Chapter IX

Freedom of Expression on Social Networks: An International Perspective

Dušan V. Popović

1. Legal aspects of content censorship on social networks

Social networks are omnipresent; yet, there is no generally accepted definition of them. In order to define 'social networks' for our current purposes, we have identified several common features of existing social media platforms, which are presented in the literature.[1] First, social networks are Web 2.0-based applications. The shift to Web 2.0 applications can be described as a shift from the user as a consumer to the user as a participant. These apps are designed to enable users to interact, create, and share content online. Second, user-generated content is the essential (but not exclusive) component of social networks. The notion of 'user generated content' is not limited to text, photos, or videos; it could well be a simple 'like.' Third, social networks connect user-specific profiles with those of other individuals or groups. User profiles are thus the pillars of every social network. The manner in which users identify themselves may vary, but every social network tracks users' Internet Protocol (IP) address. Given their similarities from a freedom of speech perspective, we shall take the same approach *stricto sensu* to social networks, such as Facebook or Twitter, and video-sharing portals, such as YouTube.

Analyzing the legal aspects of content censorship on social networks starts with the examination of the foundations of freedom of speech (Section 1.1), as well as the very

1 See for example: Obar and Wildman, 2015, pp. 745–750.

Dušan V. Popović (2021) Freedom of Expression on Social Networks: An International Perspective. In: Marcin Wielec (ed.) *The Impact of Digital Platforms and Social Media on the Freedom of Expression and Pluralism,* pp. 277–310. Budapest–Miskolc, Ferenc Mádl Institute of Comparative Law–Central European Academic Publishing.

https://doi.org/10.54237/profnet.2021.mwsm_9

notion of 'speech,' which is extensively interpreted in both an offline and online context (Section 1.2). In the first years following their creation, social networks have legally been considered private spaces. The next section examines whether they should be considered as public forums, given their social function (Section 1.3). The paper will also examine the legal basis for content censorship in comparative law. There are two main approaches to the regulation of social networks, which serve as models for other jurisdictions: the US and the EU models (Section 1.4). Further to government regulation of social networks, we witness different forms of internal rules and regulations adopted by social networks, such as terms of service, privacy policies, IP policies, and community standards (Section 1.5). However, there are two main downsides of such self-regulation: the loss of equal access to speech and the lack of accountability (Section 1.6).

1.1. The foundations of freedom of speech on the Internet

Freedom of speech allows ordinary people to participate in the spread of ideas. It undoubtedly represents an important element of democratic culture, in the sense that everyone, not only the political or cultural elite, has a chance to participate in public dialogue. Freedom of speech is interactive, since exposure to someone else's ideas influences and potentially reshapes us. Freedom of speech is also appropriative in the sense that every participant relies on, draws ideas from, and modifies and/or criticizes the existing cultural background.

The theoretical foundations of freedom of speech can be categorized in different ways.[2] Freedom of speech may be understood as a means of truth discovery. According to John Stuart Mill, the recognition of truth is a prerequisite of social development. Therefore, the limitation of freedom of speech is inadmissible, since the restricted opinion may carry the truth.[3] On the other hand, freedom of speech can be seen as an instrument of democratic self-government. According to Alexander Meiklejohn and many others, freedom of speech enables the proper operation of society. Another line of thought sees freedom of speech as a value in itself—a right to which every citizen is entitled. Ronald Dworkin is a notable representative of this individualist theory.

These theories are reflected in the case law of national and supranational courts. In the United States, the US Supreme Court adopted a landmark decision in 1964 in the case *New York Times Co. v. Sullivan*, restricting public officials' ability to sue for defamation.[4] Specifically, the court held that if a plaintiff in a defamation lawsuit is a public official or a person running for public office, not only must they prove the normal elements of defamation, i.e., publication of a false defamatory statement to a third party, they must also prove that the statement was made with actual malice, meaning that the defendant either knew the statement was false or recklessly disregarded its

2 For a more detailed analysis of different theoretical justifications of freedom of speech, see: Koltay, 2019, pp. 8–15.

3 John Stuart Mill laid down the foundations of freedom of speech in his essay *On Liberty* (1859).

4 US Supreme Court, *New York Times Co. v. Sullivan*, 376 U.S. 254 (1964).

veracity. On the other side of the Atlantic, European (national) courts' case law is under the significant influence of the views and interpretations expressed by the European Court of Human Rights (ECtHR). The right to freedom of expression, guaranteed under Article 10 of the European Convention on Human Rights,[5] is interpreted to include the right to freely express opinions, views, and ideas, and seek, receive, and impart information regardless of frontiers. Freedom of expression is applicable not only to information or ideas that are favorably received or regarded as inoffensive, but also to those that may offend or disturb. In its landmark decision in *Handyside v. the United Kingdom,* the ECtHR defined freedom of expression as one of the essential foundations of a democratic society and a basic condition for its progress and for the development of every man.[6] As noted in the Council of Europe's Guide to Human Rights for Internet Users[7] and its explanatory memorandum, the ECtHR has affirmed in its jurisprudence that Article 10 is fully applicable to the Internet.[8] Member states have a primary duty, pursuant to Article 10 ECHR, not to interfere with the communication of information between individuals, be they legal or natural persons.

The global expansion of the Internet has provided a means by which free speech can reach broader audiences than ever before. The Internet's technological superiority and affordability facilitate citizens' participation in information exchange. However, the majority of Internet users exercise their right to freedom of expression anonymously, which can lead to certain abuses or even criminal offenses that could *de facto* be impossible to persecute.

1.2. The concept of speech

International and national legal documents do not use uniform terminology to designate the right to participate in public debate. The First Amendment of the United States Constitution, adopted in 1791, employs the term 'freedom of speech:'

> Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.

It has been heavily debated whether the free speech and free press clauses are coextensive or whether one reaches where the other does not. Justice Stewart argued that the fact that the First Amendment speaks separately of freedom of speech and

5 Council of Europe, Convention for the Protection of Human Rights and Fundamental Freedoms, 1950.

6 ECtHR, *Handyside v. the United Kingdom,* 7 December 1976, § 49.

7 Council of Europe, Recommendation of the Committee of Ministers to Member States on a Guide to Human Rights for Internet users, CM/Rec(2014)6, 16 April 2014.

8 See for example: ECtHR, *Perrin v. the United Kingdom*, 18 October 2005; ECtHR, *Renaud v. France*, 25 February 2010; ECtHR, *Editorial Board of Pravoye Delo and Shtekel v. Ukraine*, 5 May 2011.

freedom of the press is no accident, but an acknowledgment of the critical role the press plays in US society. In his view, the Constitution requires sensitivity to that role and to the press's special needs in performing it effectively.[9] However, contemporary interpretations of the First Amendment analyze the speech and press clauses under an umbrella 'freedom of expression' standard. The French Declaration of the Rights of Man and of the Citizen (*Déclaration des droits de l'homme et du citoyen*), adopted in 1789, employs the term 'freedom to express thoughts and opinions:'

> The free communication of thoughts and opinions is one of the most precious of the rights of man. Every citizen may, accordingly, speak, write, and print with freedom, but shall be responsible for such abuses of this freedom as shall be defined by law.

More recently adopted legal documents employ the term 'freedom of expression' rather than 'freedom of speech.' For example, the Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR), adopted in 1948 and 1966 respectively, both state that individuals have a right to freedom of expression; this right includes the freedom to seek, receive, and impart information and ideas of all kinds.[10] The European Convention on Human Rights also employs the term 'freedom of expression.'[11]

The concept of 'freedom of speech' has been interpreted extensively, so as to include not only direct speech (words) but also symbolic speech (actions). In the United States, the freedom of speech includes *inter alia* the right not to speak,[12] the right to use certain offensive words and phrases to convey political messages,[13] the right to advertise commercial products and professional services,[14] and the right to burn the flag in protest.[15] The ECtHR also considers 'freedom of expression' to cover both direct and symbolic speech. For instance, the Court found that freedom of expression includes artistic expression such as a painting,[16] the production of a play,[17] and information of a commercial nature.[18] With regard to the so-called 'negative right' not to express oneself, the ECtHR does not rule out that such a right is protected under the European Convention on Human Rights, but it has found that this issue should be addressed on a case-by-case basis.[19] Specifically in the context of the Internet, the ECtHR has emphasized that Art. 10 of the Convention is to apply to communication

9 *Houchins v. KQED*, 438 U.S. 1, 17 (1978) (concurring opinion).
10 UDHR, art. 19; ICCPR, art. 19.
11 ECHR, art. 10.
12 *West Virginia Board of Education v. Barnette*, 319 U.S. 624 (1943).
13 *Cohen v. California,* 403 U.S. 15 (1971).
14 *Bates v. State Bar of Arizona*, 433 U.S. 350 (1977).
15 *Texas v. Johnson*, 491 U.S. 397 (1989); *United States v. Eichman*, 496 U.S. 310 (1990).
16 ECtHR, *Müller and Others v. Switzerland*, 24 May 1988.
17 ECtHR, *Ulusoy and Others v. Turkey,* 25 June 2019.
18 ECtHR, *Casado Coca v. Spain,* 24 February 1994.
19 ECtHR Guide, 2020, p. 14.

on the Internet, whatever the type of message being conveyed and even when the purpose is profit making in nature.[20]

The Internet has undoubtedly introduced new forms of communication, i.e., new forms of opinion expression. For example, a 'like' on a social network is a form of speech, as it represents an Internet user's statement. This was established in the case *Bland v. Roberts*, where a public sector employee sued because he was fired for clicking the Facebook 'like' button on his employer's re-election rival's campaign website. The judge dismissed the free speech claim stating that 'liking' web content is not 'sufficient' speech to warrant constitutional protection. However, the Fourth Circuit reversed the decision on the First Amendment issue, holding that:

> On the most basic level, clicking on the 'like' button literally causes to be published the statement that the user 'likes' something, which is itself a substantive statement. In the context of a political campaign's Facebook page, the meaning that the user approves of the candidacy whose page is being liked is unmistakable. That a user may use a single mouse click to produce that message that he likes the page instead of typing the same message with several individual key strokes is of no constitutional significance.[21]

