# ALGORITHMIC CONTENT MODERATION ON SOCIAL MEDIA IN EU LAW: ILLUSION OF PERFECT ENFORCEMENT[1] [2]

*Céline Castets-Renard*†

*Abstract*

*Intermediaries today do much more than passively distribute user content and facilitate user interactions. They now have near-total control of users' online experience and content moderation. Even though these service providers benefit from the same liability exemption regime as technical intermediaries (E-Commerce Directive, Art. 14), they have unique characteristics that must be addressed. Consequently, many debates are ongoing to decide whether or not platforms should be more strictly regulated.*

*Platforms are required to remove illegal content in the event of notice and take-down procedures built on automated processing and are equally encouraged to take proactive and automated measures to detect and remove it. Algorithmic decision-making helps scale down the massive task of content moderation. It would, therefore, seem that algorithmic decision-making would be the most effective way to provide perfect enforcement.*

*However, this is an illusion. A difficulty occurs when deciding what, precisely, is illegal. Platforms manage the removal of illegal content automatically, which makes it particularly challenging to verify that the law is being respected. The automated decision-making systems are opaque, and many scholars have shown that the main problem here is the over-removal chilling effect. Moreover, content removal is a task which, in many circumstances, should not be automated, as it depends on an appreciation of both the context and the rule of law.*

*To address this multi-faceted issue, this article offers solutions to improve algorithmic accountability and to increase the transparency around automated*

1. This paper was presented to the (Im)Perfect Enforcement Conference organized by ISP, Yale Law School, on April 6–7, 2019. The author thanks all the participants for their very helpful and relevant comments, especially Amélie Heldt for leading the panel discussion. The author is also grateful to Jack Balkin and Rebecca Crootof for this wonderful academic year as a Visiting Fellow at ISP.

2. Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) is gratefully acknowledged, as well as the Civil Law Faculty of the University of Ottawa. I also thank Harleen Kaur Kaloty for her assistance.

† Law Professor, University of Ottawa (Canada); Chair holder *Accountable AI in a Global Context*, uOttawa; Chair holder *Law, Accountability and Social Trust in AI*, ANR-3IA, ANITI (France).

*decision-making. Improvements may be made specifically by providing platform users with new rights, which in turn will provide stronger guarantees for judicial and non-judicial redress in the event of over-removal.*

TABLE OF CONTENTS



## I. INTRODUCTION

In November 2018, Mark Zuckerberg announced his vision for how content should be moderated and enforced on Facebook.[3] Zuckerberg laid out a plan detailing a new way for people to appeal content moderation decisions to an "Oversight Board."[4] This model of governance shows that Facebook is the

3. *See* Casey Newton, *Facebook will Create an Independent Oversight Group to Review Content Moderation Appeals*, VERGE (Nov. 15, 2018, 2:29 PM), https://www.theverge.com/2018/11/15/18097219/facebook-independent-oversight-supreme-court-content-moderation (explaining Facebook will have an independent oversight body in an effort to expand the rights of free speech).

4. Brent Harris, *Establishing Structure and Governance for an Independent Oversight Board*, FACEBOOK (Sept. 17, 2019), https://about.fb.com/news/2019/09/oversight-board-structure.

new lawmaker and judge when it concerns moderating content[5] and monitoring the freedom of expression. No doubt that these platforms are the new governors.[6] Even more, the separation of power is no longer respected and the system of checks and balances between the different branches of power has become fractured.[7] Worse, platforms offer no accountability over these decisions, especially when they are made by algorithmic governance.[8] They are, "politically-unaccountable technology oligarchs (who) exercise state-like censorship powers."[9] It is becoming increasingly clear that we are facing an "information fiduciary" issue.[10] Consequently, many debates are ongoing to determine whether or not platforms should be subject to more thorough regulation.[11]

### *A. Intermediaries' Liability and Content Moderation by Platforms*

The United States and European Union have different approaches to platforms regulation.[12] In the European Union, E-Commerce Directive 2000/31/EC of 8 June 2000[13] provides a liability exemption regime in favor of internet intermediaries[14] concerning illegal content and activities online.[15] The exemptions from liability only covers cases where the "information society service" provider's activity is limited to the technical operation process.[16] This is intended to provide access to a communication network over which information made available by third parties is either transmitted or temporarily stored, for the sole purpose of making the transmission more efficient.[17] The goal of the liability exemption regime was to encourage the development of the

5. *See generally* Tomer Shadmy, *The New Social Contract: Facebook's Community and Our Rights*, 37 B.U. INT'L L. J. (2019) (explaining the growing role of digital platforms on regulating its users).

6. *See generally* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018) (discussing the how digital platforms are monitoring its content and users).

7. *See* Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 101 (2008) (explaining the need for legislation and executive decision making on freedom of expression is necessary because of the growing irrelevancy of judicial doctrine regarding this topic).

8. *See* Eldar Haber, *Privatization of the Judiciary*, 40 SEATTLE U. L. REV. 115 (2016) (analyzing the judicial role delegated to search engines in the EU and the consequences of this role).

9. Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L. J. 1353, 1358 (2018).

10. *See generally* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183 (2016) (explaining how digital platforms that deal with personal information are characterized as information fiduciaries).

11. *See* TIFFANY LI, BEYOND INTERMEDIARY LIABILITY: THE FUTURE OF INFORMATION PLATFORMS (Yale Law School 2018) (discussing the possible duties of digital platforms).

12. *See generally* De Gregorio, *Democratising Online Content Moderation: A Constitutional Framework*, COMPUTER L. & SECURITY REV. (2019) (evaluating the democratic idea of freedom of speech in the digital world).

13. *See generally* Council Directive 2000/31, of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market, 2000 O.J. (L 178) 1 [hereinafter Electronic Commerce Directive] (explaining electronic commerce).

14. *Id*. at art. 14 at 13.

15. *Id*. at art. 12–15 at 12–13.

16. *Id*. at art. 12 at 12.

17. *Id*. at recital ¶ 42 at 6.

internet[18] and avoid disparities of liability rules in member states' legislation and case-law.[19] In addition to the liability exemption regime, there is no general obligation to monitor content.[20] However, such an exemption supposes that these intermediaries have neither actual knowledge of illegal activity and information nor awareness of facts and circumstances from which the illegal activity and information is apparent.[21] Upon obtaining such knowledge or awareness, these providers are required to act expeditiously to remove or disable access to the information[22] based on notice and take-down procedure.[23] The Directive also states that, "the removal or disabling of access has to be undertaken in observance of the principle of freedom of expression and within the procedural framework established for this purpose at the national level."[24] A court or administrative authority can also require the service provider to terminate or prevent an infringement.[25] Finally, the EU member states can establish procedures governing the removal of, or disabling access to, information.[26]

Beyond hosting service providers and internet access providers, the European Court of Justice (ECJ) made a broad interpretation of this liability regime for intermediaries and applied it to search engines[27] and marketplaces.[28] The ECJ interpreted recital 42[29] and considered that exemptions from liability cover cases in which the information society service provider's activity is, "of a mere technical, automatic and passive nature."[30] The ECJ applied the same

18. *Id*.

19. *See id*. at recital ¶ 42 at 6 ("both existing and emerging disparities in Member states' legislation and case-law concerning liability of service providers acting as intermediaries prevent the smooth functioning of the internal market, in particular by impairing the development of cross-border services and producing distortions of competition.").

20. *Id*. at art. 15 at 13; *see id*. at recital ¶ 47 at 6 (concluding that, "member states are prevented from imposing a monitoring obligation on service providers only with respect to obligations of a general nature; this does not concern monitoring obligations in a specific case and, in particular, does not affect orders by national authorities in accordance with national legislation.").

21. *Id*. at art. 12–15 at 12–13.

22. *Id*. at art. 14 §§ 1–2, at 13; *see id*. at recital ¶ 40 at 6 ("[S]ervice providers have a duty to act, under certain circumstances, with a view to preventing or stopping illegal activities.").

23. *See id*. at recital ¶ 40 at 6 ("[T]his Directive should constitute the appropriate basis for the development of rapid and reliable procedures for removing and disabling access to illegal information.").

24. *Id*. at recital ¶ 46 at 6.

25. *Id*. at art. 14 § 3 at 13; *see id*. at recital ¶ 45 at 6 ("[T]he limitations of the liability of intermediary service providers established in this Directive do not affect the possibility of injunctions of different kinds; such injunctions can in particular consist of orders by courts or administrative authorities requiring the termination or prevention of any infringement, including the removal of illegal information or the disabling of access to it.").

26. *Id*. at art. 14 § 4 at 13; *see id*. at recital ¶ 48 at 6 ("[T]his Directive does not affect the possibility for Member states of requiring service providers, who host information provided by recipients of their service, to apply duties of care, which can reasonably be expected from them and which are specified by national law, in order to detect and prevent certain types of illegal activities.").

27. *See* Cases C-236/08 to C-238/08, Google France SARL v. Louis Vuitton Malletier SA, 2010 E.C.R. I-02417 (holding that search engines can allow advertisers to use trademarks as keywords).

28. *See* Case C-324/09, L'Oréal SA v. eBay Int'l AG 2011 E.C.R. I-106011 (holding that digital marketplaces can allow use of trademarks).

29. Electronic Commerce Directive, *supra* note 13, recital ¶ 42 at 6 ("[T]he activity of the information society service provider . . . . is of a mere technical, automatic and passive nature, which implies that the information society service provider has neither knowledge of nor control over the information which is transmitted or stored.").

30. *Id*.

reasoning to the online marketplace, depending on whether or not the operator has played an active role which would allow the platform to have knowledge or control over the data stored.[31] Therefore, the operator plays an active role when it provides assistance which entails, in particular, promoting or optimizing the presentation of offers for sale.[32]

In the US, Section 230 of the Communications Decency Act (CDA), enacted in 1996, protects intermediaries from liability for distributing third-party user content[33] based on a "Good Samaritan" rule,[34] with the exception[35] of certain laws: criminal law, intellectual property law,[36] state law, communications privacy law, and sex trafficking law.[37] The goal was to encourage platforms to "clean up" offensive online material.[38] Moreover, a strong conception of freedom of speech pursuant to the First Amendment explains the tendency of American lawmakers to reject the idea of intermediaries' intervention on the content.[39] However, in copyright law, Section 512 of the *Digital Millennium Copyright Act* (DMCA) provides a safe harbor general regime[40] in addition to a notice-and-take-down procedure[41] similar to Article 14 of the E-Commerce Directive. The service provider must respond expeditiously to remove material that is the subject of a notice of infringement.[42] It must also have a policy to terminate the accounts of "repeat infringers."[43] The safe harbor is suspended when the provider has knowledge of specific infringement activity and still does not act.[44] This is also the case when the provider has both a "financial benefit directly attributable to" the infringement and the "right and ability to control" it.[45] American Courts have

31. *See* L'Oréal SA, 2011 E.C.R. I-106011 (holding that online marketplaces could be liable for its users).

32. *Id*.

33. Telecommunications Act of 1996, 47 U.S.C. § 230(c)(1) ("[N]o 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.").

34. *Id*. § 230(c)(2) ("[N]o provider or user of an interactive computer service shall be held liable on account of—(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or (B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).").

35. *Id*. § 230(e).

36. *See* 17 U.S.C § 512 (limiting liability on copyright law).

37. *See* Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115–164, 132 Stat. 1253 (stating that website publishers would be responsible if third parties are found to be posting ads for prostitution—including consensual sex work—on their platforms); *see also* DAPHNE KELLER, SESTA AND THE TEACHINGS OF INTERMEDIARY LIABILITY (Nov. 2, 2017), available at https://cyberlaw.stanford.edu/files/publication/files/SESTA-and-IL-Keller-11-2.pdf (discussing intermediate liability in the context of SESTA).

38. Danielle Keats Citron, *Section 230's Challenge to Civil Rights and Civil Liberties*, (University of Maryland Francis King Carey School of Law Legal Studies Research Paper No. 2018-18, 2018).

39. Michal Lavi, *Do Platforms Kill?*, 43 HARV. J.L. & PUB. POL'Y 477, 509 (2020).

40. *See generally* Matthew Sag, *Internet Safe Harbors and the Transformation of Copyright Law*, 93 NOTRE DAME L. REV. 499 (2017) (discussing the effects of copyright law on internet users).

41. *See generally* ANNEMARIE BRIDY & DAPHNE KELLER, U.S. COPYRIGHT OFFICE SECTION 512 STUDY: COMMENTS IN RESPONSE TO NOTICE OF INQUIRY (March 30, 2016), available at https://ssrn.com/abstract=2920871 or http://dx.doi.org/10.2139/ssrn.2920871 (explaining the process of notice-and-take-down).

42. 17 U.S.C. § 512(c)(1)(C).

43. *Id*. § 512(i)(1)(A).

44. *Id*. § 512(c)(1)(A)(iii).

