Digital platforms
Under the scope of the New Era in Digital Services

CONTENT MODERATION BY AI SYSTEMS – FROM THE PRODUCT LIABILITY MODEL TO INDIVIDUAL RIGHTS

Orit Fischman-Afori*

I. Introduction
II. Content Moderation in the Age of AI
A. Content Moderation and Freedom of Speech
B. Content Moderation by AI Systems
C. EU Regulations
III. The Rise of "Digital Procedural Justice" (or "Digital Due Process") Approach
IV. Regulating Content Moderation by AI Systems – From the Product Liability Model to Individual Rights
A. Giving Understandable Reasons for Decisions
B. Objective Review Involving Human Discretion
C. Hearings to Affected Individuals
V. Concluding Remarks

I. Introduction

The implications for democracies of the massive flow of content in the digital environment is at present at the heart of a fierce public debate.[1] Although the wide dissemination of content may nurture the "information society" of the 21st century and freedom of speech, it may also encourage the spread of harmful content, such as hate speech and false information, which undermines the foundations of civil society and democratic values.[2] The digital environment is operated and controlled by online

* Professor of Law, Faculty of Law, College of Management, Israel.

1 Some of the ideas expressed in this article are discussed in my previous studies; *see*: Orit Fischman Afori, *Universal Measures in the Service of Global Challenges: Proportionality, Blocking Orders, and Online Intermediaries as Hybrid Bodies*, *in* PLURALISM OR UNIVERSALISM IN INTERNATIONAL COPYRIGHT LAW (Tatiana Eleni Synodinou ed., Kluwer Law, 2020); Orit Fischman Afori, *Online Rulers as Hybrid Bodies: The Case of Infringing Content Monitoring*, 23 UNIVERSITY OF PENNSYLVANIA JOURNAL OF CONSTITUTIONAL LAW, 121 (2021); Orit Fischman Afori, *Taking Global Administrative Law One Step Ahead: Online Giants and the Digital Democratic Sphere*, INTERNATIONAL JOURNAL OF CONSTITUTIONAL LAW (ICON), (2022) http://ssrn.com/abstract=3874915; Orit Fischman Afori, *Global Digital Governance Through the Back Door of Corporate Regulation*, 33 *FORDHAM INTELLECTUAL PROPERTY, MEDIA & ENTERTAINMENT LAW JOURNAL*, 1 (2023); Orit Fischman Afori, *Regulating Online Content Moderation – Taking Stock and Moving Ahead with Procedural Justice and Due Process Rights*, *in THE EXPLOITATION OF IPR – FINDING THE RIGHT BALANCE* (Jens H. Schovsbo ed., Edward Elgar, 2023).

2 For a recent report published by the OECD on the harms caused by the spread of disinformation, *see*: ORG. FOR ECON. CO-OPERATION & DEV. FACTS NOT FAKES: TACKLING DISINFORMATION, STRENGTHENING INFORMATION INTEGRITY (2024), https://www.oecd.org/publications/facts-not-fakes-tackling-disinformation-strengthening-information-integrity-d909ff7a-en.htm. *See also*: Leslie Gielow Jacobs, *Freedom of Speech and*

platforms—private commercial entities like Google (the dominant search engine), YouTube (the dominant video-sharing service), and Facebook (the dominant social media network).[3] The online platforms are the backbone of the digital environment. They have the technical ability to control the flow of content in their pipelines, therefore can function as gatekeepers and resolve the conflict between free speech and the protection of other democratic interests, limiting access to harmful content. Online platforms conduct various types of content moderation, including monitoring, filtering, tagging, and removing content. Some of these practices are imposed on them by coercive regulation; others are the result of voluntary initiatives. In either case, online platforms enjoy extensive leeway in carrying out their content moderation policies. Thus, the task of striking a balance between conflicting interests touching on the core of democracy and limiting freedom of speech in appropriate cases has devolved to online platforms, which are private commercial entities.[4]

This reality, in which the online platforms control the channels of public discourse and with it the sphere of democratic civil society has led to increasing calls for greater guarantees for the protection of individual rights in the digital environment. Increasingly, voices advocating a "digital constitutionalism" have been heard both in academia and among policymakers.[5] One of the sub-themes in this new human rights discourse is the need to maintain procedural justice principles, known also as "due process" rights, in the operation of the online platforms in general and in content moderation systems in particular.[6] It has been argued that content moderation has a profound effect on freedom of speech, therefore it should be handled by procedures that guarantee the protection of individual rights.[7]

The controversies stemming from content moderation practices have intensified in the last few years, with the growing use of a family of technologies, collectively referred to as artificial intelligence (AI), in the implementation of these practices. AI technology provides effective automated decision making that replaces human labor, therefore it can streamline the handling of the massive traffic of content on online platforms. At the same time, the deployment of AI technologies in content moderation systems creates a serious challenge in addressing the conflict between competing interests and ensuring fair procedures.

