Regulating online content moderation: Taking stock and moving ahead with procedural justice and due process rights

Orit Fischman-Afori

INTRODUCTION

Today most aspects of our societal activities are conducted in the digital sphere, which has become the main arena where information and content flow. This environment has enabled digital free speech to flourish on a global scale, although there is already a worldwide consensus that online speech may be harmful. The devastating effects of misinformation, fake news, and illegal content pose a significant challenge to democracies around the world. To mitigate this harm, a practice of content moderation has emerged, in which online platforms monitor and remove harmful content. Content moderation practices are initiated to remove content for various reasons, including breaking the law or the policy and terms of use of the online platform. This practice, adopted...
voluntarily by the online platforms, raises another challenge to democracies, as private corporations governing the online environment in practice control digital speech. These online platforms control the backbone of democracies, and their practices have extensive effect on individuals’ fundamental rights and freedoms. The question is, therefore, whether and how the private power in the digital sphere should be regulated and whether content moderation practices should be restrained to protect users’ fundamental rights. In recent years, the online platforms have deployed advanced computerized technologies, known as artificial intelligence (AI), in the application of content moderation practices. Because the decision-making process is conducted by algorithms, new obstacles have emerged in protecting users’ fundamental rights and safeguarding the digital free speech environment. Thus, regulating artificial intelligence technologies has joined the overall attempt to regulate content moderation practices.

The EU has initiated a line of legislative efforts to regulate online digital speech, addressing content moderation practices directly and indirectly. Together, the various legislative initiatives that have emerged from distinct perspectives aimed at addressing different challenges are generating a corpus of digital governance norms for regulating content moderation practices. Three initiatives are of particular interest because they represent significant, and in two cases even landmark, moves toward building a comprehensive digital governance regime. The first is the Directive on Copyright in the Digital Single

---


6 An early report published in 2011 by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression paid special attention to safeguards of freedom of expression on social media platforms, see Special Rapporteur on the Promotion and Protection of Freedom of Opinion and Expression, A/HRC/17/27 (May 16, 2011). Another report, issued in 2018 by the UN Special Rapporteur, was based on a global survey and sought to collect empirical data of voluntary and imposed content moderation practices. The overall finding was that on a global scale, the private sector does not adequately protect freedom of speech; see Special Rapporteur on the Protection and Promotion of the Right to Freedom of Opinion and Expression, Overview of submission received in preparation of the Report of the Special Rapporteur, 2–3 (A/HRC/38/35), A/HRC/38/35/Add.1. (Jun. 6, 2018).

7 The UN Special Rapporteur’s report issued in 2018 emphasized that content moderation is often operated by algorithmic decision-making processes, which are unaccountable for any results affecting individuals’ human rights; see Special Rapporteur on the Protection and Promotion of the Right to Freedom of Opinion and Expression, Overview of submission received in preparation of the Report of the Special Rapporteur, 40–49 (A/HRC/38/35), A/HRC/38/35/Add.1. (Jun. 6, 2018).
Regulating online content moderation

Market, adopted in 2019,\(^8\) concerning the removal of content that infringes copyright. The second is the Digital Services Act,\(^9\) adopted in 2022, regulating online services for accomplishing a trustworthy digital sphere. The third is the Artificial Intelligence Act proposal,\(^10\) on the use of AI technologies both online and offline. This chapter briefly reviews these three initiatives, addressing in particular the issue of content moderation. Inspecting these initiatives together reveals that although the EU has intended to enshrine human rights guarantees in the digital sphere, it has failed to provide individual end users with procedural justice rights, known also as procedural due process rights, to protect them \textit{vis-à-vis} the online platforms. In the precedent-setting General Data Protection Regulation (GDPR), adopted in 2018, the EU has acknowledged the procedural rights of individuals \textit{vis-à-vis} operators of systems that use personal data;\(^11\) no similar acknowledgment exists with respect to content moderation. The EU digital governance regime fails to acknowledge the rights of individuals to demand full disclosure of all the information that has led to the decision of a platform to remove content, and it does not provide an adequate right to challenge such decisions through an external and objective body. Although the principles of transparency and accountability are praised at the preambles of these legislative initiatives, they are not translated into concrete obligations similar to the procedural due process rights granted to individuals in public and administrative law against the use of power by the state. Yet in the digital governance regime, such rights are essential for protecting individuals against the power of the private sector.


1. **BACKGROUND: CONTENT MODERATION AND THE AI PHASE**

Today, online platforms serve as the main arena of public discourse. Massive amounts of information, data, and content are conveyed by a range of means, including text, images, and audio. The hyper-dynamic digital environment has produced a digital information society, in which freedom of speech flourishes. This thriving environment is supported by a regulation that provides immunity to online platforms from imposition of civil liabilities for harms caused by content that users upload, for being “intermediaries” in the dissemination of the content. The two leading legal frameworks that created this safe harbor regime are Article 230 of the US Communications Decency Act (CDA), adopted in 1996,12 and the EU e-Commerce Directive, adopted in 2000.13

Yet, the massive flow of content has raised a serious challenge to democracies worldwide because it may be used as an arena for unwarranted or illegal speech as well. The most discussed examples are hate speech, the dissemination of false or misleading information, and illegal content that infringes on other individuals’ rights, such as copyright and privacy.14 Harmful digital speech encompasses many forms and types of content, each raising different social and legal concerns.

