and would not respect the fair balance between that right and the right to intellectual property"[137].

69 For the above reasons, user notifications should not trigger a stay-down obligation that is an obligation to prevent the reappearance of previously notified defamatory content. This is without prejudice to any injunction that may be issued in a specific case ordering the prevention of identical or similar infringements in line with the Glawischnig-Piesczek doctrine.

Upload filters. 2020, 45 <https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)657101> accessed 23 June 2025.

129 E Douek, Content moderation as systems thinking. *Harvard Law Review*. 2022, 136(2), 552 ("Error choice is baked in at the moment of *ex ante* system design and depends on a number of factors including the importance of speed, an assessment of the level of risk in a particular context, and the level of technological capacity for moderating a certain kind of content").

130 The authors define "form" as any observable realisation of language, such as marks on a page, pixels or bytes in a digital representation of text, or movements of the articulators.

131 EM Bender and A Koller, 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data' in D Jurafsky *et al* (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*(Association for Computational Linguistics 2020) 5193.

132 F Zufall, M Hamacher, K Kloppenborg *et al*, 'A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task' (2021) 7 <arXiv:2004.03422> accessed 23 June 2025; A Zagidullina, G Patoulidis and J Bokstaller, 'Model Bias in NLP: Application to Hate Speech Classification Using Transfer Learning Techniques' (2021) 8–11 <arXiv:2109.09725> accessed 23 June 2025.

133 E Vargas Penagos, 'ChatGPT, can you solve the content moderation dilemma?' (2024) 32 *International Journal of Law and Information Technology* 25–26.

134 'The Santa Clara Principles' <https://santaclaraprinciples.org/> accessed 23 June 2025.

135 A Marsoof, A Luco, H Tan *et al*, 'Content-filtering AI systems – limitations, challenges and regulatory approaches' (2023) 32(1) *Information & Communications Technology Law* 83.

136 European Parliament, 'Resolution of 20 October 2020 on the Digital Services Act and fundamental rights issues posed' 2020/2022(INI) para 12.

137 *Poland* (n 61) para 86. See also JP Quintais, C Katzenbach and SF Schwemer *et al*, 'Copyright Content Moderation in the European Union: State of the Art, Ways Forward and Policy Recommendations' (2024) 55 *International Review of Intellectual Property and Competition Law* 17.

D. Remedies for Non-Compliance with Due Diligence Obligations

I. Public and Private Enforcement of the DSA

70 Chapter IV of the DSA contains a set of provisions on supervision and enforcement by the competent public authorities. Digital Services Coordinators (Article 49.2 DSA) have investigative and enforcement powers (Article 51 DSA), including the power to impose fines (Article 52 DSA), in respect of conduct by providers of intermediary services falling within the competence of their Member State[138]. Digital Services Coordinators may exercise those powers on their own initiative or following a request pursuant to Article 53 DSA: "Recipients of the service and any body, organization or association mandated to exercise the rights conferred by this Regulation on their behalf[139] shall have the right to lodge a complaint against providers of intermediary services alleging an infringement of this Regulation with the Digital Services Coordinator of the Member State where the recipient of the service is located or established"[140].

71 The Member State in which the main establishment of the provider of intermediary services is located has, in general, exclusive powers to supervise and enforce the DSA (Article 56.1 DSA)[141]. However, the powers of supervision and enforcement of due diligence obligations against providers of very large online platforms (hereinafter VLOP) and of very large online search engines (hereinafter VLOSE)[142] are shared by the European Commission and by the national competent authorities (Article 56.3 DSA)[143], and the former has exclusive powers of supervision and enforcement of the additional obligations to manage systemic risks imposed on these providers (Article 56.2 DSA)[144].

72 The European Commission may initiate proceedings against a provider of a VLOP or a VLOSE if it suspects it has infringed any of the provisions of the DSA (Article 66.1 DSA). If the European Commission finds that the provider does not comply with one or more provisions of the DSA, it will adopt a non-compliance decision (Article 73.1 DSA) and may impose fines not exceeding 6 % of the provider's total worldwide annual turnover (Article 74.1 DSA). To date, the European Commission has initiated formal proceedings against AliExpress[145], Facebook/Instagram[146], Temu[147], TikTok[148], and X[149].