Regulation of Fake News, 70 THE AMERICAN JOURNAL OF COMPARATIVE LAW, 278 (2022); W. Lance Bennett & Steven Livingston, *The Disinformation Order: Disruptive Communication And The Decline Of Democratic Institutions*, 33 EUROPEAN JOURNAL OF COMMUNICATION (2018).

3 ORG. FOR ECON. CO-OPERATION & DEV., AN INTRODUCTION TO ONLINE PLATFORMS AND THEIR ROLE IN THE DIGITAL TRANSFORMATION (2019), https://doi.org/10.1787/53e5f593-en.

4 *See infra* Part II.

5 *See e.g.* Giovanni De Gregorio, *The Rise Of Digital Constitutionalism In The European Union*, 19 INTERNATIONAL JOURNAL OF CONSTITUTIONAL LAW (ICON), 41 (2021); Edoardo Celeste, *Digital Constitutionalism: A New Systematic Theorisation*, 33 INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY, 76 (2019).

6 *See e.g.* Giulia Gentile, *Between Online and Offline Due Process: The Digital Services Act*, available at: https://ssrn.com/abstract=4550655; Giancarlo Frosio & Christophe Geiger, *Taking Fundamental Rights Seriously In The Digital Services Act's Platform Liability Regime*, 29 EUROPEAN LAW JOURNAL, 31 (2023) .

7 *See infra* Part III.

The most prominent legislative initiatives aimed at meeting the challenges of the digital environment and the use of AI technologies have been in the EU. The EU has initiated a line of legislative acts that together amount to a body of digital governance norms for directly or indirectly regulating content moderation practices. Two landmark initiatives are the Digital Services Act (DSA), which was adopted in 2022 and came into force in 2024,[8] and the Artificial Intelligence Act (AI Act),[9] approved by the EU parliament in March 2024.[10] DSA regulates the various online services for creating a safe digital environment; the AI Act regulates the use of AI technologies online and offline. The two legal frameworks provide some guidance in the operation of content moderation practices, particularly in the management of AI systems. The DSA requires implementing certain fair procedures in operating online services, marking an important move in promoting "digital due process." The AI Act reflects a legal regime more akin to a product liability framework, therefore it does not address directly issues such as individual rights in the digital environment.

The present article takes content moderation practices, particularly when operated by AI systems, one step forward toward better protection of individual rights. To that end, we analyzed some basic principles in procedural justice theory that have been incorporated in several key rules in administrative laws worldwide. We argue that content moderation mechanisms should apply these rules, which function as building blocks of fair procedures in the organizational decision-making process. Adherence to such procedures can protect individual rights and lead to better handling of the complex decisions on limiting access to content. The three main procedural aspects are discussed: (a) providing understandable reasons for decisions made; (b) setting objective review mechanisms based on unfettered human discretion applied case by case; and (c) granting individuals affected by decisions an opportunity to be heard in the decision-making process and to provide relevant information. Meeting these standards can protect individual rights, promote better decisions on limiting access to content, and mitigate the potentially chilling effect associated with content moderation practices. Meeting such standards may also foster trust and legitimacy in content moderation mechanisms as warranted frameworks aimed at serving democratic goals in handling harmful speech. The various principles of fair procedure, comprising together a "good" decision-making process that promotes justice and fairness for the public good, should be implemented in the operational

8 https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en.

9 Proposal For A Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, COM/2021/206 Final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

10 The European Parliament confirmed the proposed AI Act in a plenary vote on June 14, 2023, but negotiations regarding the outstanding elements of the proposal continued until the final approval of the Act on March 13, 2024; *see* Brussels, 26 January 2024 (OR. en) 5662/24, Interinstitutional File: 2021/0106(COD); European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

systems of the online platforms, which are rightfully perceived as gatekeepers capable of promoting the democratization of the digital environment.