Despite the safe harbor enjoyed by online intermediaries, they have undertaken to voluntarily assume an active role in removing offensive content.15 The existing legal framework leaves online actors with vast discretion in the matter, without concrete guidelines. Therefore, the online platforms, in a gradual process, have adopted practices concerning monitoring, filtering, and blocking the various types of harmful content, for various reasons. Together, these practices are referred to as “content moderation.”

Article 230 to the CDA does not offer immunity based on intellectual property infringement; therefore a special clause was enacted in the US Digital Millennium Copyright Act of 1998 (DMCA).16 Section 512 to the US Copyright Act creates a safe harbor for some online intermediaries

---

14 See supra note 4.
in case of content infringing without their knowledge,\textsuperscript{17} and it establishes a notice-and-takedown mechanism for all Internet intermediaries that governs the monitoring of copyrighted content. By contrast, the EU e-Commerce Directive, which established immunity for the intermediaries on various legal grounds—albeit with no mandatory notice-and-takedown scheme—has left room for the adoption of such a scheme voluntarily, with such schemes having proliferated in EU countries as well.\textsuperscript{18} An extensively discussed outcome of the notice-and-takedown regime, created by § 512 to the Copyright Act, is the mass and easy removal of allegedly infringing copyrighted content, which has had a significant chilling effect on freedom of speech.\textsuperscript{19} Because intermediaries are risk-averse, they have an incentive to respond positively to all takedown requests, even if such requests could have been found unjustified had they been decided in court.\textsuperscript{20} As an uninvolved third party in the dispute, the intermediaries have no incentive to invest time and effort in a profound legal assessment of requests, and the outcome is massive and uncontrolled removal of content.\textsuperscript{21}

Another legal framework for content moderation are “blocking orders,” injunctions usually granted against Internet service providers, ordering them

\begin{footnotesize}
\textsuperscript{17} See Viacom Int’l, Inc. v. YouTube, Inc., 676 F.3d 19, 30 (2d Cir. 2012). For the ramifications of this case and its extensive litigation, until it was settled by the parties outside the court, see John T. Williams & Craig W. Mandell, Winning the Battle, but Losing the War: Why the Second Circuit’s Decision in Viacom Int’l, Inc. v. YouTube, Inc. is a Landmark Victory for Internet Service Providers, 41 AIPLA Q. J. 235 (2013).

\textsuperscript{18} Following the EU e-Commerce Directive, Finland is the only EU member state that has adopted a statutory notice-and-takedown mechanism, see Stefan Kulk, Internet Intermediaries and Copyright Law: EU and US Perspectives, 117 (2019).


\textsuperscript{20} See Jack Balkin, Old School/New School Speech Regulation, 127 Harv. L. Rev. 2296, 2314 (2016).

\textsuperscript{21} See, e.g., Jeffrey Cobia, The Digital Millennium Copyright Act Takedown Notice Procedure: Misuses, Abuses, and Shortcomings of the Process, 10 Minn. J. Sci. & Tech. 387, 390–93 (2009) (noting abuses of current takedown practices, particularly highlighting the fact that content that does not constitute a copyright infringement is often taken down).
\end{footnotesize}
to block access to a certain website or source of content, even directly ordering the removal of content. Blocking orders have been granted, occasionally and extensively, mainly in European countries, including Austria, Belgium, Denmark, Finland, France, Italy, Ireland, and the UK. The European Court of Justice has approved this practice, which is consistent with the European Directives on the matter. Outside Europe, blocking orders are granted, for example, in Australia, Israel, and Canada. Similarly to the notice-and-takedown framework, blocking orders have caused controversy, despite being granted by the Court and clearly being subject to judicial oversight because of their underlying policy and compatibility with applicable measures, such as efficiency, necessity, and especially proportionality, relating to constitutional remedies.
Regulating online content moderation

Content moderation practices, whether conducted through blocking orders, the notice-and-takedown framework, or any other method, raise concerns with respect to their effect on freedom of speech in the digital sphere. Because the content moderation policies of online platforms are voluntary, a key question raised in these controversies is whether some basic obligations should be imposed on online actors, aimed at safeguarding freedom of speech. These controversies also raise the question of whether to adopt some “must-carry” obligations, i.e., a rule that prevents removal of certain content, considering the key role that online platforms play in the digital speech environment.