73 Public enforcement addresses a collective action

138 'Digital Services Coordinators' <https://digital-strategy.ec.europa.eu/en/policies/dsa-dscs> accessed 23 June 2025.

139 Article 86 DSA. Rademacher argues that the *ius standi* should be extended to all parties negatively affected by an alleged infringement of a provider against provisions of the DSA, including notifiers. See T Rademacher, Article 53 Right to lodge a complaint. In B Hofmann/F Raue, (dirs.). *Digital Services Act: Article-by-article commentary*. Baden-Baden: Nomos. 2024, 937.

140 Recitals 118-119 DSA.

141 Recital 123 DSA.

142 The European Commission has designated as VLOPs: Alibaba AliExpress, Amazon Store, Apple AppStore, Booking.com, Facebook, Google Play, Google Maps, Google Shopping, Instagram, LinkedIn, Pinterest, Pornhub, Shein, Snapchat, Stripchat, Temu, TikTok, Wikipedia, X, XNXX, XVideos, YouTube and Zalando; and as VLOSEs: Bing and Google Search.

143 Recital 125 DSA.

144 I Buri, 'A Regulator Caught Between Conflicting Policy Objectives: Reflections on the European Commission's Role as DSA Enforcer' *VerfBlog* (31 October 2022) <https://verfassungsblog.de/dsa-conflicts-commission/> accessed 23 June 2025.

145 European Commission, 'Commission opens formal proceedings against AliExpress under the Digital Services Act' (14 March 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1485> accessed 23 June 2025.

146 European Commission, 'Commission opens formal proceedings against Meta under the Digital Services Act related to the protection of minors on Facebook and Instagram' (16 May 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664> accessed 23 June 2025; European Commission, 'Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act' (30 April 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373> accessed 23 June 2025.

147 European Commission, 'Commission opens formal proceedings against Temu under the Digital Services Act' (31 October 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_5622> accessed 23 June 2025.

148 European Commission, 'Commission opens formal proceedings against TikTok on election risks under the Digital Services Act' (17 December 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487> accessed 3 March 2025; European Commission, 'Commission opens proceedings against TikTok under the DSA regarding the launch of TikTok Lite in France and Spain, and communicates its intention to suspend the reward program in the EU' (22 April 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2227> accessed 3 March 2025; European Commission, 'Commission opens formal proceedings against TikTok under the Digital Services Act' (19 February 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926> accessed 23 June 2025.

149 European Commission, 'Commission sends preliminary findings to X for breach of the Digital Services Act' (12 July 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761> accessed 23 June 2025.

problem, as victims may lack sufficient incentive to pursue private actions when the impact of an infringement is minimal[150]. Nevertheless, public enforcement is inherently selective due to the limited resources of supervisory authorities[151]. As Husovec points out: "Only practice will tell how the European Commission will exercise its competence in cases when VLOPs or VLOSEs violate standard due diligence obligations. It is likely that the resource-limited Commission will prioritize cases based on their importance"[152]. The limitations of public enforcement can be mitigated by private claims[153], mainly through the claim for compensation provided for in Article 54 DSA.

II. The Right to Compensation under Article 54 DSA

74 Where the conditions for applying the safe harbor provided for in Article 6.1 DSA are not met, platforms may be held liable for hosting defamatory content in accordance with each Member State's rules on tort liability.

75 When the damage is the result of non-compliance with the due diligence obligations regulated in Chapter III of the DSA, the victim can also opt for the compensation remedy provided for in Article 54: "Recipients of the service shall have the right to seek, in accordance with Union and national law[154], compensation from providers of intermediary services, in respect of any damage or loss suffered due to an infringement by those providers of their obligations under this Regulation". As indicated by Raue, this article establishes an imperfect liability rule because it lacks provisions on the subjective requirements for damages, the burden of proof, or defenses such as the statute of limitations[155].

76 Under Article 54 DSA, victims must demonstrate: being a recipient of an intermediary service[156], the infringement of any due diligence obligations of the DSA, the existence of fault of the platform's operator[157], the existence of damages, and the causal link between the infringement and the damage.