This article proceeds as follows: Part II reviews the practice of content moderation and explains its conflict with freedom of speech; it also describes the use of AI technologies in these practices and its amplified effect on freedom of speech and reviews the main EU regulations addressing content moderation practices. Part III describes the rise of the digital procedural justice approach. Part IV proposes inserting certain key procedural standards into content moderation practices to better protect individual rights. Part V contains concluding remarks.

II. Content Moderation in the Age of AI

A. Content Moderation and Freedom of Speech

Online platforms serve as a central arena for public discourse. Massive amounts of information, data, and content are conveyed by a range of means, including text, images, and audio. The hyperdynamic digital environment has produced a digital information society, in which freedom of speech flourishes.[11] This thriving environment is supported by a regulation that provides immunity to online platforms from the 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 U.S. Communications Decency Act (CDA), adopted in 1996,[12] and the EU e-Commerce Directive, adopted in 2000.[13]

The massive flow of content on online platforms has raised serious challenges because it may include harmful 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 others' rights, such as copyright and privacy.[14] Harmful digital speech includes many forms and types of content, each raising different social and legal concerns. Despite the safe harbor enjoyed by online platforms, they have undertaken to voluntarily play an active role in removing offensive content.[15] Various interests, such as business credibility and social legitimacy in consumers' communities, serve as incentives for the development of

---

11 *See supra* note 3.

12 47 U.S.C. § 230 (2012).

13 Directive 2000/31/EC of the European Parliament and of the Council of June 8, 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) [hereinafter, EU e-Commerce Directive].

14 *See* OECD, FACTS NOT FAKES: TACKLING DISINFORMATION, STRENGTHENING INFORMATION INTEGRITY (2024), *supra* note 2. *See also* Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 31 (2018–2019); Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, (2018); Richard Ashby Wilson & Molly K. Land, *Hate Speech on Social Media: Towards a Context-Specific Content Moderation Policy*, 52 CONN. L.R. (2021).

15 47 U.S.C. § 230 (2012).

voluntary content moderation practices.[16] Therefore, the online platforms have gradually adopted a range of practices concerning monitoring, filtering, tagging, and blocking various types of harmful content. [17] Together, these practices are referred to as "content moderation."

The various voluntary content moderation practices usually address the type of content known as "awful but lawful," indicating that it does not refer to technically illegal content.[18] The only non-voluntary enactment that obligates online platforms to moderate harmful but legal content is the German Hate Speech Act (NetzDG) of 2017, which imposes an obligation to remove hate speech and provide a transparent decision-making process for it.[19] All other examples of legal frameworks that adopt mandatory content moderation measures address illegal content. For example, Article 230 of 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).[20] Section 512 of the U.S. Copyright Act creates a safe harbor for some online intermediaries in case of content that infringes without their knowledge,[21] and it establishes a notice-and-takedown mechanism for all Internet intermediaries that governs the monitoring of copyrighted content. Two decades later, in 2019, the EU also adopted a framework for the handling of copyrighted content in Article 17 of the Directive on Copyright in the Digital Single Market (DCDSM), which imposes content moderation obligations for copyright infringement.[22] Moreover, Section 230(e) of the CDA, adopted in 2018, defines another exception to the immunity granted to service providers regarding certain sex trafficking offenses.[23] Similarly, the Online Safety Act, adopted by the UK Parliament in October 2023, imposes content moderation obligations concerning highly harmful *illegal* content, such as materials related to terrorism and child sexual abuse.[24]

---

16 Aswad, *supra* note 14, at 42*;* Keats Citron, *supra* note 14, at 1047. *See also* Věra Jourova, Code of Conduct on Countering Illegal Hate Speech Online: First Results on Implementation, Eur. Comm'n (Dec. 2016), http://ec.europa.eu/information_society/newsroom/image/document/2016-50/factsheet-code-conduct-8_40573.pdf.

17 Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money* 18 (Hoover Institution, Aegis Series Paper No. 1807, 2018), www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf.

18 Daphne Keller, *Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users*, U CHI. L. REV. ONLINE (2022); Raghav Ahooja, *Section 230 and the Fediverse: The 'Instances' of Mastodon's Immunity and Liability* (April 17, 2023). Available at SSRN: https://ssrn.com/abstract=4421665

19 Netzwerkdurchsetzungsgesetz [BGBl. I p. 3352] [NetzDG] [Act to Improve Enforcement of the Law in Social Networks], art. 1(1), translation at www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf [hereinafter Hate Speech Act].