The controversies stemming from content moderation practices have been transformed in the last few years, with the growing use of a family of technologies, collectively referred to as AI, in the implementation of these practices. AI designates technologies that train algorithms to produce various outputs, including content, predictions, recommendations, and decisions. A particular AI technology, known as machine learning, trains algorithms to run on constantly updated big datasets and detect patterns used to autonomously generate outputs such as observations and decisions. Machine learning is a data-driven technology that adapts its performance to the inputs it receives. The end goal of these technologies is “to allow the computers to learn automatically without

29 See Balkin, supra note 20.  
32 The WIPO Report on AI offers a much broader view, according to which “AI systems are viewed primarily as learning systems; that is, machines that can become better at a task typically performed by humans with limited or no human intervention. This definition encompasses a wide range of techniques and applications…”, see WIPO Technology Trends 2019: Artificial Intelligence (WIPO 2019) https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf. Regarding the various ways to define AI, see also Sonia K. Katyal, Private Accountability in the Age of Artificial Intelligence, 66 UCLA L. Rev. 54 62–63, (2019); Bryan Casey & Mark A. Lemley, You Might Be a Robot, 105 Cornell L. Rev. 287, 311 (2020).  
34 Greenstein, supra note 33, at 9.
human intervention or assistance and adjust actions accordingly.\textsuperscript{35} AI has been used by the business sector for a range of tasks, from offering services replacing the human workforce to the operation of content moderation systems by online platforms.\textsuperscript{36} The use of AI systems for content moderation has amplified earlier problems related to the control that the platforms exercise over the digital sphere, now that this control has been given over to algorithms. Today, the threat to democratic values and freedom of speech comes from automated, computer-controlled devices.\textsuperscript{37} As policies regarding content moderation are translated into algorithm design and setting the defaults, the silencing mechanism has become an algorithmic one.\textsuperscript{38} This emerging reality in which online platforms control the civil digital speech environment with the aid of AI systems challenges civil society organizations, academics, and policymakers at both national and international levels.

2. EU REGULATION EFFORTS: THREE DOTS ON THE CURVE OF DIGITAL GOVERNANCE

Attempts by the EU to address the various challenges stemming from the online digital environment are reflected in a line of legal initiatives aimed at regulating the digital sphere. Three key initiatives are of interest regarding online content moderation practices. Although each initiative addresses a different aspect of the issue, they appear as three dots on the curve of the emerging EU digital governance regime, aggregately forming a clear trajectory of content moderation regulation.

2.1 Article 17 of the Directive on Copyright in the Digital Single Market

The first legal measure addressing online content moderation is the Directive on Copyright in the Digital Single Market (CDSM Directive), which was

\textsuperscript{35} Greenstein, \textit{supra} note 33, at 10.
Regulating online content moderation

adopted in the spring of 2019.\textsuperscript{39} The underlying policy of the CDSM Directive is to strengthen the position of copyright owners \textit{vis-à-vis} the various online platforms, considering what is called the “value gap,” or the gap between the economic value generated by online platforms and its fair sharing with content creators and copyright owners.\textsuperscript{40} This development reflects the first legal move for imposing mandatory \textit{active} content monitoring obligations on online platforms. The strong opposition to the CDSM Directive was based on the fear that it would impede digital freedom of speech. But despite a great public outcry\textsuperscript{41} and resistance of several member states, it was eventually approved.\textsuperscript{42}

Article 17 to the CDSM Directive contains a series of obligations imposed on the “online content sharing service provider,” including an obligation to obtain an authorization from rightholders with respect to any copyrighted work that is intended to be communicated by their services.\textsuperscript{43} In the absence of such authorization, the intermediary must demonstrate that it made best efforts to prevent the availability of works explicitly identified by rightholders.\textsuperscript{44} Upon notification by rightholders, the intermediary should act expeditiously to remove these works,\textsuperscript{45} and the intermediary should “provide rightholders, at their request, with adequate information on the functioning of their practices with regard to the cooperation referred to as concerning removal of content.”\textsuperscript{46}

Article 17(9) also obliges intermediaries to establish an effective and expeditious complaint and redress mechanism, in the service of users disputing the removal of content, e.g., wishing to benefit from exceptions or limitations to copyright. The complaints should be “processed without undue delay,” and decisions to remove content following such complaints must be “subject to


\textsuperscript{40} See Kulk, supra note 18, at 61–62.


\textsuperscript{42} The final approval in spring 2019 was not unanimous: three member states abstained and six voted against. In a joint statement, the opposing member states explained that the Directive may encroach on EU citizens’ rights, and therefore, they have voted against its adoption, see Kulk, supra note 18, at 63 (explaining that these member states felt that the text of the approved Directive would stifle innovation).

\textsuperscript{43} CDSM Directive, supra note 39, at 119.

\textsuperscript{44} Id. at 120.

\textsuperscript{45} Id.

\textsuperscript{46} Id.
human review.” The Article further stresses that “Member States shall also ensure that out-of-court redress mechanisms are available for the settlement of disputes[,]” that such mechanism shall enable disputes “to be settled impartially[,]” and that users shall not be deprived of the legal protection afforded by their national law, and have “access to a court or another relevant judicial authority.”

Article 17 to the CDSM Directive decrees the mandatory adoption of a nuanced notice-and-takedown scheme, but in practice, these schemes have already been adopted in a self-regulatory move. But Article 17 contains additional, far-reaching active obligations for content monitoring, such as making “best efforts” in fulfilling the duty to have rightholders’ authorization for the use of their content, and in accordance with “high industry standards of professional diligence,” ensuring the unavailability of content that the rightholder indicated. These obligations concern the implementation of content recognition technologies, similar to the “content ID” systems used voluntarily by YouTube. The difference, however, is that in contrast to the notice-and-takedown scheme, in which the online platform responds to a conflict initiated by a third party, Article 17 shifts the responsibility to the online platform to proactively initiate the process of obtaining authorization and thereby of content filtering and removal, using content recognition technologies. Article 17 also includes an active “staydown” obligation, aimed at making sure that there is no re-uploading of removed content.