The court also noted that the act of 'liking' a page itself results in an affirmative statement made by a Facebook user to their friends. Consequently, choosing to 'like' something on Facebook produces speech.[22]

The US courts also held that the First Amendment protects as 'speech' the results produced by an Internet search engine. In *Search King, Inc. v. Google Technology, Inc.*, the court concluded that Google's page rankings were subjective results that constituted 'constitutionally protected opinions' entitled to full constitutional protection.[23] Likewise, in *Langdon v. Google, Inc.*, the court refused to order Google and Microsoft to prominently list the plaintiff's site in their search results, reasoning that:

> The First Amendment guarantees an individual the right to free speech, 'a term necessarily comprising the decision of both what to say and what not to say.' (…) The injunctive relief sought by plaintiff contravenes defendants' First Amendment rights.[24]

Just as newspapers cannot be forced to print editorial content or advertising, the court held that search engines cannot be forced to include links that they wish to exclude. This full protection remains when the choices about how to select and arrange the material are implemented with the help of computerized algorithms.[25]

20 ECtHR, *Ashby Donald and Others v. France,* 10 January 2013.
21 *Bland v. Roberts*, No. 12-1671, 4th Cir., 18 September 2013.
22 For an extensive analysis of *Bland v. Roberts* case see: Sarapin and Morris, 2014, pp. 131-157.
23 No. CIV-02-1457-M, 2003 WL 21464568, at *4, W.D. Okla. 27 May 2003.
24 474 F. Supp. 2d 622, 629–30 (D. Del. 2007) (citing *Riley v. National Fed'n of the Blind of N.C., Inc.*, 487 U.S. 781, 796–97 (1988); *Miami Herald Pub'g Co. v. Tornillo*, 418 U.S. 241, 256 (1974).
25 Volokh and Falk, 2012, pp. 886–887.

The US legal system differentiates among several categories of speech, some of which do not fall under the freedom of speech protection. The following categories of speech are given lesser or no protection by the First Amendment: obscenity, fighting words, defamation (including libel and slander), child pornography, perjury, blackmail, incitement to imminent lawless action, true threats, solicitations to commit crimes, and plagiarism of copyrighted material. Contrary to the US legal system, the European (national) legal systems and the European Convention on Human Rights do not introduce categories of speech. Instead, they prescribe different limitations on the freedom of speech, such as protection against defamation or speech interfering with the intimate and private sphere, the maintenance of public order and national security, the protection of consumers against misleading commercial messages, the protection of children against materials that are harmful to their development, and the protection of certain social groups against hatred.[26]

1.3. Social networks as a public forum?

Since their inception, social networks such as Facebook and Twitter have been legally considered as private spaces. However, in recent years, social networks are increasingly being perceived as forums of public communication. In line with this tendency, the US courts examined whether the public forum doctrine could be applied to social networks. The nuances of the public forum doctrine were articulated in the case *Perry Education Association v. Perry Local Educators' Association* in 1983.[27] Justice Byron R. White explained three categories of government property for the purposes of access for expressive activities: (1) traditional or quintessential public forums, (2) limited or designated public forums, and (3) non-public forums. According to the public forum doctrine, the government can impose reasonable time, place, and manner restrictions on speech in all three property categories but has limited ability to impose content-based restrictions on traditional or designated public forums.

Nowadays, many politicians choose to set up official Facebook, Twitter, and Instagram accounts to communicate with citizens. These accounts are used for official purposes. Should these social network accounts be perceived as a public forum? In *Knight First Amendment Inst. at Columbia Univ. v. Trump,*[28] a group of seven citizens, represented by the Knight First Amendment Institute, sued US President Trump. Their complaint alleged that when President Trump blocked them on Twitter, he engaged in viewpoint discrimination in a public forum, an action that would violate

26 In certain situations, the ECtHR does not even examine the compatibility of a limitation with Art. 10 of the European Convention on Human Rights. This happens when the ECtHR finds an abuse of the freedom of speech, within the meaning of Art. 17 of the Convention. See: Koltay, 2019, p. 20.

27 460 U.S. 37 (1983).

28 302 F. Supp. 3d 541 (S.D.N.Y. 23 May 2018).

the First Amendment's freedom of speech guarantee. President Trump argued that because this was his private account,[29] created in 2009, it was not subject to First Amendment claims. In 2019, the 2nd and 4th Circuit Courts of Appeals ruled that government use of social media creates a designated public forum, and government officials cannot engage in viewpoint discrimination by blocking comments.[30] The Court found that President Trump violated the First Amendment by removing several individuals who were critical of him and his governmental policies from the 'interactive space' of his Twitter account. The appeals court agreed with the lower court that the interactive space associated with Trump's Twitter account is a designated public forum and that blocking individuals because of their political expression constitutes viewpoint discrimination.[31]

From a freedom of expression perspective, it is particularly relevant to determine whether social networks should be treated as tech or media companies. Social networks, such as Facebook, have repeatedly insisted that their service is a neutral tech platform, not a publisher or a media company. A publisher, after all, could be expected to make factual and qualitative distinctions, and might be responsible, reputationally or legally, for the content it publishes, whereas a platform is nothing but empty space. However, in court proceedings in the United States, when Facebook was sued by an app startup that alleged that Mark Zuckerberg developed a 'malicious and fraudulent scheme' to exploit users' personal data and force rival companies out of business, Facebook's lawyers argued that decisions about what not to publish should be protected because Facebook is a publisher. Facebook's lawyers argued in court that the social network's decisions about data access were a 'quintessential publisher function' and constituted protected activity, adding that this includes both the decision of what to publish and the decision of what not to publish.[32]

If social networks are publishers, then the manner in which they select content results from editorial decisions and should be treated as 'speech.' In addition, if a social network has an opinion, than such an opinion could, under certain legally defined conditions, be restricted.

29 President Trump maintained only one Twitter account that he used for both private and official interactions with American citizens.

30 928 F. 3d 226 – Court of Appeals, 2nd Circuit 2019.

31 The petition for rehearing was denied on 23 March 2020. On 31 July 2020, the Knight Institute filed a second lawsuit in federal court against President Trump and his staff for continuing to block followers from the @realDonaldTrump Twitter account. On 5 April 2021, the Supreme Court vacated the judgment. The case has been remanded to the United States Court of Appeals for the Second Circuit with instructions to dismiss the case as moot, given that Donald Trump is a private citizen now.

32 Sam Levin, 'Is Facebook a publisher? In public it says no, but in court it says yes' *The Guardian* (3 July 2018) at https://www.theguardian.com/technology/2018/jul/02/facebook-mark-zuckerberg-platform-publisher-lawsuit.

1.4. Legal basis for content censorship in comparative law

Typically, liability for third-party content attaches when the disseminator has the discretion to publish it or not. If a disseminator cannot exercise editorial control, the disseminator is not legally responsible for third-party content it had to disseminate. In contrast, if the disseminator can exercise editorial control over the content, the disseminator accepts legal liability for the (editorial) decisions it makes. Online intermediaries, including social networks, do not entirely fit into either category. However, that does not mean that legislators have not imposed certain content-related obligations on them.

We shall analyze two approaches to the regulation of social networks, which serve as models to other jurisdictions: US and EU law. Our comparative analysis shall start with US law, since the United States is the Internet's birthplace. The US model protects intermediaries from liability for distributing third-party user content based on the 'Good Samaritan' rule, with the exception of certain laws: criminal law, intellectual property law, communications privacy law, and sex trafficking law. The US model could be seen as more favorable to online platforms than the EU's approach. The United States' neighboring countries and traditional economic partners follow its approach. For example, the US-Mexico-Canada agreement (USMCA, also known as NAFTA 2.0), concluded in 2018, requires Canada and Mexico to adopt protections in line with US legislation.[33] On the other hand, EU law provides liability exemption in favor of Internet intermediaries, concerning illegal content and activities online. The exemptions from liability only cover cases where the information society service provider's activity is limited to the technical operation process. The EU model is followed not only by EU member states, but also by other European countries that are candidates or potential candidates for EU membership.[34]

1.4.1. US law

The Communications Decency Act of 1996,[35] particularly Section 230, is the most important piece of US legislation related to online speech. The Act is the short name of Title V of the Telecommunications Act of 1996, as specified in Section 501 of the 1996 Act. Title V has affected the Internet and online communications in two significant ways. First, it attempted to regulate both indecency (when available to children) and obscenity in cyberspace. Second, Section 230 of the Communications Act of 1934 (Section 9 of the Communications Decency Act / Section 509 of the

33 Art. 19.17 of the USMCA: "No Party shall adopt or maintain measures that treat a supplier or user of an interactive computer service as an information content provider in determining liability for harms related to information stored, processed, transmitted, distributed, or made available by the service, except to the extent the supplier or user has, in whole or in part, created, or developed the information."

34 See for example: Republic of Serbia, Law on Electronic Commerce, *Official Journal* 41/2009, 95/2013 and 52/2019, Arts. 16–20.

35 47 U.S.C. § 230.

Telecommunications Act of 1996) has been interpreted to mean that operators of Internet services are not traditional publishers. Section 230(c)(1) reads: "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." There are three elements to this immunity. First, the immunity applies to a 'provider or user of an interactive computer service.' The courts have interpreted 'providers' extensively to include any service available through the Internet. Furthermore, 'users of interactive computer services' should cover all providers' customers. Second, the immunity applies to any claims that treat the defendant as a 'publisher' or 'speaker.' However, the courts usually interpret this element more extensively so that it applies regardless of whether the claim's *prima facie* elements contain the terms 'publisher' or 'speaker.' Third, immunity applies when the plaintiff's claim is based on information provided by another information content provider, i.e., by a third party.[36]

Section 230 immunity is not unlimited. It has four statutory exclusions where it is categorically unavailable. First, prosecutions of federal crimes (e.g., obscenity, sexual exploitation of children) are not immunized by Section 230. Second, Section 230 does not apply to plaintiffs' claims based on the Electronic Communications Privacy Act (ECPA)[37] or state law equivalents. Third, Section 230 does not apply to claims based on the Fight Online Sex Trafficking Act (FOSTA),[38] related to websites that unlawfully promote and facilitate prostitution and/or facilitate traffickers in advertising the sale of unlawful sex acts involving sex trafficking victims. Fourth, Section 230 does not apply to intellectual property claims. However, the courts differ in interpreting whether this exclusion applies only to federal intellectual property claims or also to state IP claims. In *Perfect 10 v. CCBill*, the Ninth Circuit held that the exclusion only applied to federal intellectual property claims.[39] However, courts outside the Ninth Circuit do not agree with the *CCBill* ruling, so state intellectual property claims are still viable in those jurisdictions.