45. *Id*. § 512(c)(1)(B).

read Section 230 broadly,[46] deciding to include platforms,[47] and have created a useful experience for other countries[48] by making reference to due process.[49] On this legal basis, immunity was established for intermediaries who do all but "materially contribute" to the user content they distribute.[50] Moreover, as online service providers are very often insulated from liability, they have built a wide range of different applications and services.[51]

### *B. Content Moderation: from Notice and Take Down Procedure to Proactive Measures*

The immunity from liability regime demonstrates a neutral conception of how online intermediaries operate. Intermediaries today do much more than passively distribute user content and facilitate user interactions.[52] Intermediaries now have near-total control of users' online experience.[53] Moreover, platforms cover a wide-ranging set of activities, including online advertising platforms, marketplaces, search engines, social media, creative content outlets, application distribution platforms, communication services, payment systems, and platforms for the collaborative economy.[54] Even if these service providers benefit from the same liability exemption regime as technical intermediaries, they present some important and specific characteristics.[55] Consequently, many European and American scholars argue for regulating certain types of online

46. Olivier Sylvain, *Discriminatory Designs on User Data*, *in* KNIGHT FIRST AMEND. INST.'S EMERGING THREAT SERIES (David Pozen ed., 2018), https://papers.ssrn.com/abstract=3157975; Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 413 (2017).

47. Sharon Bar-Ziv & Niva Elkin-Koren, *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown*, 50 CONN. L. REV. 339, 361 (2017).

48. Jeff Kosseff, *Twenty Years of Intermediary Immunity: The US Experience*, 14 SCRIPTED 5 (2017), https://papers.ssrn.com/abstract=3225773.

49. Daphne Keller, *Toward a Clearer Conversation About Platform Liability*, *in* KNIGHT FIRST AMEND. INST.'S EMERGING THREATS SERIES (David Pozen ed., 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186867.

50. Sylvain, *supra* note 46, at 407 n.56.

51. Balkin, *supra* note 7, at 108.

52. Sylvain, *supra* note 46, at 4.

53. *Id.*

54. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Online Platforms and the Digital Single Market: Opportunities and Challenges for Europe*, COM (2016) 288 final (May 25, 2016) [hereinafter *Online Platforms*].

55. *Id.* at 2–3 ("In particular: they have the ability to create and shape new markets, to challenge traditional ones, and to organize new forms of participation or conducting business based on collecting, processing, and editing large amounts of data; they operate in multisided markets but with varying degrees of control over direct interactions between groups of users; they benefit from 'network effects', where, broadly speaking, the value of the service increases with the number of users; they often rely on information and communications technologies to reach their users, instantly and effortlessly; they play a key role in digital value creation, notably by capturing significant value (including through data accumulation), facilitating new business ventures, and creating new strategic dependencies.").

platforms more vigorously[56] and for nuancing the immunity doctrine[57] or for revising Section 230.[58]

Taking this shift into account, the European Commission maintains a predictable liability regime for online platforms in order to push forward the development of the digital economy in the EU and to facilitate investment in platform ecosystems.[59] At the same time, many specific issues relating to illegal and harmful online content and activities have been identified which need to be addressed to render this approach sustainable.[60] Nearly everywhere, the public has called for the removal of illegal content[61] based on notice and take-down procedure. Citizen groups have equally pushed for further cooperation between platforms, lawmakers, and courts.[62] Content moderation can be defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."[63] The European Commission is pursuing the goal of tackling illegal content online,[64] such as hate speech,[65] child pornography,[66] terrorist propaganda,[67] privacy,[68] and copyright infringement.[69] The goal is to prevent, detect, remove, and disable access to illegal content.[70] In addition to its communication on "tackling illegal content online" in 2017,[71] the European Commission has also focused on enhancing the liability of online

56. Balkin, *supra* note 10, at 1183.

57. Sylvain, *supra* note 46, at 19 (urging Congress to maintain the immunity but to create an explicit exception from the safe harbor for civil rights violations).

58. Citron, *supra* note 38, at 6–7.

59. *Online Platforms*, *supra* note 54, at 8.

60. *Id.*

61. Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money*, *in* HOOVER INSTITUTION'S AEGIS PAPER SERIES (2018), https://papers.ssrn.com/abstract=3262936.

62. Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035 (2018); Citron & Wittes, *supra* note 46, at 41.

63. James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 47 (2015).

64. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms*, COM (2017) 555 final (Sep. 28, 2017) [hereinafter Tackling Illegal Content].

65. Commission Code of Conduct on Countering Illegal Hate Speech Online of May 31, 2016, https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985 [hereinafter Code of Conduct].

66. Directive 2011/93/EU, of the European Parliament and of the Council of 13 December 2011 on Combating the Sexual Abuse and Sexual Exploitation of Children and Child Pornography, and Replacing Council Framework Decision 2004/68/JHA, 2011 O.J. (L 335) 1 [hereinafter Combatting Sexual Abuse].

67. Directive 2017/541, of the European Parliament and of the Council of 15 March 2017 on Combating Terrorism and Replacing Council Framework Decision 2002/475/JHA and Amending Council Decision 2005/671/JHA, 2017 O.J. (L 88) 6 [hereinafter Combatting Terrorism Framework]; *see also Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online*, COM (2018) 640 final (Sept. 12, 2018) (establishing legal framework to prevent misuse of hosting services for terrorist content) [hereinafter Terrorist Content Online].

68. Council Regulation 2016/679 of Apr. 27, 2016, on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

69. Directive 2001/29/EC, of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society and its Reform, 2001 O.J. (L 167) 10 [hereinafter Copyright Harmonisation]; Directive 2019/790, of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, 2001 O.J. (L 130) 92 [hereinafter Digital Single Market].

70. *See* Tackling Illegal Content, *supra* note 64, at 3.

71. *Id.* at 1.

platforms.[72] Without reconsidering the exemption regime, the lawmaker is now encouraging a cooperative approach with private actors to ensure respect of the rule of law. Thus, the European Commission's goals are ambiguous, falling between respect of the immunity regime and content censorship. "Governance by proxy"[73] and "collateral censorship"[74] appear when a state coerces private companies to censor speech that the government could not itself lawfully sanction.[75] Concretely, in addition to the notice and take-down procedure enacted in Directive 2000/31/EC and several additional directives,[76] the European Commission also encourages platforms to take proactive measures[77] and to detect and remove illegal online content.[78] Notice and take-down procedure involves both *ex-post* moderation and *ex-ante* proactive measures.[79] *Ex-ante* moderation supposes acting on software's architectural features and applying the same rules to all content, while *ex-post* moderation is a law-like technique which directs attention only where it is needed.[80] From the beginning, such proactive measures were provided by recommendations[81] and codes of conduct[82] which can be referred to as "soft laws." Now, they are also enacted by directives and have consequently become hard laws.[83] These provisions may or may not be binding, depending on the type of content and interests to protect. Consequently, the "open and free" speech ideal has been replaced by creating a "healthy" and a "safe" speech environment online.[84] This narrative change is also a change in substance and the acceptance of harmful speech is far closer to the European speech tradition and its bans on hate speech.[85]

---

72. *Id*. at 10–12.

73. Niva Elkin-Koren & Eldar Haber, *Governance by Proxy: Cyber Challenges to Civil Liberties*, 82 BROOK. L. REV. 105, 117 (2016).

74. Jack M. Balkin, *Old School/New School Speech Regulation*, 127 HARV. L. REV. 2296, 2298 (2014).

75. Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27 (2018).

76. *See* Combating Terrorism Framework, *supra* note 67 (obliging member states to take the necessary measures to ensure the prompt removal of online content inciting to commit terrorist acts (article 21)); *see also* Combatting Sexual Abuse, *supra* note 66 (concerning child pornography (article 25)).

77. *See* Commission Recommendations of Mar. 1, 2018, on Measures to Effectively Tackle Illegal Content Online, C(2018) 1177 final, 1, 1–2 (Mar. 1, 2018) ("In addition to notice-and-action mechanisms, proportionate and specific proactive measures taken voluntarily by hosting service providers, including by using automated means in certain cases, can also be an important element in tackling illegal content online, without prejudice to Article 15(1) of Directive 2000/31/EC.") [hereinafter Measures to Effectively Tackle Illegal Content].

78. Tackling Illegal Content, *supra* note 64, at 3.

79. Grimmelmann, *supra* note 63, at 47.

80. *Id*.

81. Measures to Effectively Tackle Illegal Content, *supra* note 77, at 1.

82. Code of Conduct, *supra* note 65, at 1; European Commission, *Code of Practice on Online Disinformation* (2018) (adoption encouraged by the E-Commerce Directive 2000/31/EC of June 8, 2000); *see* Directive 2000/31/EC, of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market (Directive on Electronic Commerce), 2000 O.J. (L 178) 1, 6–7 [hereinafter Information Society Services] ("Member states and the Commission are to encourage the drawing-up of codes of conduct; this is not to impair the voluntary nature of such codes and the possibility for interested parties of deciding freely whether to adhere to such codes.").

83. *See, e.g.*, Terrorist Content Online, *supra* note 67, at 9.

84. Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering* Systems, 119 COLUM. L. REV. 2001, 2009 (2019).

85. *Id.* at 2010.

### *C. Content Moderation and Law Enforcement by Algorithmic Decision-Making Systems*

The content moderation can be made by content moderators who are digital laborers[86] mainly located in India and the Philippines, traditional destinations for Business Process Outsourcing (BPO).[87] Social media must balance the application of automated, algorithmic systems for content moderation (like Microsoft's PhotoDNA[88], and YouTube's Content ID[89]) with teams of human moderators.[90] But all the major speech platforms use a mixture of software, humans following rules, and humans deliberating the enforcement and improvement of their content rules.[91]

Notice and take-down procedures are built on automated processing. Moreover, lawmakers encourage platforms to use these technologies to detect and remove illegal content. In this sense, in 2016,[92] and again in 2017,[93] the European Commission strongly encouraged online platforms to use voluntary, proactive measures and automatic detection technologies. According to the European Commission, principles-based, self-regulatory/co-regulatory measures—including industry tools for ensuring application of legal requirements and appropriate monitoring mechanisms—have an essential role to play.[94] The purpose is to intensify the implementation of "good practices"[95] which prevent, detect, remove, and disable access to illegal content. Moreover, when underpinned by appropriate monitoring mechanisms, they can strike the right balance between predictability, flexibility, efficiency, and the need to develop future-proof solutions.[96] The platforms appear useful when it concerns automatically enforcing the Courts' decisions and laws, especially extraterritorial provisions that don't require an *exequatur* procedure.[97] Consequently, algorithmic decision-making helps to scale content moderation

86. *See* The Verge, *Inside the Traumatic Life of a Facebook Moderator*, YOUTUBE (June 19, 2019), https://youtu.be/bDnjiNCtFk4 (explaining that Facebook employs 15,000 moderators around the world, who typically spend six full hours a day reviewing reported content and high performing moderators will look at 400 or more post per day, including graphic imagery and hates speech); *Facebook Failing to Protect Moderators from Mental Trauma, Lawsuit Claims*, GUARDIAN (Sept. 24, 2018), https://www.theguardian.com/technology/2018/sep/24/facebook-moderators-mental-trauma-lawsuit.

87. MacKenzie Common, *Fear the Reaper: How Content Moderation Rules are Enforced on Social Media*, 34 INT'L REV. OF L., COMPUTERS AND TECH. 126, 141 (2020).

88. *PhotoDNA*, MICROSOFT, https://www.microsoft.com/en-us/photodna (last visited Oct. 21. 2020).

89. *How Content ID works*, GOOGLE, https://support.google.com/youtube/answer/2797370?hl=en (last visited Oct. 21, 2020).

90. Andrew Quodling, *Anxieties Over Livestreams Can Help Us Design Better Facebook and YouTube Content Moderation*, CONVERSATION (Mar. 19, 2019), http://theconversation.com/anxieties-over-livestreams-can-help-us-design-better-facebook-and-youtube-content-moderation-113750.

91. Wu, *supra* note 84, at 2013.

92. *Online Platforms*, *supra* note 54.

93. Tackling Illegal Content, *supra* note 64.

94. European Commission, *supra* note 82.

95. *See, e.g.*, GOOGLE, *Removing Content from Google*, https://support.google.com/legal/troubleshooter/1114905?hl=en (last visited Oct. 21, 2020) (providing a mechanism for removing illegal content from Google searches).

96. European Commission, *supra* note 82.

97. *See id.* at 5 (emphasizing the importance of cooperation between authorities).

and automatically enforces European and State laws, as well as the platforms' private rules.

In Part II, I will discuss the issues around the increased stringency and cooperation between lawmakers and platforms in monitoring online content. EU soft and hard laws provide for the application of duty of care by hosting service providers when taking action for dissemination of certain illegal content online.[98] Some provisions require that service providers and hosting platforms take effective and proportionate proactive measures where appropriate to mitigate users' risk of exposure and dissemination of illegal content on their services.[99] On the other hand, by encouraging and enacting certain proactive actions, I argue that the European lawmakers risk the infringement of users' fundamental rights, particularly freedom of speech.