Member states have initiated the implementation of the CDSM Directive, but with some delay because of various reasons, including the wait for the decision of the Court of Justice of the EU (CJEU) in the case brought by Poland against Article 17. Poland claimed that Article 17 should be annulled because it is not consistent with the principle of freedom of expression and information of users of sharing services, guaranteed by Article 11 of the

---

47 Id. at 120–21.
48 See Kulk, supra note 18, at 116–120.
49 Copyright Digital Single Market Directive, supra note 39, at 120. See also Article 17(4)(b)–(c).
50 Kulk, supra note 18, at 133.
51 See Copyright Digital Single Market Directive, supra note 39, at 120 (Article 17 (4) (c)) stipulates that the service providers should make “best efforts to prevent their future uploads in accordance with point (b))”. See also Felipe Romero-Moreno, Notice and Staydown and Social Media: Amending Article 13 of the Proposed Directive on Copyright, 33 INT’L REV. L. COMPUTERS & TECH. 187, 203–04 (2019).
Regulating online content moderation

Charter of Fundamental Rights of the European Union.\textsuperscript{54} Content ID systems also raised the concern about having a chilling effect on digital speech because they may block the non-infringing use of copyrighted content that could have qualified for fair or permitted use.\textsuperscript{55} An automated system can produce false positive outcomes, which are inconsistent with freedom of speech as the default principle.\textsuperscript{56} The YouTube transparency report, published in December 2021, reveals that over-blocking is real, and that 60\% of takedown notices issued in the first half of 2021 that were disputed by users were resolved in favor of the user.\textsuperscript{57} Had these users not filed a dispute, the content would have been wrongly silenced.\textsuperscript{58}

The main problem of content recognition technologies is that they are implemented by private entities, which are not obligated to reveal the design of their algorithms. Therefore, these systems are operated without safeguards for the protection of users and of the public at large.\textsuperscript{59} Yet, in the opinion of the Advocate General in response to the case brought by Poland, submitted in July 2021, it was claimed that Article 17 sets internal guarantees for digital freedom of speech because this rule is the exception to the general approach taken by the EU, which does not mandate content moderation as a general practice. Furthermore, it was claimed that the rule is proportionate because it balances the competing interests of rightholders and users.\textsuperscript{60} The European Commission published its guidelines on the application of Article 17 a month earlier, in June 2021, stressing the necessity of safeguarding freedom of speech and other

\textsuperscript{54} \textit{Id.}
\textsuperscript{59} See, e.g., Senftleben \textit{et al.}, \textit{supra} note 55, at 4. The GDPR provides certain safeguards to users – however, only with respect to matters concerning privacy, see \textit{supra} note 11.
fundamental rights affected by Article 17.\textsuperscript{61} Therefore, it was argued that the Advocate General’s opinion contradicts to some extent that of the European Commission, which is less decisive with regard to how it is possible to reconcile Article 17 obligations with freedom of speech.\textsuperscript{62} On April 26, 2022, the CJEU handed down its decision, rejecting the Polish claim and adopting the stance of the Advocate General, by determining that Article 17 has been “accompanied by appropriate safeguards by the EU legislature in order to ensure … respect for the right to freedom of expression and information of the users of those services … and a fair balance between that right, on the one hand, and the right to intellectual property … on the other.”\textsuperscript{63} The CJEU has also stressed, however, that:

Member States must, when transposing Article 17 of Directive 2019/790 into their national law, take care to act on the basis of an interpretation of that provision which allows a fair balance to be struck between the various fundamental rights protected by the Charter. Further, … Member States must … make sure that they do not act on the basis of an interpretation of the provision which would be in conflict with those fundamental rights or with the other general principles of EU law, such as the principle of proportionality.\textsuperscript{64}

Therefore, the Court has transferred to member states the challenge of discovering how Article 17 can be implemented in a way that would generate an appropriate and balanced framework that takes into consideration freedom of speech.

2.2 The Digital Services Act

Another prominent legal initiative pertaining to the online environment in general, which also affects online content moderation practices, is the Digital

\textsuperscript{61} Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market, COM 2021, 288, June 4 2021, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1625142238402&uri=CELEX%3A52021DC0288 (stating that “Member States should implement or adapt the mandatory exceptions and limitations in Article 17(7) in a manner that allows them to be applied consistently with the Charter of Fundamental rights and ensures their effectiveness in line with the case law of the CJEU.” \textit{Id.} at p. 19).