When discussing the relationship between freedom of speech and IP rules in US law, one should bear in mind that there is also a specific 'notice and takedown' procedure related to copyrighted works, which was introduced by the Digital Millennium Copyright Act (DMCA).[40] This procedure allows a copyright owner to request the removal of content posted online. The DMCA shields online service providers from monetary liability and limits other forms of liability for copyright infringement—referred to as safe harbors—in exchange for cooperating with copyright owners to expeditiously remove infringing content if the online service providers meet certain conditions. Specifically, Subsection 512(c)(1)(A) of the DMCA requires that the service provider: (1)

36 For an overview of US case-law see: Balasubramani, 2016/2017, pp. 275–286.

37 18 U.S.C. §§ 2510–2523. The ECPA was significantly amended by the Communications Assistance to Law Enforcement Act (CALEA) in 1994, the USA PATRIOT Act in 2001, the USA PATRIOT Reauthorization Acts in 2006, and the FISA Amendments Act of 2008.

38 Public Law No: 115-164, 11 April 2018.

39 *Perfect 10, Inc. v. CCBill LLC*, 488 F.3d 1102, 9th Cir. 2007.

40 The DMCA safe harbors, codified at 17 U.S.C. § 512, are part of the Copyright Act.

does not have actual knowledge that the material or an activity using the material on the system or network is infringing; (2) in the absence of such actual knowledge, is not aware of facts or circumstances from which infringing activity is apparent; or (3) upon obtaining such knowledge or awareness, acts expeditiously to remove, or disable access to, the material. The DMCA has become a *de facto* global standard for addressing online copyright infringements, since the vast majority of removal requests are sent to global platforms that are US-based companies subject to the DMCA.

The DMCA offers Internet service providers protection from copyright liability if they expeditiously remove material in response to (essentially unverified) infringement complaints. Even if the accused poster responds with counter-notification of non-infringement,[41] the DMCA requires that the service provider keep the post offline for more than a week. Obviously, this procedure can be abused for censorship purposes. Indeed, the threat of secondary liability induces service providers to comply with the DMCA's notice and takedown provisions, making it more difficult for speakers to post material that challenges someone who can potentially make a copyright claim.[42] Since the notice and takedown procedures are implemented in a non-transparent way,[43] it is difficult to track such abuse. Moreover, because the notice and takedown procedures involve immediate removal but lack any legal oversight, there are no effective means to protect against abuse of the process. As long as the automatic enforcement system does not distinguish legitimate removal requests from non-copyright requests, there is great potential for misuse.[44] However, the DMCA does not impose a general filtering obligation, as the service provider is not required to block an allegedly infringing file from being re-uploaded to its service after the file has been taken down in response to a copyright owner's notice.[45]

1.4.2. EU law

The US DMCA legislation inspired the EU to enact the Directive on Electronic Commerce,[46] including safe harbors for mere conduits, caching, and hosting.[47] The

41 A mechanism that allows a user to contest the removal request.

42 Seltzer, 2010, p. 177.

43 The notice-and-takedown procedure is administered by private companies. Unlike copyright enforcement in court, where decisions are made public, we know very little about the actual implementation of the notice-and-takedown regime.

44 Bar-Ziv & Elkin-Koren, 2018, p. 377.

45 *UMG Recordings, Inc. v. Veoh Networks Inc.*, 665 F. Supp. 2d 1099, 1110 (C.D. Cal. 2009) at 1111: "UMG has not established that the DMCA imposes an obligation on a service provider to implement filtering technology (…)." However, some service providers have undertaken measures that exceed their legal obligations under the notice-and-takedown regime and voluntarily offer additional enforcement measures to copyright holders (e.g., YouTube's Content ID service). See also: Bridy, 2016, p. 192.

46 Directive 2000/31/EC of the European Parliament and of the Council on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, *Official Journal* L 178, 17.7.2000.

47 See Arts. 12-14 of the Directive on electronic commerce.

EU rules were modeled on the DMCA; however, they differ from the US safe harbor in two ways. First and most importantly, the directive's hosting provision governs all claims related to user-generated content, not just copyright. These claims may be derived from private law, in the form of, e.g., copyright infringement or defamation, as well as from criminal law, in the form of, e.g., incitement to violence or hate speech. Second, the notice and takedown mechanism is prescribed by a directive that allows for certain flexibility within national legislators and has resulted in 27 harmonized, albeit not identical, national legal regimes in EU member states.[48] The e-commerce directive additionally prohibits the imposition of general obligations on hosts that are protected by a safe harbor to monitor the information which they transmit or store, or to actively seek out facts or circumstances indicating illegal activity.[49]

As already noted in US case law, the expeditious removal of content may be (mis) used for censorship purposes. For that reason, the Court of Justice of the European Union in the *Promusicae* case[50] clarified that in transposing the directives and implementing the transposing measures "the Member States must (…) take care to rely on an interpretation of the directives which allows a fair balance to be struck between the various fundamental rights protected by the Community legal order."[51] This 'fair balance' doctrine was also accepted and further developed by the ECtHR, particularly in the decisions *Delfi v. Estonia*[52] and *MTE v. Hungary*.[53] Both cases concerned online hosts' liability for allegedly defamatory content posted by anonymous users in the comment sections below news articles published by the platforms. In *Delfi v. Estonia,* the ECtHR listed four specific factors to guide the balancing process: (1) the context of the comments, (2) the measures applied by the platform in order to prevent or remove the comments, (3) the liability of the actual authors of the comments as an alternative to the platform's liability, and (4) the consequences of the domestic proceedings for the platform.[54] In *MTE v. Hungary*, the Court added a fifth factor: the consequences of the comments for the victim.[55] In applying these factors to the two cases, the ECtHR came to two opposite conclusions. In *Delfi v. Estonia*, the comments were qualified as hate speech and incitement to violence. Thus, the imposition of liability on the hosting provider struck a fair balance and therefore did not entail a violation of the right to freedom of expression. However, in *MTE v. Hungary*,

48 Before Brexit – 28.

49 Art. 15 of the Directive on electronic commerce.

50 CJEU, case C-275/06, *Productores de Música de España (Promusicae) v Telefónica de España SAU* [2008] 2
CMLR 465.

51 *Ibid,* para 68. Note: Rights derived from international law are referred to as human rights, while rights derived from domestic national constitutional law, as well as from European law, are referred to as fundamental rights.

52 ECtHR, *Delfi v. Estonia*, 16 June 2015.

53 ECtHR, *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary* (hereinafter: *MTE v. Hungary*), 2 February 2016.

54 ECtHR, *Delfi v. Estonia*, para. 142.

55 ECtHR, *MTE v. Hungary,* paras. 68–69.

the Court characterized the comments as merely offensive and concluded that the liability imposed on the intermediaries for their dissemination violated the right to freedom of expression. Although the fair balance doctrine remains somewhat unclear at present, it allows for much needed flexibility in the area of intermediary liability.

As our analysis has shown, EU legislation initially limited the action expected of the intermediary to only one possibility—takedown—which applied horizontally, i.e., to all areas of law in which intermediary liability arises as a potential issue. However, the Directive on Copyright in the Digital Single Market,[56] adopted in 2019, made a subtle variation from the notice and takedown mechanism to the more flexible notice and action mechanism. Article 17 of the directive regulates 'online content-sharing service providers' (OCSSPs). These are defined as platforms with a profit-making purpose that store and give the public access to a large amount of user-uploaded works/subject matter, which they organize and promote. This includes well-known platforms like YouTube and Facebook, as well as any type of user-upload platform that fits this broad definition and is not expressly excluded, as is the case with electronic communication services, providers of business-to-business cloud services and cloud services, online marketplaces, not-for profit online encyclopedias (e.g., Wikipedia), not-for-profit educational and scientific repositories, and open source software developing and sharing platforms. The directive states that OCSSPs carry out acts of communication to the public when they give access to works/subject matter uploaded by their users. As a result, these platforms become directly liable for their users' uploads. They are also expressly excluded from the hosting safe harbor for copyright relevant acts previously available to many of them under the e-commerce directive. Consequently, the platforms have two possibilities to avoid direct liability. First, they could obtain authorization to communicate or make the user-uploaded content available. However, it seems almost impossible to obtain authorization for all user-uploaded content. Consequently, OCSSPs will have to rely on the second possibility, which allows them to avoid liability if they meet a number of cumulative conditions. They must demonstrate that they have: (1) made best efforts to obtain an authorization, (2) made best efforts to ensure the unavailability of specific works for which the right holders have provided them with the relevant and necessary information, and (3) acted expeditiously, subsequent to notice from right holders, to take down infringing content and made best efforts to prevent its future upload. These conditions have been criticized in legal theory,[57] especially the second condition, which appears to impose an upload filtering obligation, and the third condition, which introduces both a notice and takedown mechanism (already

56 Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, *Official Journal* L 130, 17.5.2019.

57 See for example: Quintais, 2020, pp. 28–41.

prescribed by the e-commerce directive) and a notice and stay down (or re-upload filtering) obligation.

In the interest of freedom of speech, the EU legislator created a special regime for certain copyright exceptions and limitations (quotation, criticism, caricature, review, parody, and pastiche).[58] However, existing content recognition technologies are not sophisticated enough, which could easily result in lawful uses of copyrighted works being blocked.