In Part III, I will present another problem that occurs when platforms use algorithmic decision-making systems as a means to guarantee perfect enforcement. These measures in favor of content control and perfect law enforcement merely create an illusion of monitoring while being inefficient.[100] Delegating law enforcement to private platforms has its costs,[101] and states are dependent on the decisions made by the platforms. Lawmakers should establish the degree of power given to private actors[102] and monitor its use to preserve users' fundamental rights.[103] As it currently stands, the European lawmaker is failing, on the one hand, to tackle illegal content and, on the other hand, to prevent the over-removal of content.[104] Moreover, lawmakers, courts, and other public authorities have no monitoring capacities regarding the respect of the law and are not able to oversee the "chilling effects," such as over-removal.[105] In practice, the decisions are carried out exclusively by online platforms and neither national courts nor independent public authorities can oversee them.[106] The trend is in favor of an over-evaluation of private rules and decisions made by online platforms, rather than consideration of the rule of law.[107] Finally, it is unclear how authorities can require more cooperation from the platforms to remove illegal content without increasing their liability and accountability. Consequently, I argue that the use of algorithmic decision-making systems is non-transparent and generates a significant risk of infringement of users' fundamental rights, as well as procedural guarantees. The "perfect enforcement" is ambiguous and even a controversial goal. In certain circumstances, an "imperfect enforcement" could be an achievable purpose when the law requires

98. *See, e.g.*, European Commission, *supra* note 82 (providing recommendations on tackling hate speech); *see also* Combatting Sexual Abuse, *supra* note 66 (requiring procedures for addressing child pornography).

99. *Id.* at 11–12.

100. Keller, *supra* note 61, at 28.

101. *Id.* at 1.

102. Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. 1149 (2018).

103. *Id*.

104. Keller, *supra* note 61, at 1.

105. *Id*. at 14.

106. Aleksandra Kuczerawy, *The EU Commission on Voluntary Monitoring: Good Samaritan 2.0 or Good Samaritan 0.5?*, CITIP BLOG (Apr. 24, 2018), https://www.law.kuleuven.be/citip/blog/the-eu-commission-on-voluntary-monitoring-good-samaritan-2-0-or-good-samaritan-0-5.

107. Keller, *supra* note 61, at 4.

discretion or a case by case interpretation, especially to apply exceptions. The problem with algorithmic systems is that they make the decisions automatically, even in irrelevant cases.

In part IV, I will make recommendations on how to ensure more accountability for algorithmic decision-making and automated enforcement actions in the EU. I will also address the issue of remedies in case of over-removal. Algorithmic decision-making systems are widely used by platforms that manage a significant amount of internet content.[108] As a tool of law enforcement, it is essential to make algorithmic systems more transparent and monitored by lawmakers and courts. Additionally, safeguards to moderate illegal content on online platforms have to be provided to protect the users' fundamental rights.[109]

## II. PROACTIVE MEASURES TO TACKLE ILLEGAL CONTENT AND RISKS FOR FUNDAMENTAL RIGHTS

Without questioning the immunity regime of the E-Commerce Directive, the European Commission considers that online platforms that mediate access to content carry a significant societal responsibility in terms of protecting users and society at large, and are therefore responsible for preventing criminals and other persons involved in infringing activities online from exploiting their services.[110] In this context, the European Commission has recently increased cooperation with platforms and intensified the implementation of good practices for removing illegal content.[111] Now, platforms are taking more proactive measures without prior notice.[112] From the beginning, proactive measures were deployed through non-binding rules (soft laws) (A). Today, binding regulations include proactive measures (hard laws) (B). These rules are dissimilar to the "Good Samaritan" American Clause and raise questions around the platform liability regime (C).

### *A. Soft Laws*

To complete the Communication on *Tackling Illegal Content Online*,[113] the European Commission enacted several Recommendations (1) and two Codes of conduct (2).

108. Balkin, *supra* note 102, at 1149.
109. *See generally* Keller, *supra* note 61 (arguing for reform in platform liability laws).
110. Tackling Illegal Content, *supra* note 64.
111. *Id.*
112. *See id.* (suggesting enhanced responsibility for platforms in online hate speech); Measures to Effectively Tackle Illegal Content, *supra* note 77 (establishing a framework to counter illegal content online); Code of Conduct, *supra* note 65 (providing recommendations for tackling online hate speech).
113. Tackling Illegal Content, *supra* note 64.

*1. Recommendations on Measures to Effectively Tackle Illegal Content Online*

The Recommendations focus on measures to effectively tackle illegal content online (2018).[114] The Member states and hosting service providers are encouraged to take effective, appropriate, and proportionate measures to address illegal content online, in full compliance with the EU Charter of Fundamental Rights, in particular the right to freedom of expression and information, and other applicable provisions of Union law, particularly regarding the protection of personal data, competition, and electronic commerce.[115] Platforms are merely "encouraged" to respect fundamental rights when they remove presumably illegal content.[116] Though the nature of soft law is to be flexible and not mandatory, the problem here is the possible infringement of fundamental rights require more stringent protection.

However, some of the other Recommendations are more practical and involve more active engagement. For instance, the Recommendations state that the mechanism to submit notices should be easy to access, user-friendly, and allow electronic means.[117] Moreover, when the contact details of the notice provider are known, the hosting service provider should send a confirmation of receipt without undue delay to inform of its decision.[118] When a hosting provider decides to remove or disable access to any content, the content provider should be informed of that decision and of the reasons for taking it, as well as the possibility to contest.[119] The term 'should' is ambiguous and can design only a desirable result and not a mandatory one, but some exceptions are provided for. This is notably the case when a competent authority requests the removal of content, or when it is manifest that the content is illegal and relates to serious criminal offenses involving a threat to life, safety, or persons.[120] The hosting service providers are put in the situation of assessing the case and applicable law, thus exercising judicial power. Finally, hosting service providers are encouraged to publish clear, easily understandable, and sufficiently detailed explanations of their policy on content removal.[121] This is only an incentive to communicate general rules, while the most important consideration here is the nature of the given explanation, and not its principle. Consequently, hosting services providers make the rules and act as lawmakers.

*2. Codes of Conduct on Hate Speech and Disinformation*

Other soft norms concern the freedom of speech, especially hate speech and disinformation. This primarily regards the Code of Conduct on *Countering Illegal Hate Speech Online* (2016).[122] This voluntary process has provided

114. Measures to Effectively Tackle Illegal Content, *supra* note 77.
115. Tackling Illegal Content, *supra* note 64, at 9.
116. *Id.*
117. *Id.* at 11.
118. *Id.*
119. *Id.*
120. *Id.*
121. *Id.* at 12.
122. Code of Conduct, *supra* note 65.

indicative targets for removal times which are twenty-four hours for the majority of cases.[123] Moreover, IT companies have committed to provide information for submitting notices, in the goal of improving the efficacy of communication between member state authorities and IT companies for removing illegal hate speech.[124] They have also committed to relying on support from member states and the European Commission to ensure access to a representative network of Civil Society Organization (CSO) partners, as well as "trusted reporters" to help provide high-quality removal notices.[125]

But many provisions of this Code of Conduct are vague and non-binding.[126] For instance, IT companies and the European Commission, "recogni[ze] the value of independent counter speech against hateful rhetoric and prejudice, [and therefore] aim to continue their work in identifying and promoting independent counter-narratives, new ideas and initiatives and supporting educational programs that encourage critical thinking."[127] Such provisions are only "good intentions" which are not detailed and, worse, could present dangers for the freedom of speech.[128]

The Communication on *Tackling Online Disinformation*, published on April 26, 2018[129] and completed in September 2018 by a Code of Practice,[130] addresses the spread of online disinformation and fake news. The goal was to increase transparency on platforms such as *Facebook*, *Google*, and *Twitter*, especially before the European Parliament elections which took place in May 2019.[131] The Code includes an annex identifying best practices that signatories will apply to implement the Code's commitments.[132] In a context where the mechanisms of disinformation are algorithmic-based, advertising-driven, and technology-enabled,[133] one of the goals of the Code is to ensure that online service providers include safeguards against disinformation by design. It includes detailed information on the behavior of algorithms that prioritize the display of content as well as the development of testing methodologies.[134] Actions pursuant to these objectives should strictly respect freedom of expression.

In December 2018, the Commission published a Report assessing the progress made in the implementation of the Communication on *Online Disinformation* to evaluate whether further actions, including measures of

123. *Id.*
124. *Id.*
125. *Id.*
126. *Id.; see also* Measures to Effectively Tackle Illegal Content, *supra* note 64, at 1 (providing similar vague language contained in the Code of Conduct).
127. Code of Conduct, *supra* note 65.
128. *Id.*
129. *Communication from the Commission to the European Parliament, the Council, and the European Economic and Social Committee and the Committee of the Regions, Tackling Online Disinformation: A European Approach*, COM (2018) 236 final (Apr. 26, 2018) [hereinafter *Tackling Online Disinformation*].
130. EU Code of Practice on Disinformation (Jun. 28, 2018), https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.
131. *Id.*
132. *Id.*
133. Measures to Effectively Tackle Illegal Content, *supra* note 64, at 5.
134. *Id.* at 8.

regulatory nature, are necessary.[135] The Commission published the first Compliance Report of the Code on January 29, 2019[136] and calls on the signatories (*Google*, *Facebook*, *Twitter*, *Mozilla*, and the trade associations) to intensify their efforts.[137] Some progress has been made, notably in removing fake accounts and limiting the visibility of sites that promote disinformation,[138] but additional action is needed, especially to ensure full transparency of political ads. However, soft norms do not provide any fines or penalties if *Facebook, Google, Twitter,* and their peers fail to comply.[139]

These soft norms are non-binding and unclear about how the platforms are supposed to moderate content before prior notice. Despite this weakness, proactive measures are extended through hard laws.

### *B. Hard Laws*

Though the general principle is provided for by Article 14 of E-Commerce Directive of 8 June 2000, other sectorial rules have been enacted. The Child Pornography Directive (EU) 2011/93 (art. 25)[140] provides for, on the one hand, provisions against websites containing or disseminating this particular illegal content and, on the other hand, blocking the access to it. The Member states shall take the necessary measures to ensure the prompt removal of child pornography content hosted in their territory and endeavor to obtain the removal of such pages hosted outside of their territory (art. 25(1)). The member states shall also take measures to block access to the internet users who seek to disseminate and access this material within their territory (art. 25(2)). Other sectorial regulations on copyright (a) and terrorism (b) state several more stringent, proactive measures.

#### *1. Proactive Measures on Copyright*

Directive 2019/790 on Copyright in the Digital Single Market, adopted on 17 April 2019,[141] revises Directive 2001/29/EC of 22 May 2001 regarding the harmonization of certain aspects of copyright and related rights in the information society.[142]

Article 17 is the main provision regarding the content moderation of copyrighted content by online content sharing service providers.[143] An 'online content-sharing service provider' is:

135. *Tackling Online Disinformation*, *supra* note 129.

136. *First Results of the EU Code of Practice Against Disinformation*, EUR. COMM'N (Jan. 29, 2019), https://ec.europa.eu/digital-single-market/en/news/first-results-eu-code-practice-against-disinformation.

137. *Code of Practice Against Disinformation: Commission Calls on Signatories to Intensify Their Efforts*, EUR. COMM'N (Jan. 29, 2019), http://europa.eu/rapid/press-release_IP-19-746_en.htm.

138. *Id*.

139. Marietje Schaake, *Critics Take On "Nonsense" EU Plan to Fight Illegal Online Content*, MARIETJE SCHAAKE: DIGIT. AGENDA (Mar. 1, 2018), https://marietjeschaake.eu/en/critics-take-on-nonsense-eu-plan-to-fight-illegal-online-content.

140. Combatting Sexual Abuse, *supra* note 66, at art. 25.

141. Digital Single Market, *supra* note 69.

142. Copyright Harmonisation, *supra* note 69.

143. Digital Single Market, *supra* note 69, at art. 17.

> a provider of an information society service of which the main or one of the main purposes is to store and give the public access to a large amount of copyright-protected works uploaded by its users, which it organizes and promotes for profit-making purposes (art. 2(6)).[144]

For instance, YouTube is an "online content-sharing service provider" which would be subject to the provision in Article 17.[145] However, many platforms are excluded, such as not-for-profit online encyclopedias (e.g. Wikipedia), not-for-profit educational and scientific repositories, open-source software-developing and sharing platforms, electronic communication service providers,[146] online marketplaces, and cloud services.[147]

When online content-sharing service providers give the public access to copyright-protected works[148] uploaded by their users, they perform an act of communication to the public[149] (art. 17(1)). Therefore, they must obtain authorization from the right holders,[150] for instance by concluding a licensing agreement to distribute and make content available to the public. When an online content-sharing service provider obtains an authorization, that shall also cover acts carried out by users of the services,[151] provided that they are not acting on a commercial basis and their activity does not generate "significant" revenues (art. 17(2)).[152]

To know if the content contains copyright-protected work, the service provider has to monitor it. This seems to create an obligation to oversight, while article 15 of E-commerce Directive excludes a general obligation of monitoring. How should this new provision be interpreted? Is it an exception of the safe harbor liability regime applicable to the intermediaries enacted by article 12 to 15 of the E-Commerce Directive? Is it only a sectorial regime of liability solely applicable in copyright without attempting to the general regime?