1. Infringement of Due Diligence Obligations

77 Non-compliance with certain due diligence obligations under the DSA may result in harm affecting a user's right to honor. This may occur when platforms fail to suspend the provision of their services to users who frequently provide defamatory content (Article 23 DSA) or when they do not designate a single point of contact to enable users to communicate directly and rapidly with them (Article 12 DSA). In this sense, before the DSA, the judgement 72/2011, of 10 February, of the Spanish Supreme Court confirmed the liability of the owner of a website for hosting defamatory messages on the grounds that the illegality was evident, and that the defendant had failed to comply with the obligation to designate a means of contact[158]. The infringement of this obligation prevented the plaintiff from being able to communicate with the defendant in an easy and direct manner to stop the dissemination of the defamatory content[159].

150 S Shavell, Liability for Harm versus Regulation of Safety. *The Journal of Legal Studies*. 13 (2) 1984, 363 ("One reason that a defendant can escape tort liability is that the harms he generates are widely dispersed, making it unattractive for any victim individually to initiate legal action").

151 A Rubí Puig, 'Problemas de coordinación y compatibilidad entre la acción indemnizatoria del artículo 82 del Reglamento General de Protección de Datos y otras acciones en derecho español' (2018) 34 *Derecho Privado y Constitución* 209.

152 M Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024) 424; D Jackson and B Szóka, 'The Far Right's War on Content Moderation Comes to Europe' *TechPolicy.press* (11 February 2025) <https://www.techpolicy.press/the-far-rights-war-on-content-moderation-comes-to-europe/> accessed 23 June 2025.

153 Z Clopton, 'Redundant Public-Private Enforcement' (2016) 69 *Vanderbilt Law Review* 308–311.

154 Recital 121 DSA clarifies that: "Such compensation should be in accordance with the rules and procedures set out in the applicable national law and without prejudice to other possibilities for redress available under consumer protection rules". The detailed procedural rules governing actions for safeguarding an individual's rights under UE law must be no less favorable than those governing similar domestic actions (principle of equivalence) and must not render practically impossible or excessively difficult the exercise of rights conferred by EU law (principle of effectiveness).

155 F Raue, 'Article 54 Compensation' in B Hofmann and F Raue (eds), *Digital Services Act: Article-by-Article Commentary*(Nomos 2024) 951.

156 Recipient of the service means any natural or legal person who uses an intermediary service, in particular for the purposes of seeking information or making it accessible (Article 3 b) DSA).

157 Within the European legal systems, fault-based liability provides the backbone of the law of torts. See G Wagner, Liability Rules for the Digital Age, *Journal of European Tort Law*, 13(3) 2022, 194.

158 See Article 5 ECD and Article 10 of the Law 34/2002 of 11 July 2002 on information society services and electronic commerce. The latter transposed the ECD into Spanish law.

159 Judgment 72/2011 (Supreme Court (Civil Chamber) 10 February 2011) para 4 (ECLI:ES:TS:2011:559).

78 The majority of DSA's due diligence obligations appear to confer rights which can be violated by a single act of non-compliance. For instance, Article 17.1 DSA obliges hosting service providers to provide a clear and specific statement of reasons to any affected user for any restriction imposed on their content. The statement of reasons shall at least contain, among other information, a reference to the legal ground relied on and explanations as to why the information is considered to be illegal content on that ground (Article 17.3 d) DSA), and clear and user-friendly information on the possibilities for redress available to the recipient of the service in respect of the decision (Article 17.3 f) DSA). Failure to comply with this provision may prevent the user from realizing that their content has been removed and, consequently, from being able to complain about the decision in time. Other DSA's due diligence obligations require an assessment of the platform's behavior on a systemic level to be able to establish violations (e.g., Articles 20.4, 21.2, 22.2, or 23 DSA)[160].

79 It is unclear whether Articles 34 and 35 DSA can be enforced through private actions[161]. In accordance with these provisions, providers of VLOPs and VLOSEs must diligently identify, analyze and assess any systemic risks in the Union, including the dissemination of illegal content through their services (Article 34.1 a) DSA), stemming from the design of their recommender systems and any other relevant algorithmic system (Article 34.2 a) DSA). After the risk assessment, the above-mentioned subjects must put reasonable, proportionate, and effective mitigation measures in place, with particular consideration given to the impact on fundamental rights. Such measures may include testing and adapting their recommender systems (Article 35.1 d) DSA).