By adopting the Directive on Copyright in the Digital Single Market, the EU started a transition toward a 'vertical' approach to intermediary liability. This new approach can also be detected in new European legislation aimed at introducing a number of measures to prevent the misuse of Internet hosting services for the dissemination of texts, images, sound recordings, or videos that incite, solicit, or contribute to terrorist offenses. The regulation on addressing the dissemination of terrorist content online[59] is designed to establish binding, uniform rules that will, above all, ensure the swift removal of terrorist online content.[60] The regulation contains a uniform definition of terrorist online content, in line with EU fundamental rights protection. Service providers will have to remove terrorist content or disable access to it in all EU member states as soon as possible and in any event within one hour after they have received a removal order from a competent authority in an EU member state. Material disseminated for educational, journalistic, artistic, or research purposes, or that aims to prevent or counter terrorism will not be considered 'terrorist content;' this also includes content expressing polemic or controversial views in a public debate. The regulation includes effective remedies for both users whose content has been removed and service providers to submit a complaint.

The EU legal framework for social networks (in a broad sense) has also expanded with the latest review of the Audiovisual Media Services Directive (hereinafter 'AVMS Directive').[61] The AVMS Directive defines a 'video-sharing platform service' as a service where (i) the principal purpose of the service or of a dissociable section thereof or an (ii) essential functionality of the service is devoted to providing programmes, user-generated videos, or both, to the general public, for which the

58 Art. 17 and § 70 of the preamble of the Directive on the Digital Single Market.

59 Regulation (EU) 2021/784 of the European Parliament and of the Council on addressing the dissemination of terrorist content online, *Official Journal* L 172, 17.5.2021.

60 The removal of content is not the only activity that hosting service providers should undertake. According to the Proposal, providers should impose specific 'proactive measures' (see Art. 6 of the Proposal), although they do not have a general monitoring obligation. The Proposal states that in light of the particularly grave risks associated with the dissemination of terrorist content, the decisions adopted on the basis of the Regulation could, in fact, derogate from the prohibition of general monitoring set in the e-commerce directive. For an in-depth analysis of the Proposal, see: Kuczerawy, 2018, pp. 1–17.

61 Directive 2010/13/EU of the European Parliament and of the Council on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services, *Official Journal* L 095, 15.4.2010; L 263, 6.10.2010; L 303, 28.11.2018.

video-sharing platform provider does not have editorial responsibility. The service must be made available by means of an electronic communications network and the organization of the service determined by the video-sharing platform provider, including by automatic means or algorithms. The AVMS Directive states that in order for the provision of audiovisual content to constitute an 'essential functionality' of the service, such content must not be 'merely ancillary to, or a minor part of' the activities of the service. The European Commission's Guidelines on video-sharing platforms[62] set out several indicators that national authorities should consider, which can be grouped into four main categories: (1) the relationship between the audiovisual content and the main economic activities of the service; (2) quantitative and qualitative relevance of the audiovisual content available on the service; (3) monetization of, or revenue generation, from the audiovisual content; and (4) the availability of tools aimed at enhancing the visibility or attractiveness of the audiovisual content. Consequently, social media services can constitute video-sharing platform services and would fall within the scope of the AVMS Directive if they meet the relevant criteria.[63] The European Commission acknowledges that social media services have become an important medium by which users (particularly young people) access audiovisual content, and both the AVMS Directive and the Guidelines emphasize that because many social media services (i) compete for the same audiences and revenues as audiovisual media services and (ii) have a considerable impact, they must comply with the same regulations where they meet the relevant criteria.[64]

Although the AVMS Directive explicitly states that the e-commerce directive's 'safe harbor' provisions remain applicable, it requires member states to ensure that video-sharing platform providers operating within their respective jurisdictions take 'appropriate measures' to protect: (1) minors from programmes, user-generated videos and audiovisual commercial communications which may impair their physical, mental or moral development; (2) the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group; (3) the general public from programmes, user-generated videos and audiovisual commercial communications containing content the dissemination of which constitutes an activity which is a criminal offence under Union law, namely public provocation to commit a terrorist offence, offences concerning child pornography and offences concerning racism and xenophobia.[65] What constitutes an 'appropriate measure' is to be determined in light of the nature of the content in question, the harm it may cause, the characteristics of the category of persons to be protected as well as the

62 Communication from the Commission Guidelines on the practical application of the essential functionality criterion of the definition of a 'video-sharing platform service' under the Audiovisual Media Services Directive 2020/C 223/02 C/2020/4322, *Official Journal* C 223, 7.7.2020.

63 Services such as YouTube, as well as audiovisual content shared on social media services, such as Facebook, are covered by the revised AVMS Directive.

64 AVMS Directive, recital 4.

65 *Ibid*, art. 28b, para. 1.

rights and legitimate interests at stake, including those of the video-sharing platform providers and the users that created or uploaded the content, as well as the general public interest.[66]

The EU's interest in regulating online intermediaries was further demonstrated in late 2020, when the European Commission submitted a new legislative proposal to the European Parliament and European Council. The package consists of proposals of two regulations: the Digital Services Act[67] and the Digital Markets Act.[68] In the context of freedom of expression, the Digital Services Act is meant to improve the existing content moderation mechanisms. The Act will apply to online intermediaries ranging from cloud services and messaging services to marketplaces, Internet providers, and social networks. Further to this, specific due diligence obligations will apply to hosting services and online platforms, which are a subcategory of hosting services. The platforms will be required to disclose to regulators how their algorithms work, how decisions to remove content are taken, and the way advertisers target users. The Digital Services Act will create stronger public oversight of online platforms, particularly for platforms that reach more than 10% of the EU's population. Some of the measures proposed by the European Commission are: (1) measures to counter illegal goods, services or content online, such as a mechanism for users to flag such content and for platforms to cooperate with 'trusted flaggers;' (2) new obligations on traceability of business users in online market places, to help identify sellers of illegal goods; (3) effective safeguards for users, including the possibility to challenge platforms' content moderation decisions; (4) transparency measures for online platforms on a variety of issues, including on the algorithms used for recommendations; (5) obligations for very large platforms to prevent the misuse of their systems by taking risk-based action and through independent audits of their risk management systems; (6) access for researchers to the largest platforms' key data, in order to understand how online risks evolve; (7) oversight structure to address the complexity of the online space. We shall not further analyze the proposed rules, given that they could (and most probably will) be modified during the legislative process that has just started.

1.5. Social networks' internal rules on content moderation

In 1997, the US Government explicitly supported self-regulation as the primary mechanism for regulating the Internet in its report 'Framework for global electronic commerce,' stating that:

66 *Ibid,* art. 28b, para. 3.

67 Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/825 final.

68 Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act), COM/2020/842 final.

> (…) governments should encourage industry self-regulation wherever appropriate and support the efforts of private sector organizations to develop mechanisms to facilitate the successful operation of the Internet. Even where collective agreements or standards are necessary, private entities should, where possible, take the lead in organizing them.[69]

Today, more than twenty years later, we are witnessing different forms of online rules and regulations, such as terms of service,[70] privacy policies,[71] IP policies,[72] and community standards.[73] Although Internet platforms tend to present these rules as users' democratic participation in their services and may occasionally seek public feedback, they actually reflect the asymmetric relationship between platforms and users. More accurately, these rules are made and closely enforced by corporate entities and are far from the 'self-governance utopia' of the 1990s.

Further to the rules' lack of democratic legitimacy, the internal content moderation mechanisms demonstrate a striking transparency deficit. Due to the extreme volume of content posted online, these mechanisms are increasingly being applied automatically by way of artificial intelligence (AI), (almost) without any human interference. Automatic detection and filtering technologies are becoming essential tools in the fight against illegal online content. Indeed, many large platforms are now making use of some form of matching algorithms based on a range of technologies, from metadata filtering to hashing and fingerprinting content. However, the asymmetry of AI is even more problematic, since the user only sees the results of its individual decisions and has no access to accurate information about the input that determined a particular output.[74] Moreover, bias may be introduced into machine learning processes at various stages, including during algorithm design. Users have no information regarding the design or instructions the platforms input into the machine, and it could easily be a source of biases and over-removal.[75]

In its 2018 Recommendation on Measures to Effectively Tackle Illegal Content Online, the European Commission endorsed the provision of effective and appropriate safeguards to ensure that decisions taken concerning the removal of content are accurate and well-founded. In the Commission's view, such safeguards should consist, in particular, of human oversight and verification where appropriate and, in any event, where a detailed assessment of the relevant context is required in order to determine whether or not the content is to be considered illegal.[76] Moreover, if

69 White House, *The Framework for Global Electronic Commerce*, 1997. See: https://bit.ly/3lDsnnm.
70 See, for example, Twitter Terms of service, https://twitter.com/en/tos.
71 See, for example, Instagram Data policy, https://help.instagram.com/519522125107875.
72 See, for example, YouTube Copyright policy, https://bit.ly/3lDdT7a.
73 See, for example, Facebook Community standards, https://www.facebook.com/communitystandards/.
74 Castets-Renard, 2020, p. 23.
75 *Ibid.*
76 Recommendation on measures to effectively tackle illegal content online, C(2018) 1177 final, § 20.

the proposed Digital Services Act is adopted, intermediary service providers will be required to provide terms and conditions that include information about any restrictions that they impose on the use of their service in respect of information provided by the service recipients. That information will have to include information about any policies, procedures, measures, and tools used for the purpose of content moderation, including algorithmic decision making and human review.[77]

Finally, once a decision on content removal is reached, pursuant to the social network's internal rules, it is usually impossible to challenge. In most cases, there is no judicial review available when platforms take action against content or activity that violates their community standards or terms of service. Although some litigants are testing the limits of this obstacle before the US courts, since most Big Tech companies are headquartered in the United States, they have not yet prevailed.[78] However, in some other jurisdictions the courts have recognized that users have remedies against platforms that wrongfully delete content. In Germany, for instance, the courts have long applied the *Drittwirkung* doctrine, which recognizes that public law values influence private rights. On several occasions, the courts held that, under the *Drittwirkung* doctrine, Facebook must respect fundamental rights when it determines whether to delete content pursuant to its terms of service.[79]

There are numerous examples social media platforms' clear mistakes or at least questionable content removal decisions. For example, in 2016, Facebook, under its child pornography policy, blocked the sharing of the iconic 'Napalm Girl' photo depicting a young Vietnamese girl running naked and panicked from a napalm attack on her village. However, following widespread criticism from news organizations and media experts across the globe, Facebook reversed its decision.[80]

In response to longstanding criticism demanding user accountability, Mark Zuckerberg, CEO and founder of Facebook, the most popular social network,[81] announced in November 2018 that his company would create an independent governance and oversight committee by the close of 2019 to advise on content policy and listen to user appeals on content decisions.[82] In September 2019, Facebook published the Oversight Board Charter, a document that delineates the structural relationship between Facebook, the Oversight Board, and the Trust that ensures the Board's financial independence from Facebook.[83] The Oversight Board has between eleven and forty members; it will increase or decrease in size 'as appropriate.'[84] Members

77 Proposal of the Digital Services Act, art. 12, para. 1.

78 *Prager Univ. v. Google LLC*, No. 17-CV-06064-LHK, 2018 WL 1471939, at *14 (N.D. Cal. Mar. 26, 2018).