20 17 U.S.C. § 512 (1988).

21 *See* Viacom Int'l, Inc. v. YouTube, Inc., 676 F.3d 19, 30 (2d Cir. 2012).

22 Directive (EU) 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, 2019 O.J. (L 130).

23 47 U.S. Code § 230; Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115-164 (2018).

24 UK Online Safety Act, 2023, https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted.

An extensively discussed outcome of the notice-and-takedown regime created by §512 of the US Copyright Act is the easy mass removal of allegedly infringing copyrighted content, which has had a significant chilling effect on freedom of speech.[25] Because intermediaries are risk-averse, they have an incentive to respond positively to all take-down requests, even if such requests would have been found unjustified had they been decided in court.[26] As an uninvolved third party in the dispute, the online platforms have no incentive to invest time and effort in profound legal assessment of the requests, and the outcome is massive and indiscriminate removal of content.[27] Likewise, the strong opposition to the mandatory content moderation framework adopted by Article 17 of the DCDSM is based on the fear that it will have a significant chilling effect to freedom of speech.[28] A particular fear was that the systems designed to perform such content moderation may produce false positive outcomes, which are inconsistent with freedom of speech as the default principle.[29] Because of such concerns, in June 2021, the European Commission published its guidelines on the application of Article 17 of the DCDSM, stressing the need to safeguard freedom of speech and other fundamental rights affected by Article 17.[30] Nevertheless, Poland has brought a case against Article 17 to the Court of Justice of the EU (CJEU), claiming that it should be annulled because it is not consistent with the principle of freedom of expression.[31] In April 2022, the CJEU rejected the Polish claim and determined that Article 17 contains appropriate safeguards to ensure a fair balance between copyright protection and freedom of

25 *See* Jerome H. Reichman, Graeme B. Dinwoodie, & Pamela Samuelson, *A Reverse Notice and Takedown Regime to Enable Public Interest Uses of Technically Protected Copyrighted Works*, 22 BERKELEY TECH. L.J. 981 (2007); Jennifer M. Urban, Joe Karaganis, & Brianna Schofield, *Notice and Takedown: Online Service Provider and Rightsholder Accounts of Everyday Practice*, 64 J. COPYRIGHT SOC'Y U.S.A. 371 (2017)); Daniel Etcovitch, *DMCA § 512 Pain Points: Music and Technology Industry Perspectives in Juxtaposition*, 30 HARV. J.L. & TECH. 547 (2017); DAPHNE KELLER, INTERNET PLATFORMS: OBSERVATIONS ON SPEECH, DANGER, AND MONEY 18 (Hoover Institution, Aegis Series Paper No. 1807, 2018); Niva Elkin-Koren & Maayan Perel, *Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law*, *in* OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY (2020).

26 *See* Jack Balkin, *Old School/New School Speech Regulation*, 127 HARV. L. REV. 2296, 2314 (2016).

27 *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).

28 *See, e.g.*, Martin Senftleben *et al*, *The Recommendation on Measures to Safeguard Fundamental Rights and the Open Internet in the Framework of the EU Copyright Reform*, 40 EUR. INTELL. PROP. REV. 149, 187 (2018).

29 Toni Lester & Dessislava Pachamanova, *The Dilemma of False Positives: Making Content ID Algorithms More Conducive to Fostering Innovative Fair Use in Music Creation*, 24 UCLA ENT. L. REV. 51, 53 (2017); Sebastian Felix Schwemer et al, *Impact of Content Moderation Practices and Technologies on Access and Diversity* (January 31, 2023), available at SSRN: https://ssrn.com/abstract=4380345; Daria Dergacheva & Christian Katzenbach, *Mandate to Overblock? Understanding The Impact of The European Union's Article 17 On Copyright Content Moderation On Youtube*, POLICY & INTERNET (2023).

30 Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market, COM 2021, 288, June 4 2021, at p. 19, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1625142238402&uri=CELEX%3A52021DC0288.

31 In CJEU Case C-401/19 – Republic of Poland v European Parliament, Council of the European Union ECLI:EU:C:2021:613.

speech.[32] The CJEU noted, however, that when implementing Article 17 into their national law, 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."[33] The exact procedures concerning how to operate the frameworks created by Article 17 are not determined in the CJEU ruling, and it is left to member states to decide what measures should be applied to prevent conflicts with fundamental rights.