\textsuperscript{64} \textit{Id.} at par. 99.
Regulating online content moderation

Services Act (DSA), adopted in 2022, to ensure a safe and accountable online environment.\textsuperscript{65} The DSA reflects an attempt to establish comprehensive and systematic regulation for the various online services and to introduce some safeguards of fundamental rights in the online environment. This initiative is the first important piece of EU legislation for the digital sector since the e-Commerce Directive of 2000.\textsuperscript{66} The new proposed regime will serve as a key element in building digital governance principles, with a potential worldwide effect.\textsuperscript{67}

The DSA is based on the principle of tailoring the obligations imposed on service providers to the nature of the service and the size of the provider.\textsuperscript{68} The proposal differentiates between various digital service providers. Some general obligations are imposed on most services, but stricter obligations are imposed on online platforms, which are defined narrowly to encompass services such as those provided by social networks (e.g., Facebook) and content storage and dissemination platforms (e.g., YouTube).\textsuperscript{69} As clarified at the outset, the proposed regulations “do not provide full-fledged rules on the procedural obligations related to illegal content and they only include basic rules on transparency and accountability of service providers and limited oversight mechanisms.”\textsuperscript{70} It is further stipulated that there is no general obligation to monitor information or a duty to actively search for illegal content.\textsuperscript{71} At the same time, however, the DSA reflects a significant move toward the expansion of various procedural obligations imposed on online service providers. Although no monitoring obligations are imposed on very large platforms, serving 10\% or more of the EU population, they are subject to some other obligations, such as risk assessment concerning the traffic of illegal content and the negative effects of content moderation on freedom of speech.\textsuperscript{72} Consistent with the general policy of setting only general accountability duties, these very large online platforms are mandated to “put in place reasonable, proportionate and effective miti-

\textsuperscript{65} See supra note 9.
\textsuperscript{67} See https://www.eff.org/deeplinks/2020/12/eu-and-digital-services-act-year-review.
\textsuperscript{68} DSA, at 6.
\textsuperscript{69} DSA, Article 2 (Definitions).
\textsuperscript{70} DSA, at 4.
\textsuperscript{71} DSA, Article 7.
\textsuperscript{72} DSA, Article 25, 26.
The exploitation of intellectual property rights
gation measures, tailored to the specific systemic risks identified.” Yet, the DSA does not specify exactly what these measures are.

Chapter III of the DSA contains “due diligence obligations for a transparent and safe online environment,” and it includes obligations concerning the procedures for the operation of the various services. For example, service providers are obligated to provide accessible information regarding any policy and contractual terms relating to content moderation, including any measures and tools used for such purpose, whether based on algorithmic or human decision-making. The DSA offers measures to open the “black box” of algorithmic content moderation with certain transparency obligations, but it does not require human determination in the process, despite many concerns expressed on this topic. It specifies that service providers are required to “act in a diligent, objective and proportionate manner,” and “with due regard to the rights and legitimate interests of all parties involved, including the applicable fundamental rights of the recipients of the service as enshrined in the Charter.” The DSA also requires the various service providers to periodically publish transparency reports. Regarding the reasoning of decisions to remove content, service providers are required to inform the recipient, i.e., the affected user, “at the latest at the time of the removal or disabling of access, of the decision and provide a clear and specific statement of reasons for that decision.” This obligation specifies the content to be included in such a notification, ensuring its substantive basis. Regarding the possibility of challenging the decision, it stipulates that the service provider must provide information about the available options of either “internal complaint-handling mechanisms, out-of-court dispute settlement and judicial redress.” With respect to online platforms in particular, the establishment of an easily accessible internal complaint-handling mechanism, and in some cases of an out-of-court dispute settlement mechanism, are mandatory. The DSA also requires to establish a “certified” out-of-court dispute settlement body that would meet basic standards of independence and apply “clear and fair rules of procedure.” Finally, a key requirement introduced by the DSA, aimed at countering the over-blocking of content that hinders freedom of speech, concerns a “put

73 DSA, Article 27.
74 DSA, at 33–4.
75 DSA, Article 12.
76 DSA, Articles 13, 23. Very large online platforms are subject to additional periodic transparency report obligations, see DSA, Article 33.
77 DSA, Article 15.
78 DSA, Article 18 (1).
79 DSA, Article 17, 18 (1).
80 DSA, Article 18.
back” obligation.\textsuperscript{81} Yet, the DSA grants discretion to the online platforms on deciding whether or not removal of content was justified, so that in borderline cases the risk of false positive decisions increases.\textsuperscript{82}

A challenging question concerns the relation between Article 17 of the CDSM Directive and the DSA requirements regarding content moderation practices. Whereas Article 17 reflects a \textit{lex specialis} regime, the DSA provides more detailed requirements on various issues.\textsuperscript{83} For example, Article 17 requires that the online platform give “a sufficiently substantiated notice” before takedown, yet the term “substantiated” is not explained. By contrast, in Article 14, the DSA devises a detailed mechanism for the notice that should be given before action, which should include “an explanation of the reasons why the individual or entity considers the information in question to be illegal content,” as well as other needed information.\textsuperscript{84} Article 14 to the DSA also requires that the services “take their decisions in respect of the information to which the notices relate, in a timely, diligent and objective manner.” Where they use automated means for that processing or decision-making, they shall include information on such use in the notification referred to.\textsuperscript{85} Although the DSA is the general framework for content moderation regulation, and the CDSM Directive is a specific norm regarding content moderation in particular cases of copyright infringement, the DSA formulates much more detailed regulations, which should illuminate the interpretation of the specific norm. This is because the DSA is intended to take the digital governance one step forward and to provide comprehensive guidelines for the conduct of the online platforms for the sake of protecting fundamental rights, such as freedom of speech.\textsuperscript{86}