79 Bloch-Wehba, 2019, p. 77.

80 See for example: The Guardian, 'Facebook backs down from 'napalm girl' censorship and reinstates photo'. Available at: https://bit.ly/3EC00yO.

81 Per number of active users.

82 Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement', 15 November 2018. Available at: https://bit.ly/2XFrwLg.

83 Facebook Oversight Board Charter. Available at: https://bit.ly/3tZgagF.

84 Facebook Oversight Board Charter, art. 1. The names of the first twenty members were announced in May 2020.

of the Oversight Board must possess and exhibit a broad range of knowledge, competencies, diversity, and expertise, and must have demonstrated experience deliberating thoughtfully as an open-minded contributor on a team, be skilled at making and explaining decisions, and have familiarity with matters relating to digital content and governance, including free expression, civic discourse, safety, privacy, and technology.[85] The Charter also instructs the Board to split into subsections, termed panels, when reviewing cases. Each panel has to contain at least one member from the region where the case arose.[86]

Excluding content that was removed in compliance with local laws[87] and requiring following an exhaustion of appeals through Facebook, a request for review can be submitted to the Board by either the original poster of the content or a person who previously submitted the content to Facebook for review.[88] Consequently, the Oversight Board has the authority to review not only content that has been removed (original poster of the content) but content that is kept up (person who previously submitted content for review). However, the Facebook Oversight Board bylaws create many exceptions to the Board's scope of review. As established at the Board's launch,[89] only single-object removals of organic content posted on Facebook and Instagram are eligible for review.[90] Within that, content decisions 'pursuant to legal obligations,' including those concerning intellectual property, the Facebook marketplace, fundraisers, Facebook dating, messages, and spam, are out of the scope.[91]

1.6. Social networks between proclaimed neutrality and value-based decisions

Following a brief period of euphoria about the possibility that social networks might facilitate global democratization, there is now widespread concern in many segments of society that social networks may instead be undermining democracy. Their specific role in a digital society does not easily fit into any of the existing categories. They cannot be qualified as 'speakers,' as they do not publish their own content, nor do they associate themselves with the content their users publish. They cannot be qualified as a traditional 'editor' either, as they do not initiate or

85 *Ibid.*
86 *Ibid.*
87 *Ibid,* art. 7.
88 *Ibid,* art. 2.
89 The type of content eligible for review can be broadened in time. For a critical assessment, see: Klonick, 2020, p. 2465 *et seq.*
90 'Organic content' is content posted by users, contrary to commercial advertising. 'Single-object' refers to a post containing a photo, video, or status message. 'Complex object' is a user profile, group, or page.
91 Facebook Oversight Board Bylaws, art. 2, § 1.2. See: https://www.oversightboard.com/sr/governance/bylaws.

commission the production of content. However, they do exercise certain editorial functions in the sense that they moderate the content their users post.[92]

The system that social networks have put in place to match users' expectations and self-regulate is indeed responsive, as demonstrated in our analysis. However, this system presents two major downsides that become more apparent over time. First, there is an evident loss of equal access to and participation in speech on these platforms.[93] Social networks are increasingly making their own choices regarding content moderation that give preferential treatment to some users over others, e.g., by designing algorithms in accordance with the network owner's preferences. Moreover, algorithms are often set to create perfect filtering in order to only show users content that meets their personal tastes. This may create a basically antidemocratic space in which people are shown things with which they already associate. As a number of social science researchers have rightfully noted,[94] although the rise of social media has made citizens much less dependent on television and traditional newspapers, this certainly does not mean that citizens have more control over the media environments in which they now operate. Media power has not been transferred to the public; instead, power has partly shifted to algorithmic selections operated by large digital platforms.

The second problem is that of accountability. Social networks should be open about their takedown rules and follow a consistent and transparent process. Under the current legal regime, the user is virtually powerless. Users are not sufficiently informed about the criteria social networks apply when moderating content. In most cases, the user cannot successfully challenge the platform's content moderation decisions either. Greater transparency in content moderation implies publication of the number of posts and accounts being removed, provision of a clear notice to users disclosing the reason for content removal, and human review of removal decisions undertaken by software.

2. Fake news as a global factor in the influence of social networks on the guarantees of freedom of speech and the truthfulness of information

In recent years, concerns about the societal consequences of the online spread of disinformation and propaganda have become widespread. New digital tools that allow anyone to easily spread political information to large numbers of Internet users can lead to a more pluralistic public debate, but they can also give a platform

92 Koltay, 2019, p. 189.
93 Klonick, 2017, p. 1665.
94 See for example: Poell and van Dijck, 2015, pp. 527–537.

to extremist voices and actors seeking to manipulate the political agenda in their own political or financial interest.[95] The problem of 'fake news' attracted substantial attention during the 2016 US presidential elections, after a series of events known as 'Pizzagate.' Namely, fake news publishers in North Macedonia circulated a false political conspiracy theory that former First Lady, Secretary of State, and presidential candidate Hillary Clinton and other prominent Democratic political figures were coordinating a child trafficking ring out of a Washington-based pizzeria by the name of Comet Ping Pong. This fake news was widely shared via social networks. In December 2016, a man who read the publication drove from North Carolina to Washington, DC and shot open a locked door at Comet Ping Pong pizzeria with his assault rifle.[96]

False statements of fact typically published on websites and disseminated via social networks for profit or social influence are usually referred to as fake news, rumors, counter-knowledge, disinformation, post-truths, alternative facts, or simply lies. Although this phenomenon is omnipresent, it is rarely defined in legal documents (Section 2.1). More recently, the concept of 'deep fakes' has been introduced (Section 2.2). The creation and/or dissemination of fake news may result in civil, criminal, or administrative liability for Internet users. Moreover, social networks have adopted their own internal rules aimed at combatting the dissemination of fake news (Section 2.3). Some governments and non-governmental organizations, either on their own or in collaboration with social networks, have introduced media literacy initiatives as an alternative approach to combatting fake news (Section 2.4).

2.1. The concept of fake news

The UK Collins Dictionary named 'fake news' the 2017 'word of the year.' According to the dictionary, usage of the phrase indicating "false, often sensational, information disseminated under the guise of news reporting" increased by 365% since 2016.

The two defining characteristics used to identify different types of fake news are, first, whether the author intends to deceive readers and, second, whether the motivation for creating or disseminating the fake news is financial.[97] By applying these two criteria, one could differentiate among at least four types of fake news. The first type is satire, that is, a news story that does not intend to deceive, although it purposefully contains false content, and is generally motivated by non-pecuniary interests, though financial benefit may be a secondary goal. The second type of fake news is a hoax, which is a news story with purposefully false content where the author intends to deceive readers into believing incorrect information and that is

95 Tucker et al., 2018, p. 15.

96 BBC, 'The saga of Pizzagate: The fake story that shows how conspiracy theories spread'. Available at: https://bbc.in/39tv59i.

97 Verstraete et al., 2017, p. 6.

financially motivated. Typically, creators of hoaxes do not have political or cultural motivations that drive the production of their fake news stories. The third type is propaganda, which is news or information with purposefully biased or false content where the author intends to deceive readers and that is motivated by promoting a political cause or point of view, regardless of financial reward. Fourth, 'trolling' presents news or information with biased or fake content where its author intends to deceive readers and is motivated by an attempt to derive personal humorous value (the lulz).[98] The term 'fake news' has a distinctively negative connotation, which is why the general public's understanding is usually limited to the second and third types of activities (i.e., hoax, propaganda).

Given its complexity and the different perceptions, the term 'fake news' is less employed in legal doctrine and legal documents in recent years. Instead, it is being replaced by the term 'disinformation.' This is particularly the case in the EU in the context of recent European Commission initiatives. Specifically, in 2018, the European Commission set up a high-level expert group on fake news and online disinformation to advise the Commission on establishing the scope of the disinformation phenomenon, defining the roles and responsibilities of relevant stakeholders, and formulating recommendations. The expert group released its final report[99] only a few months later. This was followed by the European Commission's Communication titled 'Tackling Online Disinformation: A European Approach.'[100] In September 2018, the European Commission published the Code of Practice on Disinformation (hereafter, 'the Code').[101] The Code represents a voluntary, self-regulatory mechanism agreed upon by representatives of online platforms, social networks, advertisers, and the advertising industry. The Code employs the term 'disinformation,' defined as 'verifiably false or misleading information' that is both "created, presented and disseminated for economic gain or to intentionally deceive the public" and may cause public harm, intended as "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security."[102] The term does not cover misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary.[103] Moreover, disinformation as defined here includes forms of speech that fall outside already illegal forms of speech, notably defamation, hate speech, incitement to violence, etc., but can nonetheless be harmful.[104]

98 'Lulz' is a typographical subversion of the word 'lol,' meaning to 'laugh out loud.'

99 European Commission, Final report of the High level expert group on fake news and online disinformation, 'A multi-dimensional approach to disinformation', 2018. See: https://bit.ly/3zt3bF2.

100 European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions, 'Tackling online Disinformation: a European Approach', COM(2018) 236 final.

101 European Commission, Code of practice on disinformation, 2018. See: https://bit.ly/39rdpey.

102 *Ibid*, preamble, p. 1.