All content moderation practices, whether they are part of voluntary initiatives or the result of a regulatory obligation, raise questions of conflict with freedom of speech because they may wrongfully prevent access to content that, if assessed by a court or an objective reviewing body, could have been found as legal or legitimate.[34] The fear of over-blocking and false-positive take-downs has a chilling effect on freedom of speech, and it exists regarding all types of content moderation, whatever its underlying motivation might be.[35] All content moderation practices therefore raise a significant challenge to democracies because the activities of the online platforms have an extensive effect on individuals' fundamental rights and freedoms.[36] 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.[37] Another report, issued in 2018 by the UN Special Rapporteur, was based on a global survey and sought to collect empirical data about 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.[38] The question is, therefore, whether and how the power in the hands of the private online platforms that control the digital environment should be regulated, and whether content moderation practices should be restrained to protect users' fundamental rights, first and foremost, freedom of speech.[39]

B. Content Moderation by AI Systems

The controversies stemming from content moderation practices have changed in the last few years with the growing use of a family of technologies, collectively

32 https://curia.europa.eu/juris/document/document.jsf?text=&docid=258261&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=10745892, at par. 98.

33 *Id.* at par. 99.

34 Orit Fischman Afori, *Online Rulers as Hybrid Bodies: The Case of Infringing Content Monitoring*, 23 UNIVERSITY OF PENNSYLVANIA JOURNAL OF CONSTITUTIONAL LAW, 121 (2021); David Kaye, Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. OL OTH 41/2018, 3–8 (June 13, 2018).

35 Lester & Pachamanova, *supra* note 29, at 67–72.

36 NICOLAS SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES, (2019).

37 Special Rapporteur on the Promotion and Protection of Freedom of Opinion and Expression, A/HRC/17/27 (May 16, 2011).

38 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).

39 These questions are discussed at length in some of my studies, *see supra* note 1.

referred to as AI, that train algorithms to produce various outputs, including content, predictions, recommendations, and decisions.[40] A particular AI technology, known as machine learning, trains algorithms to run on constantly updated large datasets and detect patterns used to autonomously generate outputs such as observations and decisions.[41] Machine learning is a data-driven technology that adapts its performance to the inputs it receives.[42] The end goal of these technologies is "to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly."[43] AI has been used by businesses for a range of tasks, from offering services replacing the human workforce to the operation of content moderation systems by online platforms.[44]

AI technology provides an effective decision-making system that replaces human labor.[45] Human handling of the huge amount of content flowing through online platforms appears to be impracticable. The new role of AI grants algorithms discretion about the handling of vast amounts of content uploaded by users. The use of AI systems for content moderation has exacerbated earlier problems related to the control that the platforms exercise over the flow of content in the digital environment.[46] Since the policies of online platforms regarding content moderation have been translated into algorithms and default settings, the silencing mechanism has become algorithmic.[47] At present, automated, computer-controlled devices pose a substantial threat to democratic values and freedom of speech, affecting society as a whole. For example, an AI decision to block a post uploaded by a politician—or not to block it—affects not only that individual's freedom of speech but democratic discourse as a whole.[48] The Content ID system implemented by YouTube, which automatically scans

---

40 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).

41 Stanley Greenstein, *Preserving the Rule of Law In The Era Of Artificial Intelligence*, ARTIFI. INTELL. L. 1, 9 (Springer, July 2021), https://doi.org/10.1007/s10506-021-09294-4.

42 Greenstein, *supra* note 41, at 9.

43 Greenstein, *supra* note 41, at 10.

44 *See* SELECT COMM. ON ARTIFICIAL INTELLIGENCE, AI IN THE UK: READY, WILLING AND ABLE?, 2017–19, HL 100, at 2 (U.K.), publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (UK AI Report), at p. 25.

45 ORG. FOR ECON. CO-OP. & DEV., DATA-DRIVEN INNOVATION FOR GROWTH AND WELL-BEING: INTERIM SYNTHESIS REPORT 32 (2014), http://www.oecd.org/sti/inno/data-driven-innovation-interim-synthesis.pdf.

46 Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C.D. L. REV. 1149, 1162 (2018).

47 Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, SOC. SCI. RES. NETWORK (Feb. 14, 2020), https://ssrn.com/abstract=3439261.