Article 17(3) states that:

> when an online content-sharing service provider performs an act of communication to the public under the conditions laid down in this Directive, the limitation of liability established in Article 14(1) of Directive 2000/31/EC shall not apply to the situations covered by this Article.[153]

---

144. *Id.* at art. 12(6).

145. *Id.* at art. 17.

146. Directive 2018/1972/EC of the European Parliament and of the Council of 11 December 2018 Establishing the European Electronic Communications Code, 2011 O.J. (L 321) 61 [hereinafter European Electronic Communications Code].

147. Digital Single Market, *supra* note 69, at art. 2(6); *see also* European Electronic Communications Code, *supra* note 146 ("Cloud services" include "business-to-business cloud services and cloud services that allow users to upload content for their own use.").

148. *See* European Electronic Communications Code, *supra* note 146 ("Acts of communication also include permitting public access to other user-uploaded, protected subject matter.").

149. *See id.* ("Acts by online content-sharing service providers also include making copyright-protected works available to the public.").

150. Digital Single Market, *supra* note 69, at art. 3.

151. *See id.* (referring to members of the public who are accessing the materials communicated or made available by online content-sharing service providers).

152. Digital Single Market, *supra* note 69, at art. 17(2).

153. *Id.* at art. 17(3).

Consequently, the copyright directive has its own and specific scope besides the general E-commerce directive.[154] In its sense, article 17(3) adds a new regime of liability that shall not affect the application of Article 14(1) of Directive 2000/31/EC to those service providers for purposes falling outside the scope of this Directive. However, it may probably not be easy in practice to draw the borders between both.

In other words, these provisions create a new and specific regime of liability regarding copyright without reconsidering the general regime of exemption in itself. Online content-sharing service providers no longer benefit from the exemption regime of liability.[155] If any authorization to communicate the copyrighted works is granted, online content-sharing service providers shall be liable for this unauthorized diffusion. On the other hand, article 17(8) adds that the application of this Article shall not lead to any general monitoring obligation. But how the online content-sharing service provider may know if the content is protected by copyright without monitoring the full content?

However, article 17(4)) provides an exoneration of liability if the service provider demonstrates that he has:

> (a) made best efforts[156] to obtain an authorization, and (b) to ensure the unavailability of specific works for which the right holders have provided with the relevant and necessary information; and in any event (c) acted expeditiously, upon receiving a sufficiently substantiated notice from the right holders, to disable access to, or to remove from, their websites the notified works, and made best efforts to prevent their future uploads.[157]

This new liability regime enacts a notice and take-down procedure (under c) but also confirms the requirement of proactive measures to remove illegal content (under b)). Even if "proactive measures" and "automated decision systems" are not explicitly mentioned, the intention of the European lawmaker is obvious. Only an algorithmic system is able to ensure the unavailability of copyrighted works with the cooperation of the right holders.[158] For instance, a system such as YouTube's Content ID blocks uploaded videos that match an extensive list of copyrighted works.[159]

Consequently, this new liability regime obliges de facto the intermediaries in the deployment of filtering systems.[160]

Regarding the notice and take-down procedure, what are the differences between the general regime of the E-Commerce Directive (art. 14) and this specific liability regime applying for copyright (art. 17)? First, the principle is

154. Information Society Services, *supra* note 82.

155. Digital Single Market, *supra* note 69, at art. 17.

156. *Id* (explaining that best efforts must be made "with high industry standards . . . .").

157. *Id*.

158. Maxime Lambrecht, *Free Speech by Design: Algorithmic Protection of Exceptions and Limitations in the Copyright DSM Directive*, 11 J. INTELL. PROP., INFO. TECH., & E-COMMERCE L. 68, 71 (2020).

159. *See Using Content ID*, YOUTUBE, https://support.google.com/youtube/answer (last visited Oct. 21, 2020) (noting that Content ID is an automated system that enables copyright owners to identify videos that use their content).

160. Giancarlo Frosio, *From Horizontal to Vertical: An Intermediary Liability Earthquake in Europe*, 12 OXFORD J. OF INTELL. PROP. & PRAC. 565, 570 (2017).

the liability of online content-sharing service providers with conditions of exoneration, while article 14(1) states that the hosting service providers are not liable for the information stored at the request of a recipient, except if they have an actual knowledge of illegal content or awareness of facts.[161]

Second, the conditions of the exemption are different. They are mainly based on the authorization of communication to the public by the rights holders in article 17(4), while article 14(1) takes into account the knowledge or awareness of illegal activity or content.[162]

Third, there is a common provision regarding the notice and take-down procedure, which requires an expeditious action to remove the illegal content if the service provider receives a sufficiently substantiated notice. However, article 17(4) also requests that the service provider makes the best efforts to prevent their future uploads.[163] This additional requirement, so-called "notice and stay down" procedure, is particularly challenging to fulfill. This requirement is not set out in article 14.

Fourth, article 17(4) only provides an obligation of means. The content-sharing providers have to make their "best efforts" for obtaining authorization, ensuring the unavailability of specific works, and preventing their future uploads.[164] In comparison, article 14(1) does not indicate if the removal of illegal content is an obligation of means or results, but the European Court of justice interpreted it as an obligation of results.[165] The Court makes them liable in case of inaction in a very short time after the notice.

Fifth, and most importantly, article 17(4) enacts explicitly the principle of a cooperation between service providers and rights holders. The right holders have to provide for the relevant and necessary information to allow the unavailability of copyright-protected works.[166] Reciprocally, service providers give to the right holders, at their request, adequate information on the functioning of their practices[167] and, where licensing agreements are concluded between them, information on the use of content covered by the agreements (art. 17(8)). The cooperation is mentioned, and, even more, the conclusion of licensing agreements is encouraged.[168]

Finally, the obligations of the online content-sharing service providers may seem strict. However, they are reasonable in light of the exclusion of not-for-profit activities. Moreover, they have to be interpreted proportionally depending on:

161. Digital Single Market, *supra* note 69, at art. 17. *But see* Information Society Services, *supra* note 82, at 13.

162. *Id.*

163. Digital Single Market, *supra* note 69, at art. 17.

164. *Id.*

165. Case C-131/12, Google Spain SL v. Agencia Espanola de Proteccion de Datos and Mario Costeja Gonzalez, 2014 E.C.R. 317.

166. Digital Single Market, *supra* note 69, at art. 17.

167. *See id.* (stating online content-sharing services providers must provide right holders with information on how they disable access to and remove flagged material).

168. *Id.*

> (a) the type, the audience and the size of the service and the type of works or other subject matter uploaded by the users of the service; and (b) the availability of suitable and effective means and their cost for service providers (art. 17(5)).[169]

The "high industry standards of professional diligence" are flexible and may be appreciated following a case by case approach and regarding the context of the infringement.[170]

Furthermore, article 17(6) enacts a "light" regime of liability for small companies:

> the services available to the public for less than three years and which have an annual turnover below 10 million euros have to comply with article 17(4)(a), meaning they only have to do their best efforts to obtain an authorization. They also have to comply with the notice and take-down procedure.[171]

In comparison with the regime under 17(4)(c), no matter if the notice comes from the right holders or not. Most importantly, they are not supposed to make their best efforts to prevent the future uploads, except where the average number of monthly unique visitors of the service exceeds 5 million (art. 17(6)).[172] Consequently, this regime of liability applicable to small companies is dual, depending on the number of monthly unique visitors.[173] In any case, the small companies are exempted to comply with article 17(4)(b) and take proactive measures for removing illegal content.[174]

In brief, the copyright reform creates a new and sectorial regime of liability in addition to the general regime of the E-Commerce Directive. This regime is complex because it enacts new obligations, conditions of exoneration and burden of proof in the scope of the copyright law, as well as different level of obligations, regarding the context, the size and the activity of these online content-sharing service providers.[175] Three degrees of liability can be drawn within this new specific regime. The first level of liability is applicable to the smallest companies[176] and produces the same results as the general regime of E-Commerce Directive. The service providers are exempted from taking pro-active measures and preventing future uploads.[177] They are only expected to remove the infringed content after prior notification. At the second level of liability, the small companies, which capture a certain audience,[178] have to

169. *Id.*
170. *Id.*
171. *Id.*
172. *Id.*
173. *Id.*
174. *Id.*
175. *Id.*
176. *Id.* at 20. The smallest companies that the liability regime established by Directive 2019/790 applies to are online content-sharing service providers whose services have been "available to the public in the Union for less than three years and which have an annual turnover below EUR 10 million."
177. *Id.*
178. *See id.* (noting that in addition to the requirements imposed upon the smallest companies, services providers that average more than five million monthly visitors must also make "best efforts to prevent further uploads of the notified works" and other covered material.).

prevent the future uploads of copyrighted works. At the third level, the most prominent service providers have to comply with the same obligation as the others and, in addition, taking proactive measures for ensuring the unavailability of copyrighted works.[179]

On another matter, in order to prevent the over-removal problem, article 17(7) states that:

> the cooperation between online content-sharing service providers and right holders shall not result in the prevention of the availability of works uploaded by users, which do not infringe copyright, including where such works are covered by an exception or limitation.[180]

Such measure tempts to respect the balance between, on the one hand, the interests of the rights holders and, on the other hand, the exceptions and limitations in favor of the public. But interpreting the counterfeiting is particularly difficult, while article 5 of the Directive 2001/29/EC of 22 May 2001 on copyright and related rights in the information society[181] enacts twenty-one exceptions and limitations to the right of reproduction and communication to the public. Moreover, only one of them is mandatory in all member states,[182] while other exceptions have only been enacted in some member states. Consequently, there is not a single way to respect copyright within the EU. Faced to the reality of this fragmented landscape and patchwork of copyrights, how could the online content-sharing service providers guarantee the respect of the rule of law? How can the interpretation be made at the national level, when the service providers used unlocated algorithmic decision-making systems? Why does the lawmaker permit the interpretation of the law by private actors, especially when fundamental rights protected by the EU Charter are involved (art. 17(2))? Here is probably the most crucial problem and ambiguity of this new copyright directive. On the one hand, the use of proactive measures to moderate the content and, consequently, algorithmic decision systems are encouraged, and, on the other hand, intellectual property rights and exceptions have to be balanced.[183] Article 17(8) adds that:

> the member states shall ensure that users are able to rely on some existing exceptions or limitations when uploading and making available content generated by users on online content-sharing services, especially quotation, criticism, review, and use for the purpose of caricature, parody or pastiche.[184]

---

179. *See* Ali Amirmahani, *Digital Apples and Oranges: A Comparative Analysis of Intermediary Copyright Liability in the United States and European Union*, 30 BERKELEY TECH. L.J. 865, 866 (2015) (comparing the différences in copyrights laws between the United States and European Union).

180. Copyright Harmonisation, *supra* note 69.

181. *Id*. at 14.

182. *Id*. at 16 (proving three new mandatory exceptions concerning the text and data mining (art. 3 and 4), the teaching activities (art. 5), and the preservation of cultural heritage (art. 6)).

183. *Id*.

184. *Id*. at 19.

It especially concerns the content generated by users, such as "reaction GIF" and "meme" images, fan fiction, "libdubs,"[185] and "supercuts."[186] In some member states, this user-generated content could be covered by a wide-reaching quotation exception, while others countries are more restrictive.[187] The interpretation has to be harmonized, but the terms of the directive are vague and allow some flexibilities in favor of the member states and, finally, the platforms.