80 Following the well-established case law of the CJUE, individuals who have been harmed have a right to compensation where the rule of EU law infringed intended to confer rights on them, and those rights arise not only where provisions of EU law expressly grant them, but also by reason of positive or negative obligations which those provisions impose in a precise, clear and unconditional manner, whether on individuals, on the Member States or on the EU institutions[162]. The problem is that Articles 34-35 DSA grant the providers of VLOP and VLOSE and the Commission broad discretion. Therefore, as Husovec notes: "Prior to the Commission concretizing what risk mitigation measures are appropriate given the practice of the provider, it is hard to infer specifically what an individual can personally expect from such rules"[163].

2. Fault

81 Article 82 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (hereinafter GDPR) provides a right to compensation to any person who has suffered material or non-material damage as a result of an infringement of the GDPR.

82 This provision, like Article 54 DSA, does not specify the subjective requirements for damages. Some authors argue that Article 82 GDPR does not require the existence of fault when establishing the liability of controllers or processors[164], while others maintain that a strict liability rule should apply in cases of infringement of obligations of result[165]. However, the CJEU has held that it is apparent from a combined analysis of Articles 82.2 and 82.3 that this article provides for a fault-based regime, in which the controller is presumed to have participated in the processing constituting the breach of the GDPR in question[166]. Article 54 DSA is much narrower than Article 82 GDPR, so it is uncertain whether the CJEU will reach the same conclusion.

83 It should be noted that, unlike Article 54 DSA, Article 74 DSA does include a reference to the requirement

160 M Husovec, Principles of the Digital Services Act (Oxford University Press 2024) 434.

161 In favor of this possibility, see F Raue, Article 54 Compensation. In B Hofmann/F Raue (dirs.). *Digital Services Act: Article-by-article commentary*. Baden-Baden: Nomos. 2024, 958.

162 The CJUE established the principle of direct effect of EU law in Case C-26/62 *van Gend & Loos* (5 February 1963).

163 M Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024) 431. See also M del Moral Sánchez, 'The Devil is in the Procedure: Private Enforcement in the DMA and the DSA' (2024) 9(1) *University of Bologna Law Review* 33.

164 G Zanfir-Fortuna, 'Article 82. Right to compensation and liability' in C Kuner, L Bygrave and C Docksey (eds), *The EU General Data Protection Regulation* (Oxford University Press 2020) 1176.

165 MJ Santos Morón, 'Reflexiones en torno a la jurisprudencia del TJUE sobre la acción indemnizatoria del art 82 RGPD (asuntos C-300/21; C-340/21; C-456/22; C-667/21; C-687/21; C-741/21)' (2024) 16(2) *Cuadernos de Derecho Transnacional* 1420; A Rubí Puig, 'Daños por infracciones del derecho a la protección de datos personales. El remedio indemnizatorio del artículo 82 RGPD' (2018) 5(4) *Revista de Derecho Civil* 62–63.

166 Case C-667/21 *Krankenversicherung Nordrhein* (CJEU, 21 December 2023) para 103; Case C-687/21 *MediaMarktSaturn*(CJEU, 25 January 2024) para 52; Case C-741/21 *Iuris* (CJEU, 11 April 2024) para 46.

of fault when it establishes that: "The Commission may impose on the provider of the very large online platform or of the very large online search engine concerned fines not exceeding 6 % of its total worldwide annual turnover in the preceding financial year where it finds that the provider, *intentionally or negligently*: (a) infringes the relevant provisions of this Regulation (...)".

3. Damages

84 Since Article 54 DSA does not contain any provision intended to define the rules on the assessment of damages, it is for the legal system of each Member State to prescribe the criteria for determining the extent of the compensation payable in that context, subject to compliance with principles of equivalence and effectiveness[167].