48 Former US President Trump found himself at the center of the debate regarding politicians' digital speech in May 2020, when Twitter (today "X") declared that Trump's online messages were potentially false or glorifying violence. *See Permanent Suspension of @realDonaldTrump*,

and identifies copyrighted content in user-uploaded videos, has a similar effect.[49] Under the Content ID system, copyright owners can choose to monetize, track, or take down infringing content.[50] Thus, to a large extent, copyright enforcement is operated today by algorithms[51] that significantly control the flow of content and information in society.[52]

The combination of the two core characteristics of the online speech environment, that it is operated by online platforms owned by private businesses and that content moderation is carried out by algorithms, creates a challenge to regulators.[53] The algorithms that control participation in various social arenas use a language and logic that are not understood by ordinary humans.[54] This is part of the known "black box" problem of AI technologies.[55] A similar trend exists in the public

---

TWITTER BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html [https://perma.cc/ADH3-RLL2]. Former President Trump perceived this move as hindering freedom of speech. *See* Maggie Haberman & Kate Conger, *Trump Signs Executive Order on Social Media, Claiming to Protect 'Free Speech'*, N.Y. TIMES (June 2, 2020), https://www.nytimes.com/2020/05/28/us/politics/trump-order-social-media.html [https://perma.cc/W22L-SBMT]. As a result of massive public pressure, Facebook followed Twitter's move and adopted a new proactive policy monitoring speech that encourages violence, resulting in the suspension of Trump's account for two years. *See* Mike Isaac & Sheera Frenkel, *Facebook Says Trump's Ban Will Last at Least 2 Years,* N.Y. TIMES (June 7, 2021), https://www.nytimes.com/2021/06/04/technology/facebook-trump-ban.html [https://perma.cc/HRS8-57HE]. The Facebook oversight board inspected the case and upheld the decision to suspend Trump's account but found that a proper penalty, which should have been suspension for a limited time, failed to be imposed, *see* https://www.oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty/.

49 *See* https://support.google.com/youtube/answer/2797370?hl=en (last visited January 21, 2024).

50 *Id.*

51 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).

52 For the influence of general content moderation practices on freedom of expression, *see* Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, (2019); Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 528, (2022).

53 Katyal, *supra* note 40, at 107–08.

54 *See generally* Martin Ebers, *Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s), in* NORDIC YEARBOOK OF LAW AND INFORMATICS 2020: LAW IN THE ERA OF ARTIFICIAL INTELLIGENCE (Liane Colonna & Stanley Greenstein eds., 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3901732 [https://perma.cc/C68A-T898]; Greenstein, *supra* note 41, at 18 ("The rule of law up until now has been dependent on its form being in the format of natural language—it entails governance by natural language as opposed to the governance of the algorithm. The rule of law is dependent on natural language in order to be comprehended.").

55 Greenstein, *supra* note 41, at 292 ("The threat to the rule of law lies in the fact that most of these decision-making systems are 'black boxes' because they incorporate extremely complex technology that is essentially beyond the cognitive capacities of humans and the law too inhibits transparency to a certain degree."); *see generally* Sylvia Lu, *Data Privacy, Human Rights, and Algorithmic Opacity*, 110 Cal. L.R. 2087 (2022); Jennifer Cobbe & Jatinder Singh, *Reviewable Automated Decision-Making*, 39 COMPUTER LAW AND SECURITY REVIEW 105475, 105478 (2020); Seth Katsuya Endo, *Technological Opacity & Procedural Injustice*, 59 BOSTON COLLEGE LAW REVIEW, 821 (2018); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

sector, where decisions made by public agencies are increasingly generated by computational systems.[56] But in the public sector, there have been significant developments regarding the legal framework that should govern the use of AI technologies to ensure adequate administrative procedural standards and public law principles of accountability.[57] The emerging reality in the business sector calls for legislatures, policymakers, and civil society stakeholders to address this challenge as well. The global digital environment, which is run by the business sector, needs a comprehensive and effective regime of digital governance to ensure that the rule of algorithms is consistent with the rule of law worldwide.[58]

C. EU Regulations

As noted, the most prominent legislative initiatives aimed at regulating the digital environment and the use of AI technologies have been in the EU. The EU initiated a line of legislative efforts that together, directly and indirectly, are generating a body of digital governance norms for regulating content moderation practices, the two landmark initiatives in this regard being the DSA and the AI Act.