To prevent misuse or limitation in the exercise of exceptions and limitations to copyright, article 17(9) enacts that:

> online content-sharing service providers have to put in place an effective and expeditious complaint and redress mechanism that is available to users of their services in the event of disputes over removal of works uploaded by them. Where right holders request to have access to their specific works removed, they shall duly justify the reasons for their requests. Complaints submitted under this mechanism shall be processed without undue delay, and decisions to remove uploaded content shall be subject to human review. Member states shall also ensure that out-of-court redress mechanisms are available for the settlement of disputes. Such mechanisms shall enable disputes to be settled impartially and shall not deprive the user of the legal protection afforded by national law, without prejudice to the rights of users to have recourse to efficient judicial remedies. In particular, Member states shall ensure that users have access to a court or another relevant judicial authority to assert the use of an exception or limitation to copyright and related rights.[188]

In brief, redress mechanisms guarantee a right to contest the content removal and obtain a non-judicial review based on a human decision, as well as a judicial redress, in compliance with article 47 of the EU Charter of fundamental rights on the right to an effective remedy and to a fair trial.[189]

Finally, article 17(10) states that:

> the Commission, in cooperation with the Member states, shall organize stakeholder dialogues to discuss best practices for cooperation between online content-sharing service providers and right holders. When discussing best practices, special account shall be taken, among other things, of the need to balance fundamental rights and of the use of exceptions and limitations. For the purpose of the stakeholder dialogues, users' organizations shall have access to adequate information from online content-sharing service providers on the functioning of their practices.[190]

185. Digital Single Market, *supra* note 69, at art. 17; *see also* Julia Reda, *What the EU Parliament May Add to Copyright Reform Plans*, JULIA REDA, https://juliareda.eu/eu-copyright-reform/parliament-additions (last visited Oct. 21, 2020) ("Recording yourself acting as if performing a pop song.").

186. *Id* (making compilations of movie scenes sharing a particular characteristic).

187. Digital Single Market, *supra* note 69, at art. 17.

188. *Id*.

189. EU Charter of fundamental rights 2012/C 326/02, 2012 O.J. 326, 405 (article 47 on right to an effective remedy and to a fair trial).

190. Digital Single Market, *supra* note 69, at art. 17.

This last provision could be interpreted as a way to guarantee a transparency of the modus operandi and maybe the rules of automated decision-making systems used but there is not really a strong requirement here and the service providers are not expected to respect a transparency commitment.

This new copyright reform is the most achieved EU law regarding content moderation issues.[191] This new, complex, and specific liability regime obliges the most important web actors of using algorithmic decision-making systems to remove the copyrighted content and interpret it, in the hope to achieve a perfect enforcement. Unfortunately, perfectly automated enforcement is not possible here.[192] Nevertheless, the European lawmaker continues to go forward with these types of requirements.[193] The proposal of regulation on the moderation of terrorist content is going in the same way.

### *2. Proactive Measures on Terrorism*

Terrorist attacks on EU territory have demonstrated how terrorists misuse the internet to groom and recruit supporters, prepare and facilitate the terrorist activity, and glorify in their atrocities and urge others to follow suit and instill fear in the general public. Given the fact that terrorist content online continues to be easily accessible, the European Union took some measures.[194]

The terrorism directive (EU) 2017/541 is the first text enacted which obliges the member states of taking necessary measures for ensuring the prompt removal of online content inciting to commit terrorist acts hosted in their territory.[195] The member states shall also endeavor to obtain the removal of such content hosted outside their territory (art. 21(1)).[196] The purpose is the removal of illegal content, and the member states could adopt a notice and take-down procedure, as well as proactive measures.[197] In the context of the EU Internet Forum, platforms remove voluntarily terrorist content from referrals sent by the Europol Internet Referral Unit (IRU).[198] Furthermore, the member states shall take measures to block access to web pages containing or disseminating terrorist content, when the content removal is not feasible (art. 21(2)).[199] This directive creates a priority order, and the internet access providers act only and potentially in the event of a hosting service providers' inaction towards the removal of the illegal content.[200] The territorial scope of the directive is limited to the territory of the European Union. However, to protect Europeans, there is an expectation that the internet access providers block illegal content and ensure the directive's efficacy.[201] This directive provides safeguards to limit the over-removal thanks

191. *Id*.
192. *Id*.
193. *Id*.
194. Combatting Terrorism Framework, *supra* note 67.
195. *Id*.
196. *Id*.
197. *Id*.
198. *Id*.
199. *Id*.
200. *Id*.
201. *Id*.

to a transparent procedure, and principles of necessity and proportionality.[202] Rights in favor of the users have also been recognized: a right to be informed of the reason for the removal and a right to obtain judicial redress.[203]

Moreover, the European Commission proposed for a regulation on 12 September 2018 on preventing the dissemination of terrorist content online.[204] After the Communication on Tackling illegal content online, the European Commission pursues its goal to moderate the content online, especially regarding terrorism, taking into account the freedom of expression and information in an open and democratic society.[205] This regulation lays down a set of measures to be put in place by member states to identify terrorist content, to enable its swift removal by hosting service providers, and to facilitate cooperation with the competent authorities in other member states, hosting service providers, and where appropriate, relevant Union bodies.[206] The proposal directly imposes duties of care on hosting service providers (but excludes access providers) to prevent the dissemination of terrorist content through their services and ensure, where necessary, its swift removal (art. 1(1)).[207] Given its preventive nature, it covers not only material inciting terrorism but also material for recruitment or training purposes, reflecting other offences related to terrorist activities, also covered by directive 2017/541.[208] Its territorial scope applies to hosting service providers offering services in the Union, irrespective of their place of main establishment (art. 1(2)).[209]

This proposal also introduces a number of necessary safeguards designed to ensure full respect for fundamental rights, in addition to judicial redress possibilities guaranteed by the right to an effective remedy as enshrined in article 19 TEU and Article 47 of the Charter of fundamental rights of the EU.[210] However, Articles 3 to 7 of the proposal state several measures for preventing the dissemination of content terrorist online, while the definition of the "content terrorist" is particularly broad: the fact to incite,[211] advocate, or glorify the commission of terrorist offences is included,[212] as well as to promote the activities of a terrorist group.[213]

Consequently, it may be very difficult to proactively appreciate this kind of content and its context, especially by automated decision-making systems. How to distinguish the terrorist propaganda and the work of civil rights

202. *Id*.
203. *Id*.
204. *PhotoDNA*, *supra* note 88.
205. Tackling Illegal Content, *supra* note 64.
206. Terrorist Content Online, *supra* note 67.
207. Combatting Terrorism Framework, *supra* note 76 at art. 1.
208. *See id* (defining the 'terrorist content' as one or more of the following information: (a) inciting or advocating, including glorifying, the commission of terrorist offences, thereby causing a danger that such acts be committed; (b) encouraging the contribution to terrorist offences; (c) promoting the activities of a terrorist group within the meaning of Article 2(3) of Directive (EU) 2017/541; (d) instructing on methods or techniques for the purpose of committing terrorist offences).
209. *Id*.
210. *Id*.
211. *Id*. at 13.
212. *Id*.
213. *Id*.

association, advocacy groups, or journalists who struggle and denounce it? How preserve this second type of content to an over-removal? Such legal provisions create a high risk for freedom of speech.

Article 3(1) states that:

> hosting service providers (…) are expected to act in a diligent, proportionate and non-discriminatory manner, and with due regard to the fundamental rights of the users, taking into account the fundamental importance of the freedom of expression and information in an open and democratic society.[214]

Hosting service providers should strike a fair balance between public security needs, and the affected interests and fundamental rights including, in particular, the freedom of expression and information, freedom to conduct a business, and protection of personal data and privacy.

However, the balance of rights is not an easy work of reasoning and is not by itself compatible with an automatization of the decision-making. A casuistic interpretation doesn't fit with solely automated decision-making. No pattern can be found because each case is different and deserves human reasoning to appreciate the context. Automated processing could be useful for supporting the decision and not for making it and a human monitoring at least necessary.[215] But by doing that, with or without human oversight, the lawmaker delegates the power of judgement to the hosting service providers.[216] Even more, these providers have to take proactive measures and act before the dissemination of the content. They should anticipate a risk that could maybe not occur.[217] That is problematic for the freedom of expression. How could these service providers know an intention or an effect in advance? Does it mean they have to monitor the content a priori? Is it compliant with Article 15 of the E-commerce Directive which excludes general content monitoring? The European Commission indicates that the proposal is consistent with the E-commerce Directive.[218] Notably, any measures taken by the hosting service provider, including any proactive measures, should not, in principle, lead to the imposition of a general obligation to monitor, as defined in Article 15(1) of E-commerce Directive.[219] However, in its explanatory memorandum, the European Commission considers that this balance of rights is led by the particularly grave risks associated with the dissemination of terrorist content.[220] Consequently, the decisions under this regulation may exceptionally derogate from this principle under an EU framework. Should we understand this provision as an exception to article 15 of the E-commerce Directive?

Article 3(2) adds that:

---

214. *Id*. at 24.
215. Omer Tene, *A New Harm Matrix for Cybersecurity Surveillance*, 12 COLO. TECH. L.J. 391, 401–02 (2014).
216. Bloch-Wehba, *supra* note 75, at 24.
217. *Id*.
218. *Id*. at 16.
219. *Id*.
220. *Id*.

> hosting service providers shall include in their terms and conditions and apply, provisions to prevent the dissemination of terrorist content.[221]

This contractual requirement is vague and doesn't provide any platforms' obligation to explain neither a general explanation based on their content moderation policy nor an individual one.[222]

Besides, Article 4 provides a stringent notice and take-down procedure and states that "competent authority shall have the power to issue a decision requiring the hosting service provider to remove terrorist content or disable access to it."[223] Moreover, "[h]osting service providers shall remove terrorist content or disable access to it within one hour from receipt of the removal order."[224] This obligation is very difficult to satisfy.

Article 6(1) states that:

> Hosting service providers shall, where appropriate, take proactive measures to protect their services against the dissemination of terrorist content . . . [and protect] the fundamental rights of the users, . . . [especially the] freedom of expression and information in an open and democratic society.[225]

The specific proactive measures it requires include "automated tools" and, "preventing the re-upload of content which has previously been removed," as well as to, "detect[], identify[], and expeditiously remov[e]" the terrorist content.[226] Moreover, hosting service providers shall, "submit a report, within three months after receipt of the request and thereafter at least on an annual basis, on the[se] specific proactive measures it has taken."[227] "The reports shall include all relevant information allowing the competent authority . . . to assess whether the proactive measures are effective and proportionate, including to evaluate the functioning of any automated tools used as well as the human oversight and verification mechanisms employed."[228]

Despite this platform's system of reporting, the effectiveness of this accountability will depend on the severity of the sanctions, as well as the kind of control. It would be difficult for the "competent authority" to oversee automated tools of content moderation, while the removal itself and the nature of content removed are not traceable.[229]

Article 9(1)–(2) states:

> Where hosting service providers use automated tools . . . , they shall provide effective and appropriate safeguards to ensure that decisions taken . . . are accurate and well-founded. Safeguards shall consist, in particular, of human oversight and verifications where appropriate

221. *Id*. at 24.
222. *Id*.
223. Terrorist Content Online, *supra* note 67, at art. 4(1).
224. *Id.* at art. 4(2).
225. *Id.* at art. 6(1).
226. *Id.* at art. 6(2).
227. *Id.*
228. *Id; see also id.* at art. 17(1)(c) (defining competent authority).
229. *See id.* at art. 6(2).

> 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 terrorist content.[230]

Such provisions are interesting but only if the human controller has significant power to decide and if her oversight concerns most of the content. The problem here is the cost of human intervention.[231]

Article 10(1) also provides interesting complaint mechanisms:

> Hosting service providers shall establish effective and accessible mechanisms allowing content providers whose content has been removed . . . as a result of proactive measures . . . to submit a complaint requesting reinstatement of the content.[232]

A mechanism of redress is necessary but has to be ensured by an independent body to provide a true right of appeal.

Article 11(1) obliges the hosting service providers to give information to individual content providers where they removed terrorist content.[233] Most interestingly, Article 11(2) states that, "[u]pon request of the content provider, the hosting service provider shall inform . . . about the reasons for the removal . . . and possibilities to contest the decision."[234] However, these obligations "shall not apply where the competent authority decides that there should be no disclosure for reasons of public security, such as the prevention, investigation, detection[,] and prosecution of terrorist offences, for as long as necessary, but not exceeding [four] weeks from that decision."[235] This exception to the obligation of providing individual information could be often invoked.

In brief, this proposal of regulation on preventing the dissemination of terrorist content online creates a new obligation to take proactive measures, which could be dangerous for the respect of freedom of speech. However, this text also provides interesting safeguards and rules of accountability, such as the right to human oversight, information, remedies, and reinstatement of the content removed.[236] More rights have to be given to the content providers, but attention has to be paid to the effectiveness of these safeguards in practice.

### *C. Proactive Measures and Platform Liability Regime*

According to the Commission, taking such voluntary, proactive measures does not automatically lead to the online platform losing the benefit of the liability exemption provided for in Article 14 of the E-Commerce Directive.[237] However, the European Commission refers to a weak legal argument to justify proactive measures. According to Recital 40, which is non-binding, "this Directive should constitute the appropriate basis for the development of rapid

230. *Id.* at art. 9(1)–(2).
231. Keller, *supra* note 61, at 5.
232. Terrorist Content Online, *supra* note 67, at art. 10(1).
233. *Id.* at art. 11(1).
234. *Id.* at art. 11(2).
235. *Id.* at art. 11(3) (brackets in original).
236. *Id.* at art. 9.
237. Tackling Illegal Content, *supra* note 64.

and reliable procedures for removing and disabling access to illegal information."[238] Article 15 prohibits Member States from "impos[ing] a general obligation on providers . . . to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity."[239] At the same time, Recital 47 of the Directive recalls that this only concerns "monitoring obligations of a general nature" and does not automatically cover, "monitoring obligations in a specific case and, in particular, does not affect orders by national authorities in accordance with national legislation."[240] That is the only legal argument to encourage proactive measures.