103 *Ibid.*

104 Final report of the High level expert group on fake news and online disinformation, p. 10.

The reason the EU seems to prefer the term 'disinformation' to 'fake news' is explained in the Final Report of the High Level Expert Group on Fake News and Online Disinformation. First, the latter term is considered to be inadequate to capture the complex problem of disinformation, which involves content that is not actually or completely 'fake' but is rather fabricated information blended with facts and practices that go well beyond anything resembling 'news' to include some forms of automated accounts used for astroturfing, networks of fake followers, fabricated or manipulated videos, targeted advertising, organized trolling, visual memes, and much more. Second, the term 'fake news' has been appropriated by some politicians and their supporters, who use it to dismiss coverage that they find disagreeable.[105]

2.2. The concept of deep fakes

'Deep fakes' are face-swapping technologies that enable the quick creation of fake images or videos that appear very realistic. Deep fake technology can also be used to create 'voice clones,' usually of public figures. Typically, deep fakes rely on artificial neural networks, which are computer systems that recognize patterns in data. Developing a deep fake photo or video involves feeding hundreds or thousands of images into the artificial neural network in order to 'train' it to identify and reconstruct patterns. Coinage of the term 'deep fakes' is attributed to a Reddit[106] user called 'deepfakes,' who published several videos in which famous actresses' faces were swapped into pornographic videos in late 2017.[107] The increased availability of deep fakes, especially through apps, raises a number of legal, social, and ethical questions. Indeed, their very existence is blurring the line between what is true and what is fake.

Legal theory distinguishes among four main types of deep fakes.[108] The first type is deep fake pornography, for which technology is used either to create celebrity deep fakes or revenge porn. Celebrity deep fakes refer to content where celebrity images are superimposed on the bodies of individuals engaged in sexual acts. Revenge porn is created by persons seeking revenge for terminated relations. The second type of deep fake comprises fake photos or videos created during a political campaign. This type of deep fake can have significant negative consequences for democratic processes, as deep fakes can target certain individuals' reputation or portray fake events. The third type of deep fake comprises fake photos, videos, or voices created for commercial purposes. For example, the technology can be used to translate a video by enabling the recorded person to 'speak' in different languages.

105 *Ibid.*

106 Reddit is a website comprising user-generated content (including photos, videos, links, and text-based posts) and discussions of this content in what is essentially a bulletin board system.

107 The New York Times, 'Here come the fake videos, too'. See: https://nyti.ms/3AtUH1X.

108 See for example: Meskys et al., 2020, pp. 24–31.

Finally, the fourth type could be referred to as a creative deep fake. This category comprises fake content created purely for creative purposes, usually as parody or satire.

2.3. Legal framework for combatting the creation and dissemination of fake news

The creation and/or dissemination of fake news may result in civil, criminal, or administrative liability for Internet users. Further to these 'traditional' legal instruments, legislators in certain jurisdictions have adopted specific legislative acts aimed at combatting the creation and dissemination of fake news. We will analyze in further detail the existing legal framework related to fake news in the United States, on the one hand, and in the EU and its member states, on the other hand. Moreover, social networks have adopted their own internal rules aimed at combatting the dissemination of fake news.

2.3.1. US law

Fake news creators and/or disseminators are frequently sued by private individuals or businesses seeking to collect monetary damages or injunctive relief in civil law proceedings. The most frequent claim invoked against fake news creators and/or disseminators is the common law tort of defamation.[109] In the United States, false publications of fact concerning a public figure (e.g., a government official) are actionable only if the publisher acted with actual malice, i.e., either with knowledge of the statement's falsity or reckless disregard for the same. However, strictly private figures do not need to prove actual malice; they are only required to prove that the defamatory statements were published with negligence. If we define fake news restrictively, so as to include only intentional or knowingly false statements, it is reasonable to conclude that such statements would satisfy the requirements for defamation claims. However, fake news in a broad sense need not always satisfy these requirements. For example, a satire or parody is actionable only if it could be reasonably understood to describe actual facts or events, which is typically not the case. Finally, it should be recalled that Section 230 of the Communications Decency Act of 1996 protects online publishers[110] from defamation claims in situations where another Internet user provided the information.

After defamation, intentional infliction of emotional distress (IIED) is a common law tort that is regularly alleged against fake news creators and/or disseminators under state law. IIED occurs when a person intentionally or recklessly engages in

109 Defamation is the communication of a false statement of fact that harms another person's reputation or character. Spoken (unrecorded) defamation is referred to as slander, while defamatory statements that are written or otherwise recorded are known as libel.

110 However, it does not protect the original author of a defamatory or otherwise tortious publication.

extreme or outrageous behavior that causes another person to suffer severe emotional distress. Unlike defamatory statements, which may be actionable for simply being harmful and false, statements supporting IIED claims must be "so outrageous in character, and so extreme in degree, as to go beyond all possible bounds of decency, and to be regarded as atrocious, and utterly intolerable in a civilized community."[111] Consequently, particularly extreme fake news content remains susceptible to IIED claims, especially when involving non-public figures.

Moreover, creating fake news content could easily violate a third party's intellectual property rights, typically a copyright or trademark right. The creators of text, photographs, videos, and other original works of authorship are granted exclusive rights to reproduce, distribute, display, and create derivative works from such content. Consequently, creators and/or disseminators of fake news content using third-party materials have to seek the copyright owners' permission (unless the work is in the public domain or the doctrine of fair use applies). In addition, the creators of fake news content should refrain from using third-party trademarks or logos that may confuse consumers as to the origin of products, since the Lanham Act and state unfair competition law prohibit trademark infringements and false representations of fact in commercial advertising that misrepresent the nature or characteristics of another's goods, services, or commercial activities.[112] Creators and/or disseminators of fake news content may also be sued for the violation of the right of publicity, i.e., respect for a person's name and likeness, which most US states recognize.[113] The right of publicity grants an individual the right to control the *commercial* use of their identity.

In addition to civil law liability, fake news creators and/or disseminators may be accused of crimes or the violation of other specific regulations. For example, the Federal Trade Commission (FTC) is given broad discretion to investigate questionable trade practices and take appropriate enforcement action. Entities found to have engaged in consumer fraud or deception can be permanently enjoined by a court from continuing such conduct in the future. They may also be ordered to pay civil penalties and provide consumer redress.[114] Further to this, criminal libel statutes exist in several US states and territories.[115] The elements of criminal libel are similar to the elements of civil defamation. Criminal libel consists of defamation of an individual (or group) made public by a printing or writing. The defamation must

111 Restatement (Second) of Torts § 46 cmt. d (Am. Law Inst. 1965). For a critical analysis of IIED see: Fraker, 2008, pp. 983–1026.

112 15 U.S.C. § 1125(a).

113 See for example: N.Y. Civ. Rights Law § 50.

114 Within the FTC is the Bureau of Consumer Protection, which is designed to protect consumers from deceptive or unfair business practices. The Bureau of Consumer Protection focuses on protecting consumers' privacy, fighting identity theft, regulating advertising and marketing practices, regulating business practices in the financial industry, and protecting US citizens from telemarketing fraud.

115 For example, in Florida (see: Chapter 836 of the Florida Statutes).

tend to excite a breach of the peace or damage the individual (or group) in reference to their character, reputation, or credit.[116]

Finally, in October 2017, Congress announced a bill that would require digital platforms with at least 50,000,000 monthly visitors to maintain a public file of all electioneering communications purchased by a person or group who spends more than $500.00 in total on ads published on their platform. This file must contain a digital copy of the advertisement, a description of the audience the advertisement targets, the number of views generated, the dates and times of publication, the rates charged, and the purchaser's contact information. The bill, called the Honest Ads Act, was introduced by US senators Mark Warner, Amy Klobuchar, and Lindsey Graham, with the aim of preventing foreign interference in future elections and improving the transparency of online political advertisements.[117] The proposed legislation addresses a loophole in the existing campaign finance laws that regulate television and radio ads, but not Internet ads. The Honest Ads Act would help close that gap by subjecting Internet ads to the same rules as television and radio ads.

2.3.2. European Union and its member states

The problem of disinformation on the Internet is a source of growing concern for EU policymakers. As previously mentioned, in September 2018, the European Commission published the Code of Practice on Disinformation, which is a voluntary, self-regulatory mechanism agreed upon by representatives of online platforms, social networks, advertisers, and the advertising industry. The Code observes that social networks facilitate the dissemination of disinformation, impacting a broad segment of actors in the ecosystem. For this reason, all stakeholders have roles to play in countering the spread of disinformation.[118] The Code considers advertising and monetization incentives as leading to behaviors such as misrepresentations about oneself or the purpose of one's properties.[119] In response, the Code's signatories have committed to deploying policies and processes to disrupt such incentives. The signatories have acknowledged, in particular, that there is a need to significantly improve the scrutiny of ad placements.[120] All parties involved in the online advertising market need to work together to improve transparency across the ecosystem. This means that they should effectively scrutinize, control, and limit the placement of advertising on accounts and websites belonging to purveyors of disinformation.[121] The signatories, moreover, should make commercially reasonable efforts to ensure that they do not accept remuneration from or promote accounts and websites that

116 Brenner, 2007, p. 714.
117 The full text of this legislative proposal is available here: https://bit.ly/2XBWKCA.
118 Code of Practice on Disinformation, p. 1.
119 *Ibid,* p. 5.
120 *Ibid,* p. 4.
121 *Ibid,* p. 4.

consistently misrepresent information about themselves.[122] The Code acknowledges the need to ensure transparency in the area of political and issue-based advertising. In particular, such transparency means that users should be able to understand why they have been targeted for a given advertisement.[123]

Some of the self-regulatory standards introduced by the Code are reflected in the European Commission's proposal of the Digital Services Act, published in December 2020.[124] The Act is supposed to impose greater transparency obligations for platforms in the field of targeted advertising, amongst other requirements in the field of content regulation. Penalties for violations of the rules include fines of up to 6% of a company's annual income.[125] In the field of online advertising, the European Commission has proposed rules that would give online platform users immediate information about the sources of the ads they see online, including granular information about why an individual has been targeted with a specific advertisement.[126] Moreover, very large online platforms[127] that display advertising on their online interfaces will have to compile and make publicly available through application programming interfaces a repository containing the following information: (1) the content of the advertisement; (2) the natural or legal person on whose behalf the advertisement is displayed; (3) the period during which the advertisement was displayed; (4) whether the advertisement was intended to be displayed specifically to one or more particular groups of recipients of the service and if so, the main parameters used for that purpose; (5) the total number of recipients of the service reached and, where applicable, aggregate numbers for the group or groups of recipients whom the advertisement targeted specifically. The information will have to remain publicly available until one year after the last time the advertisement was displayed on their online interfaces.[128]

Several EU member states have complemented the EU's current self-regulatory approach, which is best demonstrated in the Code of Practice on Disinformation, with its mandatory rules and harsher sanctions for non-compliance. Germany reacted first, although its reaction was directed more toward hate speech than fake news. In September 2015, the German Minister of Justice first initiated a task force composed of representatives of the service providers Facebook, Twitter, and Google (with respect to its service YouTube), and several nongovernmental organizations (NGOs) to jointly fight illegal speech. The self-regulatory measures they agreed upon included user-friendly notification mechanisms, an immediate review of notified content for

122 *Ibid.*

123 *Ibid,* p. 5.