85 When the infringement of the DSA's due diligence obligations affects the right to honor, the extent of non-pecuniary losses should be assessed considering factors such as the number of views of the defamatory content and the duration it remained publicly accessible[168]. In online defamation cases, Spanish courts have awarded damages against online intermediaries ranging from 1000€ to 18000€ depending on these criteria[169].

86 For example, the judgement of the Court of Appeals of Malaga, section 4, 82/2018, of 5 February, ordered a news portal to pay compensation of 1200€ for hosting defamatory comments. The judgement took into account that the defamatory comments were a minority compared to the rest of the comments, that the number of users was not significant, and that the news item referred to a limited territorial scope (Marbella)[170]. In contrast, the judgement of the Court of Appeals of Murcia, section 1, 9/2020, of 13 January, ordered another news portal to pay compensation of 20000€ for hosting defamatory comments. The Court of Appeals considered that the comments received a total of 89462 visits in order to quantify non-pecuniary losses[171].

E. Concluding Remarks

87 Online platforms are liable for hosting defamatory content only once they have knowledge or awareness of its illegality, as harm is not foreseeable before that moment. They have reactive duties which, pursuant to Article 16 DSA, include implementing notices and take-down mechanisms to react rapidly against defamatory content. In contrast, platforms should not be subjected to proactive prevention duties, as this would entail general monitoring or active fact-finding obligations, both expressly prohibited under Article 8 DSA. Given the current state of the art, platforms should also not be required to prevent the reappearance of previously notified defamatory content. However, the lack of notice and stay-down obligations does not preclude courts from ordering measures to prevent the publication of identical or similar illegal content.

88 When platforms cannot benefit from the safe harbor, their liability for hosting defamatory content must be based on the Member State's rules on tort liability. Additionally, liability may be established under Article 54 DSA if the damage results from the platform's actions or omissions. In such cases, the victim must demonstrate that they are the recipient of an intermediation service, that the platform infringed one or more due diligence obligations under the DSA, the fault of the platform operator, the damage suffered, and a causal link between the infringement and the damage.

167 Case C-300/21 Österreichische Post (CJEU, 4 May 2023) para 54; Joined Cases C-182/22 and C-189/22 *Scalable Capital*(CJEU, 20 June 2024) para 27.

168 *Kozan v Turkey* no 16695/19 (ECtHR, 1 March 2022) para 66; *Sanchez* (n 75) para 193; *Danileț v Romania* no 16915/21 (ECtHR, 20 February 2024) para 76.

169 Judgment of the Court of Appeal of Madrid, sec 19, 50/2006, 6 February, ECLI:ES:APM:2006:266 (compensation of €18,000); Judgment of the Court of Appeal of Islas Baleares, sec 3, 65/2007, 22 February, ECLI:ES:APIB:2007:200 (compensation of €6,000); Judgment of the Court of Appeal of Madrid, sec 13, 420/2008, 22 September, ECLI:ES:APM:2008:18214 (compensation of €6,000); Judgment of the Court of Appeal of Badajoz, sec 3, 280/2010, 17 September, ECLI:ES:APBA:2010:871 (compensation of €2,000); Judgment of the Court of Appeal of Barcelona, sec 14, 707/2010, 29 November, ECLI:ES:APB:2010:8805 (compensation of €12,000); Judgment of the Court of Appeal of Madrid, sec 11, 221/2011, 31 March, ECLI:ES:APM:2011:2467 (compensation of €9,000); Judgment of the Court of Appeal of Málaga, sec 4, 540/2011, 24 October, ECLI:ES:APMA:2011:1605 (compensation of €10,000); Judgment of the Court of Appeal of Madrid, sec 12, 47/2015, 4 February, ECLI:ES:APM:2015:4445 (compensation of €10,000); Judgment of the Court of Appeal of Santa Cruz de Tenerife, sec 3, 1/2022, 13 January, ECLI:ES:APTF:2022:97 (compensation of €5,000).

170 The Spanish Supreme Court upheld this judgement. See Judgement 235/2020, of 2 June (ECLI:ES:TS:2020:1534).

171 The judgement of the Spanish Supreme Court 226/2021, of 27 April (ECLI:ES:TS:2021:1570) reduced the compensation to 15000€ considering that the right to privacy was not violated, only the right to honour.