By adopting all these proactive measures, the European lawmakers claim to provide clarification to platforms regarding their liability when they take proactive steps to detect, remove, or disable access to illegal content, following the framework of "Good Samaritan" actions. But with others,[241] I argue that voluntary monitoring could generate awareness and knowledge of facts or circumstances from which the illegal activity or information is apparent and known by the platform. In this case, the platform has to act promptly and remove the illegal content.[242] Consequently, platforms could lose the benefit of the liability exemption regime.[243] Therefore, concerns related to losing the benefit of the liability exemption should neither deter nor preclude the application of the effective proactive voluntary measures that this Communication seeks to encourage.[244]

Besides, copyright reform is more ambiguous because online content sharing service providers need licenses to communicate to the public the copyrighted content they store.[245] They are liable if they do not obtain authorization from copyright holders.[246] Consequently, the new copyright directive provides a new regime of liability outside the scope of the E-Commerce Directive and besides the exemption regime of liability. European Union lawmakers proclaim these measures are providing clarifications on liability to platforms, but this sectorial approach is particularly confusing.[247] Even more, the new obligation of "notice and stay down," as well as the set of three different rules of liability related to the size and audience of the platforms, introduce unnecessary normative complexity.[248] If the rule of law is misleading, the platforms could play with it. In a context where opaque algorithmic tools are used to apply the rule of law, the lack of transparency will be reinforced if the norm is unclear. The transparency of the application of the rule of law by algorithms first supposes a clear rule to respect.[249]

238. Electronic Commerce Directive, *supra* note 13, at recital 40.
239. *Id.* at art. 15(1).
240. *Id.* at recital 47.
241. Kuczerawy, *supra* note 106.
242. Tackling Illegal Content, *supra* note 64, at 15.
243. Kuczerawy, *supra* note 106.
244. Tackling Illegal Content, *supra* note 64, at 12.
245. Copyright Harmonisation, *supra* note 69, at recitals 61, 64.
246. *Id.* at recital 66.
247. *Id.* at recital 3.
248. *Id.* at art. 17(4) – (6).
249. Tackling Illegal Content, *supra* note 64, at 12.

Furthermore, the European Commission uses the term "Good Samaritan" action, which is a reference to Section 230 of the 1996 Communications Decency Act (CDA) from the United States.[250] But this is irrelevant in the European context.[251] Such provisions have allowed the internet to thrive on user-generated content without holding platforms and ISPs responsible for whatever those users might create. Section 230(c)(2) of the CDA protects intermediaries when they take voluntary measures to restrict access or availability of certain content in the "Good Samaritan" spirit.[252] On the other hand, the hosting service providers are also protected when they fail to notice such content and do not take any action at all.[253] This provision provides moderators with immunity, "both for the content they moderate and the content they miss."[254] Platforms are protected against under-removal and over-removal.[255] By giving this assurance, the United States Congress effectively encouraged intermediaries to implement proactive measures that are disconnected from any rule of liability.[256] Conversely, the E-Commerce Directive states a liability rule if internet service providers do not act "expeditiously to remove" content flagged as illegal.[257] Consequently, neither proactive measures nor liability rules of the E-Commerce Directive are similar to the "Good Samaritan" clause of Section 230 of the CDA.[258]

## III. AUTOMATED DECISION-MAKING SYSTEMS: ILLUSION OF CONTROL AND (IM)PERFECT ENFORCEMENT

European lawmakers purport to control the removal of illegal content in different fields of law. However, the first difficulty arises when deciding what is illegal. As the European Commission points out, the legal framework does not define what constitutes "illegal" content.[259] Indeed, specific legislation at both the EU level and at national levels determine what is illegal,[260] creating a

250. 47 U.S.C. § 230(c).

251. Kuczerawy, *supra* note 106.

252. *See* 47 U.S.C. § 230(c)(2) (explaining that online service providers shall not be held liable based on "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected").

253. *Id.* The "good faith" provision was introduced to overrule *Stratton Oakmont, Inc. v. Prodigy Servs. Co.*, No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995). An online service provider was found liable for defamatory content by third parties because service provider removed the content but failed to do so entirely. Kuczerawy, *supra* note 106.

254. Grimmelmann, *supra* note 63, at 103.

255. *Id.*

256. Kuczerawy, *supra* note 106.

257. Electronic Commerce Directive, *supra* note 13, at art. 14.

258. *Id.*

259. Tackling Illegal Content, *supra* note 64, at 15.

260. Each Member State enacts its own regulations pertaining to removing online and offline illegal content. For instance, German Law reinforced the sanctions to require the platforms to remove the illegal content in the case of hate speech, including defamatory "fake news." Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], July 12, 2017, *translation at* https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2. On June 30, 2017, German Parliament approved the Network Enforcement Act, commonly known as NetzDG, a bill criminalizing hate speech on social media sites, such as Facebook, Instagram, Twitter, and YouTube. *Id.* The law requires

fragmented approach within and outside the EU.[261] Moreover, a certain amount of illegal content is related to opinions and the principles of freedom of opinion and speech and is protected in Articles 10 and 11 of the EU Charter of Fundamental Rights.[262] Consequently, the fight against illegal content online must be carried out with proper and robust safeguards to ensure protection of the different fundamental rights at stake.[263] Platforms manage the removal of illegal content automatically, which makes it particularly challenging to verify that the law is being respected. The automated decision-making systems are opaque[264] and do not prevent an effect of under or over-removal.[265] Many scholars have shown that the main problem here is the over-removal chilling effect (2).[266] Finally, automated decision-making is imperfect. In some circumstances, automatic removal is neither possible, nor desirable, nor relevant. That is, for instance, the case when exceptions are applied with contextual consideration.[267] Nevertheless, even in these cases, platforms continue to apply their automatic tools and methods (3).[268]

### *A. Opaque Algorithmic Decision-Making*

In the EU, the courts and national authorities have the ability to prosecute crimes and impose criminal sanctions relating to the illegality of any given activity or information online.[269] However, the European Commission considers that online platforms are entitled to prevent their infrastructure and business from being used to commit crimes.[270] Consequently, they are responsible for protecting their users and preventing illegal content from appearing on their

large social media platforms to promptly remove "illegal content," as defined in 22 provisions of the criminal code, ranging widely from the insult of public office to actual threats of violence. Among criminalizing hate speech, the law states that social networking sites may be fined up to €50 million (US$56 million) if they persistently fail to remove illegal content within a week. French Law also provides specific rules concerning child-pornography, terrorism, and, recently, fake news. *See, e.g.*, Loi 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique [Law 2004-575 of June 21, 2004 for Confidence in the Digital Economy], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], June 22, 2004, art. 6.I.7; Loi 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information [Law 2018-1202 of December 22, 2018 on the Fight Against the Manipulation of Information], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], Dec. 23, 2018. These laws require cooperation of web giants with respect to the information distributed on their platforms.

261. See Citron, *supra* note 38; Annemarie Bridy, *Remediating Social Media: A Layer-Conscious Approach*, 24 B.U. J. SCI. & TECH. L. 193, 215–16 (2018).

262. Charter of Fundamental Rights of the European Union, arts. 10–11, 2012 O.J. (C 326) 391, 397–98 ("Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers . . . . The freedom and pluralism of the media shall be respected.").

263. Tackling Illegal Content, *supra* note 64, at 3.

264. Bar-Ziv & Elkin-Koren, *supra* note 47, at 343.

265. *Id*.

266. Keller, *supra* note 61, at 5; Thomas E. Kadri & Kate Klonick, *Facebook v. Sullivan: Building Constitutional Law for Online Speech*, 93 S. CAL. L. REV. 37 (2019); Danielle Keats Citron, *What to Do About the Emerging Threat of Censorship Creep on the Internet*, 828 CATO INST. POL'Y ANALYSIS 1 (2017); Bloch-Wehba, *supra* note 75, at 27.

267. Keller, *supra* note 61; Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019).

268. *See* Keller, *supra* note 61 (explaining that major platforms rely increasingly on automation rather than human review).

269. Tackling Illegal Content, *supra* note 64, at 3.

270. *Id.*

platforms.[271] If a human-decision making system applies at Facebook for instance,[272] automatic tools and filters are also used to identify quickly[273] any potentially infringed content. Even more, the European Commission encourages the use of automated tools, which would seem to guarantee perfect enforcement.[274] In regards to the volume of material intermediated by online platforms, as well as the technological progress made in information processing and machine intelligence, automatic detection, and filtering technologies are becoming essential tools in the fight against illegal online content.[275] Many large platforms are now making use of some form of matching algorithms based on a range of technologies, from simple metadata filtering to hashing and fingerprinting content.[276]

Nevertheless, while swift decisions concerning the removal of illegal content are important, there is also a need to apply adequate safeguards. If the ways in which the Big Tech companies design their services are unchecked, it could easily reinforced the conditions of inequality.[277] This also requires a balance of roles between public and private bodies.[278] A software system can be effective and deliver relevant results without guaranteeing transparency, even for cases which are not complex.[279] Consequently, software can be unpredictable to those it regulates.[280] Moreover, an algorithm is not neutral[281] and can perpetuate existing stereotypes and social segregation.[282] There is a bias problem when a computer system systematically and unfairly discriminates against groups of individuals in favor of others, based on social or ethical criteria.[283] But even more problematic, the software is asymmetrical when the user only sees the results of the software's individual decisions, but has no access to accurate information on the input set that determined a particular output.[284] Moreover, users are confined to dealing with software on a case-by-case basis concerning their individuality.[285] It may be difficult to detect systemic discrimination against a group, especially where unpredictable software is

271. *Id*.

272. Simon Van Zuylen-Wood, *"Men Are Scum": Inside Facebook's War on Hate Speech*, VANITY FAIR (Feb. 25, 2019), https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech.

273. Common, op. cit., *supra* note 87.

274. Tackling Illegal Content, *supra* note 64, at 3.

275. *Id*.

276. *Id*.

277. Olivier Sylvain, *Recovering Tech's Humanity*, 119 COLUM. L. REV. 252, 261 (2019); Olivier Sylvain, *Intermediary Design Duties*, 50 CONN. L. REV. 203, 232 (2018).

278. Tackling Illegal Content, *supra* note 64, at 3.

279. James Grimmelmann, *Regulation by Software*, 114 YALE L.J. 1719 (2005).

280. *Id*.

281. Common, op. cit., *supra* note 87 ("Both human and algorithmic moderation are not neutral and hold biases because the prejudices and assumptions that organically occur in humans are held by both moderators and the programmers who create algorithms.").

282. Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1055 (2017).

283. Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANSACTIONS ON INFO. SYS. 330, 331 (1996).

284. *Id*.

285. *Id*. ("If the system is complex, and most are, biases can remain hidden in the code, difficult to pinpoint or explicate, and not necessarily disclosed to users or their clients.").

involved.[286] Moreover, bias may be introduced into machine learning processes at various stages, including during algorithm design,[287] which is the case for platforms which implement automated processes to remove content. But the users have no information regarding the design or instructions given to the machine by the platforms, and it could easily be a source of biases, errors and over-removal.[288]

More powers for platforms have to come with more responsibility. At this time, the EU and national authorities are not able to control decisions made by platforms via algorithms, nor are they able to hold the dominant social media platforms accountable.[289] Furthermore, automated enforcement reinforces the risk of biases and over-removal that cannot be reviewed. In order to combat opacity, the most basic regulatory strategy is an explanation, in which the regulator says what it is she is doing as she makes a decision.[290] But two problems occur: first, the relevance of such strategy will depend on one's ability to verify the explanation's accuracy; second, software itself does not always give a clear explanation, even if the programmer has tried to code the means for such explanation.[291] These two limits are particularly visible in the context of online platforms. For instance, YouTube's ContentID system is a moderation technique that is difficult to oversee, and it's difficult to know if the matching algorithms are too aggressive or not aggressive enough.[292] Moreover, there is a language challenge, and it is not guaranteed that communication from the machine language to natural human language will be comprehensible.[293] Machines have difficulty replying appropriately.[294] Moreover, unlike human responses, it is difficult to individualize a machine-generated answer. A human support system could be added to better understand the machine decision-making process, but only if the programmer is able to understand and justify it in the first place.