124 Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/825 final.

125 *Ibid,* arts. 42 and 59.

126 *Ibid,* art. 24.

127 Online platforms that provide their services to a number of average monthly active service recipients in the Union equal to or higher than 45 million.

128 Digital Services Act, art. 30.

compatibility with German law (within 24 hours of notification), adequate responses to illegal hate speech including the blocking of access to domestic users without undue delay, and transparent notice and takedown policies.[129] In spite of leading social networks' willingness to implement this self-regulatory mechanism, Germany proceeded with the adoption of harsher mandatory rules against illegal content online. In 2017, German Parliament adopted the Law Improving Law Enforcement on Social Networks (NetzDG).[130] This federal law aims at improving law enforcement regarding social networks by calling 'telemedia service providers'[131] to account regarding acting on online speech that is punishable under domestic criminal law. The NetzDG applies to all telemedia service providers that, for profit-making purposes, operate Internet platforms designed to enable users to share any content with other users or make such content available to the public.[132] Social network operators with at least two million registered users within Germany are required to implement an effective, transparent complaints management infrastructure and have the duty to compile reports on complaints management activity.[133] The law distinguishes between content that is manifestly illegal and that which is illegal. Manifestly illegal content must be deleted or removed within 24 hours of receiving a complaint, while for merely illegal content, a period of seven days is granted for action.

As neither hate speech nor the dissemination of fake news as such are statutory offenses under German criminal law, the NetzDG lists a catalogue of offenses considered to be illegal content requiring access blocking: (1) dissemination of propaganda material of unconstitutional organizations; (2) usage of symbols of unconstitutional organizations; (3) preparation of a serious violent offense endangering the State; (4) encouraging the commission of a serious violent offence endangering the state; (5) treasonous forgery; (6) public incitement to crime; (7) breach of the public peace by threatening to commit offense; (8) forming criminal or terrorist organizations; (9) incitement to hatred; (10) dissemination of depictions of violence; (11) rewarding and approving of offenses; (12) defamation of religions, religious and ideological associations; (13) distribution of child pornographic performances by broadcasting, media services or telecommunications services; (14) insult; (15) defamation; (16) violation of intimate privacy by taking photographs; (17) threatening the commission of a felony; and (18) forgery of data intended to provide evidence.[134]

129 Schmitz-Berndt and Berndt, 2018, p. 15.

130 *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken*, Bundesgesetzblatt Teil 1 (BGB 1), n° 61, 7 September 2017.

131 Telemedia service providers are defined as electronic information and communications services, insofar as they do not provide telecommunications services, which consist of the transmission of signals via telecommunications networks, telecommunications-based services, or broadcasting services.

132 NetzDG, § 1(1).

133 *Ibid,* § 2-3.

134 *Ibid,* § 1(3).

Paradoxically, although the battle against fake news has been one of the main arguments to pass the NetzDG, the notion does not appear in the law itself.[135]

In November 2018, neighboring France adopted the Law Against the Manipulation of Information,[136] which targets the widespread and extremely rapid dissemination of fake news by means of digital tools, especially through the dissemination channels offered by social networks and media outlets influenced by foreign states. The law requires online platforms with more than five million unique users per month in France to adhere to the following conduct during the three months preceding general elections: (1) provide users with honest, clear, and transparent information about the identity and corporate address of anyone who paid to promote informational content related to a 'debate of national interest;' (2) provide users with honest, clear, and transparent information about the use of personal data in the context of promoting content related to a 'debate of national interest;' (3) make public the amount of payments received for the promotion of informational content when these amounts are above a certain threshold.[137] Moreover, the law provides that, during the three months preceding an election, a judge may order 'any proportional and necessary measure' to stop the deliberate, artificial, or automatic and massive dissemination of fake or misleading information online.[138] A public prosecutor, candidate, political group or party, or any person with standing can bring a fake news case before a judge, who must rule on the motion within 48 hours.[139] An interim judge will qualify the fake news, as defined in the 1881 Law on the Freedom of the Press, in accordance with three criteria: (1) the fake news must be manifest, (2) be disseminated deliberately on a massive scale, and (3) lead to a disturbance of the peace or compromise the outcome of an election.[140] Further to this, the Law Against the Manipulation of Information requires that online platform operators implement measures to prevent the dissemination of false information that could disturb public order or affect the validity of an election.[141] They must also establish an easily accessible mechanism for users to flag fake information, and they are required to submit a yearly report to the French Superior Council on Audiovisual (CSA)[142] detailing the measures they have taken to curb the dissemination of fake news.[143]

Italy also reacted to the online spread of disinformation by introducing a specific enforcement mechanism to combat fake news during the election period. In January

135 Schmitz-Berndt and Berndt, 2018, p. 21.

136 *Loi n° 2018–1202 relative à la lutte contre la manipulation de l'information,* Official Journal n°0297 of 23 December 2018. This 'ordinary law' is paired with the 'organic law' against the manipulation of information: *Loi organique n° 2018–1201 relative à la lutte contre la manipulation de l'information,* Official Journal n°0297 of 23 December 2018.

137 (Ordinary) law against the manipulation of information, art. 1.

138 *Ibid.*

139 *Ibid.*

140 Law on the freedom of press (*Loi du 29 juillet 1881 sur la liberté de la presse*), art. 27.

141 (Ordinary) law against the manipulation of information*,* art. 11.

142 *Conseil supérieur de l'audiovisuel.*

143 (Ordinary) law against the manipulation of information, art. 11.

2018, the minister of the interior introduced the Operating Protocol for the Fight Against the Diffusion of Fake News through the Web on the Occasion of the Election Campaign for the 2018 Political Elections.[144] General elections were scheduled for March 2018.[145] The protocol introduced a 'red button' reporting service where users "may indicate the existence of a network of content attributable to fake news." The Polizia Postale, a unit of the Italian State Police that investigates cybercrime, were tasked with reviewing reports and acting accordingly. The web portal allowed users to submit links to content and social networks (if they found the content on a social network), as well as further information. The portal also required users to provide their email address. The police then reviewed submissions with the aim of 'directing the next activity' for content that is 'manifestly unfounded and biased' or 'openly defamatory.' The police were supposed to carry out in-depth analysis using specific techniques and software in order to identify significant indicators allowing for the qualification, with maximum certainty, of the news as fake news (presence of official denials, false content already proven by objective sources, provenance of the alleged fake news from sources not accredited or certified, etc.). The Polizia Postale were also empowered to independently collect information "in order to identify early on the network of news markedly characterized by groundlessness and tendency that is openly defamatory." After reviewing the information, the authorities would pursue legal action if they determined that the content was unlawful. In cases where content was deemed to be false or misleading, but not unlawful, authorities would publish public denials.

The operating protocol contained references to defamation, which the Italian Penal Code defines as "injuring the reputation of an absent person via communication with others" and to which it attaches penalties of up to one year of imprisonment for members of the general public.[146] If the defamatory act or insult consisted of the allegation of a specific fact, the potential penalty increased to imprisonment for up to two years or a fine of 2,065 euros.[147] If committed by the press or otherwise publicly, violators could face penalties of at least 516 euros or imprisonment from six months to three years.[148] The penal code also provided for increased penalties for defamation against public officials. For example, the code imposed enhanced penalties of one to five years of imprisonment for criminal defamation of the president.[149] The Italian enforcement mechanism introduced in 2018 was criticized by the United Nations Human Rights Council (UN HRC) for failing to precisely define the type of

144 Press release: *Protocollo Operativo per il contrasto alla diffusione delle Fake News attraverso il web in occasione della Campagna elettorale per le Elezioni politiche 2018,* 18 January 2018. Available at: commissariatodips.it.

145 More on Italy's failed attempts to regulate 'fake news' prior to the adoption of the Operating protocol: Pollicino and Somaini, 2020, pp. 171–193.

146 Penal Code (*Codice Penale*), Official Journal n. 251/1930, art. 595.

147 *Ibid.*

148 *Ibid.*

149 *Ibid,* art. 278.

disinformation it targeted. The operating protocol aimed at combatting "manifestly unfounded and biased news, or openly defamatory content" left significant discretionary power to the police, according to the UN HRC special rapporteur on the promotion and protection of the right to freedom of opinion and expression.[150] Following widespread criticism, the authorities stopped enforcing the protocol.

2.3.3. Social networks' internal rules against fake news

Most social networks do not have a blanket rule against posting false material, but they do ban certain kinds of disinformation. Some allow specific types of false claims. For example, Facebook admits in its Community Standards that it does not totally ban fake news:

> Reducing the spread of false news on Facebook is a responsibility that we take seriously. We also recognize that this is a challenging and sensitive issue. We want to help people stay informed without stifling productive public discourse. There is also a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook but instead, significantly reduce its distribution by showing it lower in the News Feed.[151]

Twitter also stated that it is not addressing all false material:

> We are not attempting to address all misinformation. Instead, we prioritize based on the highest potential for harm, focusing on manipulated media, civic integrity, and COVID-19. Likelihood, severity and type of potential harm — along with reach and scale — factor into this.[152]

Social networks' internal rules against fake news are often vague and allow for a significant discretionary power as to whether the content will be blocked/permanently removed or not. TikTok, one of the newest social media, provides a good example of such rule ambiguity: "We do not permit misinformation that causes harm to individuals, our community, or the larger public regardless of intent."[153]

Social networks generally prohibit deep fakes, a specific type of manipulated content. For example, TikTok, Instagram, and Facebook explicitly prohibit AI-modified content: "Videos cannot be modified with AI tools in ways that are not apparent

150 Comments of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on the Operating protocol for the fight against the diffusion of fake news through the web on the occasion of the election campaign for the 2018 political elections, p. 4. Available at: https://bit.ly/3CxdMRw.