\subsection{2.3 The Artificial Intelligence Act}

In the last few years there has been a growing movement calling for regulating the use of AI systems. Civil society organizations, policymakers, scholars, and

\begin{itemize}
\item DSA, Article 17 (3).
\item DSA, Article 14 (2).
\item DSA, Article 14 (6).
\item Peukert et al., supra note 81.
\end{itemize}
various grassroots initiatives have been generating an increasing number of reports stressing the need to regulate the use of AI technologies in all sectors.\textsuperscript{87} Two prominent examples of international attempts to formulate principles for recommended AI regulation are the OECD recommendations to policymakers, in 2019,\textsuperscript{88} and the UNESCO recommendation on the ethics of AI, in 2021.\textsuperscript{89} Both recommendations address the need to provide guarantees for the protection of human rights in the operation of AI systems in the various sectors. Another international initiative has been led by the Council of Europe \textit{ad hoc} committee on AI (CAHAI), which promotes an international legal framework for regulating AI, based on the standards of human rights and the rule of law.\textsuperscript{90}

The EU made its first move in 2020, with the White Paper on AI, stressing the need to promote an overarching regulatory setting that would allow the development of a trustworthy AI environment.\textsuperscript{91} The EU Commission noted that although AI technology has the potential to change human lives by promoting the public good in a range of aspects, it entails potential risks, “such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes.”\textsuperscript{92} Consistent with this statement, the European Parliament has adopted several resolutions relating to AI, including on copyright.\textsuperscript{93} In April 2021, the EU Commission presented its proposal for an Artificial Intelligence Act (AI Act),\textsuperscript{94} aimed at guaranteeing that the function of AI technologies conforms to the “Union values, fundamental rights and principles.”\textsuperscript{95} This proposed legislation covers

\begin{thebibliography}{99}
\item\textsuperscript{87} For the various AI initiatives conducted at the national level worldwide, see https://www.coe.int/en/web/artificial-intelligence/national-initiatives. For various voluntary programs for self-regulating the use of AI, see Carlos Ignacio Gutierrez, \textit{Identifying Incentives for the Enforcement of Artificial Intelligence Soft Law Programs} (July 31, 2021), available at SSRN: https://ssrn.com/abstract=3897486.
\item\textsuperscript{89} UNESCO 2021, 41C/73 Recommendation on The Ethics of Artificial Intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000379920.page=14.
\item\textsuperscript{90} See https://www.coe.int/en/web/artificial-intelligence/cahai.
\item\textsuperscript{92} \textit{Id.} at p. 1.
\item\textsuperscript{93} European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI).
\item\textsuperscript{95} AI Act, at p. 1.
\end{thebibliography}
a wide array of topics, from prohibition of use of various types of AI systems\textsuperscript{96} to the imposition of operational requirements according to the type of the AI system and its categorization. AI systems falling in the category of “high risk” will be subject to a strict standard of requirements,\textsuperscript{97} including implementation of a risk management system\textsuperscript{98} and data management governance standards.\textsuperscript{99} Certain transparency obligations are also required for the limited purpose of enabling the operator of an AI system “to interpret the system’s output and use it appropriately,”\textsuperscript{100} or to provide the operator of the AI system with appropriate instructions.\textsuperscript{101} The AI Act does not grant the end user, i.e., the non-professional user, any individual rights, and its focus is on regulating either the producer or the operator of AI systems. The only obligation relating to end users’ rights concerns the disclosure of the fact that an interaction is conducted with a machine, particularly when it pertains to a chat, so that when a chat is conducted with a bot, the nature of the machine–human interaction must be disclosed up front.\textsuperscript{102}

The proposed AI Act represents a comprehensive attempt to regulate the operation of AI systems, whether they are deployed as part of the online environment or in the course of “offline” life, as part of products or services. But because the use of AI systems by the online platforms has become pervasive, the AI Act will serve as a complementary measure in building the online digital governance regime, in particular, content moderation practices implemented by AI systems, such as the Content ID systems used by YouTube, discussed above. Thus, algorithmic content moderation systems would be regulated by the AI Act, like all other “high-risk” AI systems, and therefore would need to meet the various requirements imposed on producers and operators of AI systems. The mandatory standard of the AI Act pertaining to the technical dimensions of operational content moderation systems would serve as an additional layer, on top of the DSA provisions. Whereas the AI Act is focused on the producers’ and operators’ obligation to comply with adequate standards, the DSA reflects a more consumer-centric approach, addressing the conduct of the online platforms vi\textit{s-à-vis} end users. Together, the AI Act and the DSA provisions will serve as an overarching regime, joining Article 17 of the CDSM Directive, which enshrines the obligation to monitor and filter online content to prevent copyright infringement.

\textsuperscript{96} AI Act, Article 5.
\textsuperscript{97} AI Act, Articles 7, 8.
\textsuperscript{98} AI Act, Article 9.
\textsuperscript{99} AI Act, Article 10.
\textsuperscript{100} AI Act, Article 13(1).
\textsuperscript{101} AI Act, Article 13(2).
\textsuperscript{102} AI Act, Article 52.
In sum, these three legal initiatives may be perceived as three elements of the digital governance regime at the EU level, which will govern the practices used for online content moderation in general, and the monitoring of infringing content in particular.