The first *Compliance Reports* of the *Code of Practice against Disinformation*, published by the European Commission on January 29, 2019,[295] confirms this opacity problem. The European Commission Security Chief, Julian King, criticized the "patchy, opaque and self-selecting" reporting provided by Facebook and other tech giants following their bids to comply with the EU's *Code of Practice Against Disinformation*.[296] But the *Code of Practice*

286. *Id.*

287. WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES (Harv. U. Press ed., 2018).

288. Common, op. cit., *supra* note 87; Chander, *supra* note 282; Friedman & Nissenbaum, *supra* note 283.

289. Keller, *supra* note 61.

290. Grimmelmann, *supra* note 279.

291. *Id.*

292. Grimmelmann, *supra* note 63.

293. Grimmelmann, *supra* note 279.

294. *Id.*

295. *First Results of the EU Code of Practice Against Disinformation*, EUROPEAN COMMISSION, https://ec.europa.eu/digital-single-market/en/news/first-results-eu-code-practice-against-disinformation (last visited Oct. 21, 2020).

296. Samuel Stolton, *EU Commission Hits Out at Facebook's Disinformation Report*, EURACTIV (Jan. 30, 2019), https://www.euractiv.com/section/data-protection/news/facebook-singled-out-in-eu-disinformation-report.

is merely a voluntary process which encourages compliance, and it is both vague and non-binding.[297] Even when the EC tries to specifically regulate the removal of illegal content and the expected behavior of the platforms, the lawmaker can neither monitor automated systems nor challenge the decisions made.[298] Without tools and data access to oversee removal decisions, pursuing safeguards to protect fundamental rights is a utopian goal. The European Commission recommends the provision of effective and appropriate safeguards to ensure that decisions taken concerning the removal of content are accurate and well-founded.[299] 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.

However, several limits can be pointed out. First, the platform monitors recognize when human verification is "appropriate" and when it is relevant to provide a detailed assessment of the context, as technical filters cannot assess context.[300] In many cases, it is impossible to appreciate the legality of the content without consideration of the context.[301] Second, human intervention provides no guarantees, especially when the person has no power or understanding of the decision made by the machine.[302] Indeed, "seeing is not knowing"[303] and access doesn't necessarily mean a true power of control. The safeguards recommended by the European Commission are not strong enough and, even worse, non-binding.

Consequently, the *Compliance Reports* of the *Code of Practice against Disinformation* have a limited impact, because of their focus sectorial and their lack of accountability.[304]

### *B. Over-Removal Chilling Effect*

The trend is in favor of the removal of legal content, so-called over-removal.[305] In the US, the rise of the social demand for removing illegal content generates a routine social media practice toward an automated over-removal for the last twenty years.[306] In Europe, European and national lawmakers are developing additional ways to monitor the online content and strengthen law enforcement.[307] In this context, the platforms are thus empowered by automated decision-making systems, especially when deciding disputes regarding harmful

297. Grimmelmann, *supra* note 279.
298. *Id.*
299. Tackling Illegal Content, *supra* note 64, at 5.
300. Daphne Keller, *Inception Impact Assessment: Measures to Further Improve the Effectiveness of the Fight Against Illegal Content Online* (2018), https://papers.ssrn.com/abstract=3262950.
301. *Id.*
302. *Id.*
303. Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973 (2018).
304. *Id.*
305. Keller, *supra* note 61.
306. *Id.*
307. Tackling Illegal Content, *supra* note 64, at 3.

speech,[308] while many rules are limited by exceptions or specific material scope which is not in line with the automation of procedures. Moreover, considering the notice and take-down procedure, as well as the risk of liability, over-removal was encouraged to mitigate the risk.[309] Platform behavior is partially the result of these national intermediary liability laws.[310] Additionally, the right to be forgotten, recognized in the *Google Spain* case by the European Court of Justice,[311] based on Data Protection Directive 95/46/EC[312] and now enacted by the General Data Protection Regulation,[313] also encourages the removal of inadequate, irrelevant or excessive content in light of the time that has elapsed.[314] This rule does not concern illegal content but provides another way to moderate the content.[315] All these circumstances create a risk of over-removal where the digital single market legal framework is complex.

European guidelines and principles seek to address specific concerns in relation to illegal content.[316] But very few considerations and solutions have been proposed concerning the over-removal chilling effect, while it impacts freedom of expression and media pluralism.

First, this risk of over-removal is related to the territoriality of law. Platforms with international presence must comply not only with US laws, but also with those of other countries.[317] They can offer different versions of their product in different countries, which is both inconvenient and costly, or they can settle for an acceptable amalgam of national laws and comply with these as a matter of "voluntary" policy.[318] Many of the laws which have recently been adopted to remove certain types of content, for instance to respect the "right to be forgotten,"[319] would not be enforceable in the United States.[320] But *de facto*, the European notion of free speech leads the debate and the platforms' policy,[321] despite not being as strong as the US' notion of free speech.[322] Content covered by free speech in the First Amendment in the US Constitution can be illegal in Europe, such as denial of the Holocaust.[323] Even within the EU, most topics fall

308. Thomas E. Kadri & Kate Klonick, *Facebook v. Sullivan: Building Constitutional Law for Online Speech*, 93 S. CAL. L. REV. 37, 46 (2019).

309. *Id.*

310. *Id.*

311. C-131/12, Google Spain v. APED, 2014 ECLI:EU:C:2014:317, ¶ 91 (May 13, 2014) [hereinafter *Google Spain*].

312. GDPR, *supra* note 68.

313. *Id.* at art. 17.

314. *Google Spain*, *supra* note 311, at ¶ 7.

315. *Id.*

316. Zuylen-Wood, *supra* note 272.

317. Keller, *supra* note 61, at 8.

318. *Id.*

319. *Google Spain*, *supra* note 311, at ¶ 91.

320. Securing the Protection of our Enduring and Established Constitutional Heritage Act, Pub. L. 111-223, 124 Stat. 2380 (2010) (amending Title 28 U.S.C. so as to prohibit recognition and enforcement of foreign defamation judgments in United States Courts where those judgments undermine the first amendment to the Constitution of the United States, and to provide a cause of action for declaratory judgment relief against a party who has brought a successful foreign defamation action whose judgment undermines the first amendment).

321. Citron, *supra* note 266.

322. Balkin, *supra* note 7, at 107.

323. *See* Gertz v. Robert Welch, Inc., 418 U.S. 323, 339 (1974) (ruling "there is no such thing as a false idea" being subject to compensation in a defamation action).

under the jurisdiction of national laws. There is no single and coherent approach to remove illegal content, as all depends on national laws, nature of the content and type of online platform.[324] But online platforms operate worldwide and usually apply the same rules everywhere. Consequently, content is frequently over-removed in consideration of the stricter national law, sometimes going against the US interpretation of free speech. Besides, once platforms achieve the technical ability to do things like filter user speech, governments around the world will want to use these technologies too.[325] Governments outsource certain decision-making functions and attempt to extend domestic laws and norms beyond territorial limits.[326] Moreover, some courts have ruled that they have jurisdiction to enforce compliance on a global scale, leading to the deletion of information that is legal in one jurisdiction and illegal in the next.[327]

Second, social media platforms can decide to remove content without legal basis, based on their own private rules, which are more or less clearly disclosed in their terms and conditions.[328] While the question of whether certain content is legal or illegal is governed by European and national laws, online platforms' terms of service can consider specific types of content as being undesirable, regarding for instance moral considerations.[329] Moreover, the advertisers have the power to drive platform content policies as the "adpocalypse" proved it when YouTube changed its policies after the revealing of hate speech and extremist content that could also generate over-removal.[330] Such private self-regulation is not legitimate *per se*, and there is no reason to admit it, but it is performed automatically.[331]

Third, another problem occurs when abusive requests appear, including efforts to silence commercial competitors or ideological rivals.[332] The European Commission recommends taking effective and appropriate measures to prevent the submission of notices and counter-notices in bad faith and other forms of abusive online behavior.[333] But this framework is not robust. The rules are not detailed, and the platforms are free to decide which measures to take. There is no guarantee here that the over-removal is avoid with certainty.

Fourthly, algorithmic governance guarantees direct enforcement and reinforces the over-removal risk effect. Moreover, due to potential bias in the data sets, algorithmic governance itself can introduce biases and discrimination.[334]

---

324. Digital Single Market, *supra* note 69.
325. Keller, *supra* note 61, at 8.
326. Bloch-Wehba, *supra* note 75, at 27.
327. *Id.*
328. *See* Tribunal de grande instance [TGI] [ordinary court of original jurisdiction] Paris, 4e chamber 2e section, March 15, 2018, n° 12/12401 (Fr.) (ruling in favor towards a painting of Gustave Courbet showing nudity).
329. Keller, *supra* note 61, at 4.
330. *Id.* at 1.
331. *Id.* at 4.
332. *Id.*
333. Tackling Illegal Content, *supra* note 64.
334. Keller, *supra* note 61, at 6.

In brief, there are many reasons why over-removal is *de facto* encouraged. Even when it is predictable, there is no way that lawmakers guarantee its prevention or reduction. The platforms thus maintain total control of the content and action that they host.

### *C. Non-Relevant Algorithmic Decision-Making Process*

The European Commission considers that algorithmic governance achieves a "perfect law enforcement," thanks to the platforms' cooperation.[335] However, in many circumstances, automatization is not the best way to comply with the law, and a human interpretation is often necessary to appreciate the context. For instance, in copyright law, the reproduction or representation of work without the rights holder's permission can be covered by an exception, such as fair use.[336] In these cases, the law cannot be perfectly enforced by automated copyright enforcement, and human intervention is required. However, platforms such as YouTube use an automated process to remove copyrighted content and leave little possibility of appealing for an exception.[337] Moreover, the European copyright reform enacted in April 2019 reinforces this kind of automated enforcement system.[338] Consequently, the platforms systematically use automated processing with few considerations for the niceties of the rule of law involved. Moreover, neither platforms nor lawmakers can throw a switch and halt the flow of particular kinds of speech or content. Artificial intelligence and technical filters cannot do it either.[339]

The new directive (EU) 2019/790 on copyright and related rights in the Digital Single Market tries to provide guarantees concerning the respect of exceptions and limitations to copyright.[340] It states that cooperation between online content service providers and rights holders shall not lead to preventing the availability of non-infringing works, including those covered by an exception or limitation to copyright.[341] But an automatic removal system, such as Content ID used by YouTube, is not able to interpret the content or context to decide whether or not there is a violation of the copyrighted material.[342] Consequently, any use of such content will generate automated removal. Then, the users could be in the situation of not benefiting from their exception and have to exercise their right to contest.[343] Redress is not guaranteed, and the period of removal can be harmful and prejudicial to freedom of expression. Over-removal can hurt businesses as much as it hurts individual users.[344]

---

335. *Id*.
336. Burk, *supra* note 267, at 283.
337. Keller, *supra* note 61, at 5–6.
338. Digital Single Market, *supra* note 69.
339. Keller, *supra* note 61, at 3.
340. Digital Single Market, *supra* note 69, at art. 17 §7.
341. *Id.*
342. *See* Jao Carrasqueira, *YouTube to rely more heavily on automatic content moderation due to Covid-19*, NEOWIN (Mar. 16, 2020), https://www.neowin.net/news/youtube-to-rely-more-heavily-on-automatic-content-moderation-due-to-covid-19.
343. *Id.*
344. Keller, *supra* note 61, at 1.

Another example can be found in the EC recommendation,[345] which provides exceptions to the obligation of the hosting service provider to inform the content provider of any removal of content. One of these exceptions applies when it is "manifest" that the concerned content is illegal and relates to serious criminal offenses involving a threat to life or safety of persons.[346] Consequently, the hosting service providers appreciate the situation. But in fact, such appreciation is made by an algorithmic process programmed to remove content, as is the criteria of appreciation of what is "manifest" and a "threat to life."[347] In this context, an automated decision is neither possible nor desirable. Furthermore, another issue occurs when the removal made automatically was challenged and finally declared unjustified. In this case, the decision has to be reversed, but very often, it is not materially possible to reverse a removal.[348] Moreover, the reversal might have an effect on other content that is less traceable and verifiable.[349]

Consequently, these examples show that content removal is a task which, in many circumstances should not be automated, because it directly depends on an appreciation of the context and the rule of law. The role of the automated function is *de facto* judging, which is not compatible with a necessary case by case approach. Not only is such function delegated to a private actor, but more precisely, to a machine, which has no casuistic way of functioning and few processes that can be held accountable.[350] If the European Commission seems aware of these problems, its recommendations are far too weak.