151 Facebook Community Standards, § 20. See: https://www.facebook.com/communitystandards/false_news.

152 Tweeted on 'Twitter Safety' profile on 3 June 2020.

153 Tik-Tok Community Standards, section: Misinformation. See: https://bit.ly/3Cxdmul.

to an average person, and would likely mislead an average person to believe that a subject of the video said words that they did not say."[154] On the other hand, YouTube has more lenient rules regarding deep fakes: "Videos must not be technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm."[155]

Advertisers face stricter rules than 'ordinary users' on almost every platform. For example, Facebook only fact checks 'regular posts' (written by 'ordinary users') under special circumstances, while paid advertisements are always checked before being published. Paid ads need to comply with Facebook's advertising policies, which cover misinformation *inter alia,* but also with its Community Standards, which apply to regular posts as well. Under Facebook's advertising policies, ads that include claims debunked by third-party fact checkers or, in certain circumstances, by organizations with particular expertise, are prohibited. Advertisers that repeatedly post information deemed to be false may have restrictions placed on their ability to advertise on Facebook.[156] However, ads are rarely checked by human moderators. Instead, Facebook uses an algorithmic ad screening system. Similarly to Facebook, Twitter claims that it does not allow ads that are false, deceptive, misleading, defamatory, or libelous.[157]

2.4. Alternative approaches to combatting fake news

The self-regulatory approach, which social networks prefer, as well as the co-regulatory approach, which the EU favors, typically face several challenges. First, conflicts of interest may occur between the social networks' need to keep users engaged and monetize their engagement, and the public authorities' need to the safeguard the integrity of democratic processes. Second, the amount of content that has to be monitored is enormous, which necessarily implies the use of algorithmic content screening and consequently introduces possible errors in that process. Third, the efficiency of fact checking mechanisms is limited, as algorithms cannot be relied upon to control the extremely vast amount of online content. On the other hand, direct state-imposed regulation, which is preferred by certain European and non-European countries, focuses on illegal content, while ignoring many other variants of disinformation. Moreover, there is no commonly accepted definition of 'fake news,' which leaves significant discretionary power to enforcers.

Given that it has recently become increasingly difficult to recognize fake news and particularly deep fake materials, some alternative approaches to combatting disinformation have also been designed and implemented. Many governments and

154 Facebook Community Standards, § 21. Similar rules are adopted by other two networks.

155 YouTube Policies, section: Spam, deceptive practices, and scams policies. See: https://bit.ly/3tXuKW0.

156 Facebook Advertising Policies, § 13. See: https://www.facebook.com/policies/ads/.

157 Twitter Ads Policies. See: https://business.twitter.com/en/help/ads-policies.html.

NGOs have launched different media literacy initiatives, sometimes in collaboration with social network operators. Media literacy is usually defined as an informed, critical understanding of the prevalent mass media, and it involves examining the techniques and institutions involved in media production, as well as the ability to critically analyze media messages. One of the aspects of digital media literacy is the ability to recognize disinformation or partially false digital content.

The European Commission has also recognized that media literacy is a crucial skill for all European citizens, as it helps them to counter the effects of disinformation campaigns and the spreading of fake news through digital media. The revised AVMS Directive strengthens the role of media literacy. It requires EU member states to promote measures that develop media literacy skills.[158] The AVMS Directive also obliges video-sharing platforms to provide effective media literacy measures and tools. This is a crucial requirement due to the central role such platforms play in providing access to audiovisual content. Platforms are also required to raise users' awareness of these measures and tools.[159] Additionally, the European Commission has established a media literacy expert group that brings media literacy stakeholders together. This group meets annually to (1) identify, document and extend good practices in the field of media literacy; (2) facilitate networking between different stakeholders; and (3) explore ways of coordinating EU policies, support programmes and media literacy initiatives.[160]

An alternative approach to combatting fake news consists of fact checking projects oriented toward monitoring the factual accuracy of news, political statements, and interviews. Fact checking web portals offer counter-narratives to untrue and manipulated information. Facebook and Instagram have also established a fact checking program, in partnership with independent third-party fact checkers who are certified through the non-partisan International Fact-Checking Network (IFCN). The fact checking program, launched in 2016, enables fact checking partners to review content across both Facebook and Instagram, including organic and boosted posts. They can also review videos, images, links, and text-only posts.

158 AVMS Directive, art. 33a.

159 *Ibid,* art. 28b.

160 European Commission, Directorate-General for Communications Networks, Content and Technology, Mandate of the Expert Group on Media Literacy, 6 July 2016. Available at: https://bit.ly/39tv19y.

Bibliography

Balasubramani, V. (2016/2017) 'Online Intermediary Immunity Under Section 230', *The Business Lawyer,* 72(1), pp. 275–286.

Bar-Ziv, S. and Elkin-Koren, N. (2018) 'Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown', *Connecticut Law Review,* 50(2), pp. 339–385.

Bloch-Vehba, H. (2019) 'Global platform governance: Private power in the shadow of the state', *SMU Law Review,* 72(1), pp. 27–80.

Brenner, S.W. (2007) 'Should online defamation be criminalized?', *Mississippi Law Journal,* 76(3), pp. 705–787.

Bridy, A. (2016) 'Copyright's Digital Deputies: DMCA-Plus Enforcement by Internet Intermediaries' in Rothchild, J. A. (ed.) *Research Handbook on Electronic Commerce Law.* 1st Cheltenham: Edward Elgar, pp. 185–209.

Castets-Renard, C. (2020) 'Algorithmic content moderation on social media in EU law: Illusion of perfect enforcement', *University of Illinois Journal of Law, Technology & Policy*, 2020(2), pp. 283–323.

Durach, F. et al. (2020) 'Tackling Disinformation: EU Regulation of the Digital Space', *Romanian Journal of European Affairs,* 20(1), pp. 5–20.

ECtHR (2020) *Guide to Article 10 of the Convention – Freedom of expression.* Strasbourg: Council of Europe (online edition).

Fraker, R. (2008) 'Reformulating Outrage: A Critical Analysis of the Problematic Tort of IIED', *Vanderbilt Law Review*, 61(3), pp. 983–1026.

Frosio, G. Geiger, C. (2021) 'Taking fundamental rights seriously in the Digital Services Act's platform liability regime'. Available at: http://cyberlaw.stanford.edu/publications/taking-fundamental-rights-seriously-digital-service-act%E2%80%99s-platform-liability-regime (Accessed: 15 April 2021).

Klonick, K. (2017) 'The new governors: the people, rules and processes governing online speech', *Harvard Law Review,* 131(6), pp. 1598–1670.

Klonick, K. (2020) 'The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression', *Yale Law Journal,* 129, pp. 2418–2499.

Koltay, A. (2019) *New media and freedom of expression: rethinking the constitutional foundations of the public sphere*. London: Hart Publishing.

Kuczerawy, A. (2018) 'The proposed Regulation on preventing the dissemination of terrorist content online: safeguards and risksfor freedom of expression', *CDT Working Paper,* pp. 1–17.

Meskys, E. et al. (2020) 'Regulating Deep Fakes: Legal and Ethical Considerations', *Journal of Intellectual Property Law & Practice*, 15(1), pp. 24–31.

Mill, J. S. (1859) *On liberty*. London: John W. Parker and Son, West Strand.

Obar, J.A., Wildman, S. (2015) 'Social media definition and the governance challenge: An introduction to the special issue', *Telecommunications policy,* 39(9), pp. 745–750.

Poell, T., van Dijck, J. (2015) 'Social media and activist communication' in Atton, C. (ed.) *The Routledge Companion to Alternative and Community Media.* 1st London: Routledge, pp. 527–537.

Pollicino, O., Somaini, L. (2020) 'Online disinformation and freedom of expression in the democratic context: The European and Italian responses' in Baume, S. et al. (eds.) *Misinformation in referenda.* 1st London and New York: Routledge, pp. 171–193.

QUINTAIS, J. P. (2020) 'The new copyright in the Digital Single Market Directive : a critical look', *European Intellectual Property Review,* 42(1), pp. 28–41.

TUCKER, J. A. ET AL. (2018) 'Social media, political polarization and political disinformation: a review of the scientific literature', Hewlett Foundation, pp. 1–95. Available at: https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf (Accessed: 3 May 2021).

SARAPIN, S., MORRIS, P. (2014) 'When "Like"-Minded People Click: Facebook Interaction Conventions, the Meaning of "Speech" Online, and Bland v. Roberts', *First Amendment Studies,* 48(2), pp. 131–157, https://doi.org/10.1080/21689725.2014.962557.

SCHMITZ-BERNDT, S., BERNDT, C. (2018) 'The German Act on Improving Law Enforcement on Social NEtworks (NetzDG): A Blunt Sword?', *University of Luxembourg Working Paper,* pp. 1-41. Available at: https://orbilu.uni.lu/handle/10993/45125 (Accessed: 7 May 2021).

SELTZER, W. (2010) 'Free Speech Unmoored in Copyright's Safe Harbor: Chilling Effects of the DMCA on the First Amendment', *Harvard Journal of Law & Technology,* 24(1), pp. 171–232.

VERSTRAETE, M. ET AL. (2017) 'Identifying and Countering Fake News', *Arizona Legal Studies Discussion Paper,* 17(15), pp. 1–38.

VOLOKH, E., FALK, D. M. (2012) 'Google First Amendment Protection for Search Engine Search Results', *Journal of Law, Economics & Policy*, 8(4), pp. 883–899.

ZUCKERBERG, M. 'A Blueprint for Content Governance and Enforcement', 15 November 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634