3. THE MISSING LINK: PROCEDURAL JUSTICE AND DUE PROCESS RIGHTS

The new digital governance regime designed by the EU seeks to safeguard human rights and preserve the rule of law in the digital sphere. The main criticism of the new regime is the absence of measures aimed at protecting individuals from the use of power by the operators of the digital sphere, and the failure to provide adequate guarantees for human rights, in particular of digital freedom of speech. In this respect, the EU has gone the extra mile to protect individuals’ privacy rights, when in 2018 it adopted the General Data Protection Regulation (GDPR), acknowledging the concept of data subject rights, that is, the rights of individuals vis-à-vis operators of systems that use personal data.103 No similar measure was enacted regarding content moderation practices. Consequently, there are no adequate protective measures of the personal rights of end users that may be injured by content removal. The “missing link” of the EU digital governance regime pertaining to online content moderation practices is, therefore, the lack of procedural justice rights, known as procedural due process rights, which should provide adequate guarantees for individuals’ rights to digital free speech.104

Procedural justice measures, referred to in US terminology as procedural due process, have been developed in the realm of public law, particularly in its sub-branch of administrative law. The underlying rationale of public law procedural justice principles is to provide individuals with measures aimed at protecting them from the power of the state.105 Basic administrative law principles are headed by the general notion of “accountability.” Accountability in administrative law means that public authorities should act according to basic

---

103 See GDPR, supra note 11, at Chapter III – Rights of the Data Subjects.
104 For a similar stance, see SUZOR, supra note 5, at 144–145. For the view that the CDSM Directive lacks procedural safeguards to users’ rights and that there is a need to anchor such measures based on the DSA provisions, see Sebastian Felix Schwemer & Jens Schovsbo, What Is Left of User Rights: Algorithmic Copyright Enforcement and Free Speech in the Light of the Article 17 Regime, in INTELLECTUAL PROPERTY LAW AND HUMAN RIGHTS (P. Torremans ed., Kluwer Law International 2000).
Regulating online content moderation

standards of transparency, providing reasons for their decisions, and their decisions should be subject to objective external review.106 These are the three key procedural rights providing guarantees that the decision-making process of public authorities is “accountable,” that is, conforming with the rule of law.107 Procedural justice measures, headed by the principle of transparency, are also aimed at promoting trust in public authority and at supporting the legitimacy of the public sector.108 Therefore, the public sector is subject to far-reaching disclosure obligations, some of which are enshrined in the freedom of information legislations.109

The present chapter proposes that content moderation practices deployed by the online platforms incorporate the three basic procedural principles that are embedded in administrative law, to provide individuals with adequate protection from the power of the private corporations that control the digital sphere. To accomplish the end goal of a trustworthy digital environment, genuine transparency standard, as applied in administrative law, is essential. The justifications for public law norms acknowledged in modern democracies in the 20th century should be imported into the new digital environment, where private corporations are the main source of sovereignty.110 Thus, when a decision is made regarding content on a platform, the affected individual should be given all the relevant information that was used in the decision-making process; a reasoned explanation of the decision, in a manner that the individual can understand the decision and dispute it; and a dispute mechanism managed by an objective, external, independent entity. Due process with regard to the activities of the platforms should be based on these three basic procedural principles.


This does not mean that content moderation policies of the platforms must be approved by a court or an administrative tribunal, but rather that the decision-making process of the online platforms should be subject to the three key procedural principles applicable in administrative law. The goal is to provide individuals with procedural measures enabling them to challenge decisions that they perceive as prejudicing their personal rights.

All three EU legislative initiatives described above stop short of imposing full-fledged procedural justice principles on online platforms. Article 17 of the CDSM Directive imposes a mandatory obligation to monitor content and to take down allegedly infringing items. The sections aimed at protecting the users’ interests by balancing some of the rightholders’ benefits do not provide procedural rights for individuals. Article 17(7) requires that “Member States shall ensure that users in each Member State are able to rely on any of the following existing exceptions or limitations when uploading and making available content generated by users on online content-sharing services.” As far as the exceptions are concerned, there is a general obligation to allow the uploading of content regarded to be permitted for use, but this obligation is not elaborated into individual rights vis-à-vis the online platforms or into a clear procedural mechanism in case the platform did not allow permitted content to be uploaded. The only aspect that pertains to users’ rights concerns the obligation of the online platform to establish “an effective and expeditious complaint and redress mechanism.” But the CDSM Directive does not require that this mechanism be based on basic principles of independence and therefore of objectivity, and as such, it does not meet the procedural justice threshold. Such a complaint mechanism may be in the form of a complaint office for business customers, which, although it may be highly efficient, is not a mechanism aimed at providing adequate guarantees for human rights, and is not competent to rule in freedom of speech conflicts.