## IV. RECOMMENDATIONS

To address these issues, a set of mandatory measures have to be built for online platforms in cooperation with national authorities, EU Member states, and other relevant stakeholders.[351] In its recommendations, the European Commission does not make proposals concerning algorithmic transparency towards platforms.[352] Two kinds of measures may improve monitoring of how platforms moderate content: the platforms should disclose their rules of moderation, their process of notification, as well as their results. "Auditability" should also be mandatory.[353] I propose solutions to improve the algorithmic accountability and transparency of automated decision-making, as well as "auditability" (1). Furthermore, though the European Commission has made some recommendations concerning the judicial redress in the event of over-removal, such proposals are not mandatory and are dependent on the national judicial systems of the member states.[354] To guarantee a better redress in case

345. Measures to Effectively Tackle Illegal Content, *supra* note 77, at 1.
346. *Id.* at 7.
347. *Id.*
348. *Id.* at 6.
349. *Id.*
350. Carrasqueira, *supra* note 342.
351. Digital Single Market, *supra* note 69, at art. 17 §7.
352. *Id.*
353. *Id.*
354. Tackling Illegal Content, *supra* note 64, at 17–18.

of over-removal of legal content, I will make several recommendations to create new rights and guarantee remedies in favor of platform users to empower them (2).

### *A. Solutions for More Accountability and Transparency Towards the Public*

Transparency makes moderators' actions explicit and public, reveals the general moderation policies, and makes it clear how they are applied in each specific case.[355] Inversely, secret moderation hides the details. Moderation can be transparent about the *what* (that is, the content which was removed) but not the *why*, or it may be transparent only some of the time.[356] Most of the time, more transparency around content moderation by platforms is necessary, even if it does not automatically provide all the guarantees we could expect.[357] Furthermore, algorithms challenge accountability in three ways: (1) access (denied on the basis of competitive advantages, intellectual property, trade secrets…); (2) understanding; and (3) mutability (revision of the code, change of training data, contingency in a situation or use case).[358]

Fully automated deletion or suspension of content can be particularly effective and should be applied where the circumstances leave little doubt regarding the illegality of the material.[359] For instance, this may be applied when content removal is required by law enforcement authorities, or in cases involving known illegal content which has previously been removed.[360] In these cases, there should be few issues on accountability or transparency. But in other cases, more transparency is required concerning the rules created and applied by the platforms to moderate and remove the content. The European Commission maintains that online platforms should disclose their detailed content policies in their terms of service and clearly communicate this to their users.[361] These terms should not only define the policy for removing or disabling access to content, but also spell out the safeguards that ensure that content-related measures do not lead to over-removal.[362] While these measures are functioning as they should, their weakness is the simple fact they are not mandatory. More stringent rules have to be enacted by the European lawmaker to create greater transparency. It is not enough to merely encourage the platforms to disclose their rules. The reasons why content is removed have to be given not only generally but also case by case to the users who deserve a specific explanation. This would involve the requirement of clear display of the rules of removal based on the law, and also on the contract. It also means providing a simple way for users to obtain further explanation.

---

355. Grimmelmann, *supra* note 63.
356. *Id.*
357. Paul B. de Laat, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?*, 31 PHIL. & TECH. 1, 1–17 (2017).
358. Helen Nissenbaum, *Accountability in a Computerized Society*, 2 SCI. ENGINEERING ETHICS 25, 25–42 (1996).
359. Tackling Illegal Content, *supra* note 64, at 14.
360. *Id*.
361. *Id.* at 16.
362. *Id.*

In this context, an obstacle to disclosure, as well as "auditability" can occur. The platforms could invoke their intellectual property rights or trade secret[363] to avoid all kinds of communication. Nevertheless, the goal of disclosure does not concern the algorithmic rules by themselves but only their results, that is, explanation of removal as a result of their application.[364] In other words, there is no risk of breaching IP rights and trade secrets. Moreover, such rights should not be used to avoid disclosure, nor general rules of removal, nor should they be used to prevent individual explanation of a removal.

To provide an example, the disclosure of general rules of removal could involve the use of a "trusted flaggers" system. "Trusted flaggers" are notice providers who offer particular expertise in identifying illegal content, serving as a dedicated structure for detecting and identifying such online content.[365] The removal of illegal online content thus becomes more efficient and transparent, and happens more quickly and reliably when online platforms put such mechanisms in place.[366] Consequently, the European Commission encourages online platforms to make use of existing networks of trusted flaggers.[367] For instance, concerning terrorist content, Europol's Internet Referral Unit has the expertise to assess whether any given content constitutes terrorist or violent extremist online content, and uses this expertise to act as a trusted flagger, in addition to its law enforcement role.[368] The INHOPE network of hotlines for reporting child sexual abuse material is another example of a trusted flagger system.[369] For illegal hate speech content, Civil Society Organizations (CSOs) and semi-public bodies are specialized in the identification and reporting of illegal racist and xenophobic online content.[370]

Trusted flaggers help determine which content is illegal. However, with the exception of official flaggers, such as those previously cited, credibility depends on which criteria is used to qualify third parties as "trusted flaggers."[371] Such a mechanism has credit only if the process is transparent, independent and objective. The European Commission recommends that hosting service providers publish clear and objective conditions for determining which individuals or entities they consider as trusted flaggers.[372] Those conditions should aim to ensure that the concerned individuals or entities have the necessary expertise and carry out their activities as trusted flaggers diligently and objectively.

These requirements are both necessary and relevant, but they are unfortunately non-binding, which constitutes a first limitation. A second limitation concerns the necessity of providing for protective measures in the

363. Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1348 (2018).
364. *Id.*
365. Tackling Illegal Content, *supra* note 64, at 8.
366. *Id.*
367. *Id.*
368. *Id.*
369. *Id.*
370. *Id.*
371. *Id.* at 8–9.
372. *Id.* at 9.

event of abusive behavior from a trusted flagger. In the case of bad faith notices and counter-notices, a procedure of revoking the trusted flagger has to be provided according to well-established and transparent criteria.[373] These policies should be clearly described in online platforms' terms of service, and should be part of the general transparency reporting of online platforms[374] in order to increase public accountability. Such measures have to be mandatory to function properly.

Additionally, to allow true transparency, an "auditability' of content moderation results should also be mandatory upon the requests of users, organizations which defend public interests, as well as judges and the government. The goal is to monitor not only the rules of moderation (inputs), but also their results in concrete situations (outputs), in order to improve the monitoring process as a whole.[375] An independent "auditability" is necessary which could be provided by a system of peer-reviewing by experts.

Further steps should be taken to ensure greater transparency towards the public, not only in regard to the proper functioning of the platform as such, but also to compensate for damages in the event of the malfunction of the platform. Rights and remedies must be offered to users of the platforms.

### *B. Rights and Remedies in Favor of Platforms' Users*

Usually, the decisions on flagged content are rarely communicated to those who flagged it, leaving them consequently without any explanation regarding which human or automated decision made such decisions and as a result depriving them of any opportunity to respond.[376] Generally speaking, those who provide the content should be given the opportunity to contest any removal decisions via a counter-notice.[377] This is also valid when content removal has been automated.[378] The Child Pornography Directive (EU) 2011/93[379] and the Terrorism Directive (EU) 2017/541[380] provide some rights in favor of the users in the event of removal and blocking, especially a right to be informed of the reason for the removal and a right to obtain judicial redress. The copyright reform also provides some redress mechanisms, such as the right to contest the removal of content before an independent body (alternative dispute resolution

373. *Id.* at 18.
374. *Id.*
375. Common, op. cit., *supra* note 87;
376. *Id.*
377. Tackling Illegal Content, *supra* note 64, at 17.
378. *Id.*
379. Combatting Sexual Abuse, *supra* note 66 ("These measures must be set by transparent procedures and provide adequate safeguards, in particular to ensure that the restriction is limited to what is necessary and proportionate, and that users are informed of the reason for the restriction. Those safeguards shall also include the possibility of judicial redress.").
380. Combatting Terrorism Framework, *supra* note 76 ("Measures of removal and blocking must be set following transparent procedures and provide adequate safeguards, in particular to ensure that those measures are limited to what is necessary and proportionate and that users are informed of the reason for those measures. Safeguards relating to removal or blocking shall also include the possibility of judicial redress.").

system), the right to obtain a review based on a human decision, and the right to obtain judicial redress.[381]

The balance between automation, flexibility, and justice must be particularly guaranteed by the rules of due process,[382] as well as the right to appeal,[383] the right to reconsideration, and the right to obtain human intervention. While these measures provide a certain degree of protection, it should not be limited to specific content and should be extended to all kinds of illegal content. For instance, YouTube's content ID filtering system incorporates a counterclaim process.[384] However, such a right of appeal should be mandatory and not merely discretionarily ensured by private actors. Moreover, to detect any potential over-removal, human review should be the norm to take context into account.[385] Even if human-administered notice and takedown systems consistently err on the side of removing information in the face of legal risk or complexity,[386] it is better to have interaction with a human, rather than with a machine, only if the human can freely review the initially made decision. Such rights should be ensured by governmental laws and not only soft laws or internal rules from the platforms.

Besides, to redress over-removal, some fast and efficient alternative procedures must be provided. The ability to obtain a decision out-of-court process should not affect the right to obtain judicial redress in any case. In the EU Charter of Fundamental Rights, the conditions of Article 47 state a right to an effective remedy and to a fair trial.[387]

This out-of-court dispute settlement is legitimate and recommended by the European Commission if the procedure provides guarantees, such as the ICANN or WIPO Domain Names Dispute Resolution System.[388] Article 17§ 9 of the Directive 2019/790 (EU) on copyright and related rights in the Digital Single Market enacts that member states should ensure that users have access to an independent body for an amiable resolution of disputes.[389] Online content-sharing service providers put in place an effective and expeditious complaint and redress mechanism that is available to users of their services in the event of disputes over the disabling of access to, or the removal of, works or other subject matter uploaded by them.[390] Complaints submitted shall be processed without undue delay, and decisions to disable access to or remove uploaded content shall

381. Digital Single Market, *supra* note 69.

382. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1262 (2008).

383. Bridy, *supra* note 261, at 193.

384. *Id.*

385. *Id*.

386. Keller, *supra* note 61. *See also* Daphne Keller, *Empirical Evidence of "Over-Removal" by Internet Companies under Intermediary Liability Laws*, CENTER FOR INTERNET AND SOCIETY (Oct. 12, 2015), http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws (commenting on "Notice and Takedown" systems and how they function by minimizing costs).

387. Charter of Fundamental Rights of the European Union, *supra* note 262.

388. Digital Single Market, *supra* note 69; *Uniform Domain Name Dispute Resolution Policy*, INTERNET CORPORATION FOR ASSIGNED NAMES AND NUMBERS (Oct. 24, 1999), https://www.icann.org/resources/pages/policy-2012-02-25-en.

389. Digital Single Market, *supra* note 69.

390. *Id*.

be subject to human review. Such mechanisms shall enable disputes to be settled impartially and shall not deprive the user to have recourse to efficient judicial remedies. In particular, users should have access to a court or another relevant judicial authority to assert the use of an exception or limitation to copyright and related rights.

The provisions enacted in the copyright directive are interesting, but such a system should be generalized. If the counter-notice provides reasonable grounds to prove that removed activity or information is not illegal, and if we are to fight against over-removal, the platform provider should restore the removed content without undue delay and allow for the re-upload of content by the user, without prejudice to the platform's terms of services. This kind of mechanism should be easily accessible, effective, transparent, impartial, fair, and compliant with applicable law and fundamental rights.

Furthermore, to effectively guarantee the respect of this right and efficiency of judicial redress, the burden of proof must be reversed. The platforms must disclose their general rules and prove the respect of them, case by case. Regarding the asymmetry of information, the applicant is not able to prove the over-removal. Additionally, they must provide an explanation about the reasons why a removal was deemed necessary, according to the specific circumstances.

These recommendations are close to the Santa Clara principles on Transparency and Accountability in Content Moderation generated from a conference held in February 2018 and updated in May 2018.[391]

## V. CONCLUSION

In a nutshell, I have argued that the European Commission has made some interesting recommendations concerning conditions of platform content moderation and removal of illegal content. Nevertheless, there is significant room for improvement, especially through the creation of new obligations to the platforms, despite the liability regime of the E-Commerce Directive. To concretely ensure the respect of these obligations, such commitments should mainly provide more algorithmic transparency, give new rights to the users, and equally ensure access to an independent and efficient alternative procedure. The access to court should also be guaranteed, as well as the respect of fundamental rights. To fight against the problem of asymmetry, the burden of proof should be reversed. Finally, the monitoring of platforms' systems of content moderation should concern all the illegal content. Moreover, the territorial scope of all these measures should be the European Union territory and not the territory of the member states. If any of these measures could guarantee a perfect

391. *The Santa Clara Principles*, HIGH TECH LAW INSTITUTE (May 7, 2018), https://santaclaraprinciples.org ("These principles are meant to serve as a starting point, outlining minimum levels of transparency and accountability that we hope can serve as the basis for a more in-depth dialogue in the future: (1) Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines; (2) Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension; (3) Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.").

enforcement and effective monitoring of the platforms, the sum of them would help to improve the trust of the platforms' users on the content moderation by social media.