The proposed DSA represents a landmark legal move, by promoting a digital governance regime that imposes extensive obligations on various online service providers to enhance their accountability as gatekeepers of the digital sphere. But it also stops short of imposing full-fledged administrative law-like procedures. The most far-reaching duties it imposes are to act with “due regard” for human rights and to provide services that would meet the threshold of procedural “objectivity” and “proportionality.” These standards are an important step in the process of introducing public law principles

111 Kyle Langvardt, Regulating Online Content Moderation, 106 Geo. Mason L. Rev. 1353, 1357, 1376 (2018) (proposing that the policy for content monitoring in the US, legislated at a federal level, should be approved by court or an administrative body for the platform to enjoy a safe-harbor clause).

112 Article 17 (9) CDSM Directive.
into the digital governance environment, but the term “due regard” reflects the overall policy of the DSA, which is to impose only partial procedural obligations on online platforms, which do not rise to the level of genuine public law standards. The DSA also requires that various service providers periodically publish transparency reports. Although these reports are quite detailed, they do not include explicit transparency obligations concerning individual entitlements. The DSA requirement for establishing either “internal complaint-handling mechanisms” or “out-of-court dispute settlement and judicial redress” means that although external objective judicial review may be available in accordance with applicable laws, it is not mandatory. Thus, the proposed DSA does not aspire to emulate administrative law, and it does not seek to implement the “fuller package” of procedural guarantees of individuals’ rights. The only DSA requirement that may be regarded as a significant step toward full implementation of public law procedural standards is the notice that online platforms must provide regarding their decision. Article 14 of the DSA requires “an explanation of the reasons why the individual or entity considers the information in question to be illegal content.” But without a corresponding right of the individual to dispute such explanations before an external and independent body, transparency does not accomplish its procedural justice end. It is doubtful whether transparency for its own sake can generate the trustworthy digital environment the EU is seeking.

Finally, the proposed AI Act exacerbates the failure to provide individuals with rights concerning procedural justice. Both the OECD and UNESCO have recommended several basic principles that should underlie any AI regulation, including transparency and explainability, as part of the larger principle of accountability. Yet, the proposed AI Act, which regulates the manufacturer and operator of high-risk AI systems, grants no rights to the individual end user. The transparency and explainability obligations are imposed on the relations of the manufacturer and operator of high-risk AI systems, leaving

---

113 Naomi Appelman, João Pedro Quintais & Ronan Fahy, Using Terms and Conditions to apply Fundamental Rights to Content Moderation, in To Break up or Regulate Big Tech? Avenues to Constrain Private Power in the DSA/DMA Package, 29 (eds. Heiko Richter, Marlene Straub & Erik Tuchtfeld, Max Planck Max Planck Institute for Innovation and Competition, 2021).

114 DSA, Articles 13, 23. Very large online platforms are subject to additional periodic transparency report obligations, see DSA, Article 33.

115 Appelman et al., supra note 111, at 32.

116 OECD 2019, supra note 86. UNESCO 2021, supra note 87, at p. 8–11.

117 Except from the obligation of the operator of AI systems to notify the end user that the interaction is conducted with a machine, as in the case of a chat with a bot; see AI Act Article 52.

118 AI Act Article 13.
the end user with only potential general law claims *vis-à-vis* the operator of the AI system.\(^{119}\) In this regard, the DSA proposes a much more human rights-oriented approach to digital governance. As part of content moderation practices that are implemented in AI systems, the proposed AI Act requires the manufacturer and the operator of the AI systems to meet various regulatory standards, but it does not provide an additional layer of procedural guarantees to end users, such as a personal right to obtain all the information used by the AI system in its decision-making process, a right to an explanation of the decision, and a right to challenge the decision before an objective body.

**CONCLUDING REMARKS**

The digital sphere has become the backbone of civic life in modern societies, and is operated entirely by the private sector, i.e., by technology companies, some of which have grown to unprecedented size. Therefore, the private sector, and in particular the gigantic companies operating the online platforms, are the rulers of the online digital sphere. One of the greatest challenges of our time is to constrain the power of the online rulers. Constraining the power of the state over individuals is at the heart of public law, which adheres to both substantive fundamental rights and to procedural justice rights, manifested as the basic principles of administrative law. Because today civic life is conducted in the digital sphere, the rationale underlying human rights protection, together with procedural justice guarantees, should be replicated to the new digital arena. The more extensively procedural justice rights are imported into the digital sphere, the greater the adherence to substantive human rights and the rule of law will be accomplished.

Content moderation practices serve as a case study for designing a digital governance regime that aspires to protect human rights. Although the reasons for these practices have merit, as, for example, removing illegal or harmful content, they may still conflict with freedom of speech. The online platforms that use these practices become the gatekeepers of digital free speech on a global scale. The various legal initiatives introduced in the EU are aimed at imposing the general obligation of accountability on the online platforms, especially when AI technologies are deployed in content moderation, amplifying existing problems of digital free speech. Yet these initiatives lack a basic element essential to providing full-fledged protection of human rights, that is, ⚫.

---

Regulating online content moderation

procedural justice rights granted to individuals. Imposing various obligations on the operators of the online platform, which generates a duty of compliance, overseen by the regulator, is not enough to create a trustworthy digital environment. The way to build trust is through personal due process rights enforced by the injured individuals.