The DSA aims to establish a comprehensive and systematic regulation of various online services and introduce safeguards for fundamental rights in the online environment. This initiative is the first important EU legislation for the digital sector since the e-Commerce Directive of 2000.[59] The new regime serves as a key element in building digital governance principles, with a potential worldwide effect.[60] The DSA is based on the principle of tailoring the obligations of service providers to the nature of the service and the size of the provider.[61] The proposal differentiates between various digital service providers. Some general obligations apply to most services, but

---

56 Paul Daly, *Artificial Administration: Administrative Law, Administrative Justice and Accountability in the Age of Machines*, *Forthcoming* AUSTRALIAN JOURNAL OF ADMINISTRATIVE LAW & PRACTICE (2023), Available at SSRN: https://ssrn.com/abstract=4434238; Monika Zalnieriute et al., *The Rule of Law and Automated Government Decision-Making*, 82 MODERN L. REV. 425, 444 (2019); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1205–13 (2017); Van Loo, *supra* note 102, at 1321.

57 Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 657–60 (2017); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 4–14 (2019); JOE TOMLINSON, JUSTICE IN THE DIGITAL STATE: ASSESSING THE NEXT REVOLUTION IN ADMINISTRATIVE JUSTICE, 69-70 (2019); Responsible Use of Artificial Intelligence (AI), GOV. OF CANADA (Nov. 1, 2022), https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html [https://perma.cc/FG9X-UX5Q]; *Digital Nations Charter*, GOV. OF CANADA (Nov. 18, 2021), https://www.canada.ca/en/government/system/digital-government/improving-digital-services/digital9charter.html [https://perma.cc/NJJ5-5WNU].

58 *See e.g.* LUCIANO FLORIDI, THE FOURTH REVOLUTION: HOW THE INFOSPHERE IS RESHAPING HUMAN REALITY (Oxford Univ. Press, 2014); Luciano Floridi et al., *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 MINDS & MACHINES 689, 694 (2018), https://link.springer.com/article/10.1007%2Fs11023-018-9482-5 [https://perma.cc/A796-H3FG].

59 Proposal for a Regulation Of The European Parliament And Of The Council on a Single Market For Digital Services (DSA) and amending Directive 2000/31/EC https://eur-lex.europa.eu/legal-ontent/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en, at 1.

60 https://www.eff.org/deeplinks/2020/12/eu-and-digital-services-act-year-review.

61 DSA, at 6.

stricter obligations are imposed on online platforms, which are defined narrowly to include services such as those provided by social networks (*e.g.*, Facebook) and content storage and dissemination platforms (*e.g.*, YouTube).[62] As clarified at the outset of the DSA proposal, this regulation does 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."[63] Furthermore, no general obligation exists for online platforms to monitor information or a duty to actively search for illegal content.[64] Nevertheless, the DSA significantly expands various procedural obligations of online service providers. Although no monitoring obligations are imposed on large platforms, serving 10% or more of the EU population, they are subject to some additional obligations, such as risk assessment concerning the traffic of illegal content and the negative effects of content moderation on freedom of speech.[65] Consistent with the general policy of imposing only general accountability duties, large online platforms are mandated to "put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified."[66] 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," including 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 the 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 and imposes certain transparency obligations[67] but does not require human determination in the process. Nevertheless, the platforms are obligated to disclose the use of automated tools and algorithmic decision-making processes in content moderation.[68] Moreover, the DSA 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."[69] The DSA also requires that service providers periodically publish transparency reports.[70] Regarding the reasoning of decisions to remove content, service providers are required to inform the recipient, *that is*, 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."[71] The

62 DSA, Article 2 (Definitions).
63 DSA, at 4.
64 DSA, Article 7.
65 DSA, Article 25, 26.
66 DSA, Article 27.
67 DSA, at 33-4.
68 DSA, Articles 14(1), 16 (6), 17 (3).
69 DSA, Article 12.
70 DSA, Articles 13, 23. Very large online platforms are subject to additional periodic transparency report obligations, *see* DSA, Article 33.
71 DSA, Article 15.

DSA specifies the content to be included in the notification, ensuring its substantive nature. Regarding the possibility of challenging the decision, it stipulates that the service provider must supply information about the available options of either "internal complaint-handling mechanisms, out-of-court dispute settlement and judicial redress."[72] 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.[73] The DSA also proposes 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."[74] Finally, a key requirement of the DSA, aimed at countering the over-blocking of content that hinders freedom of speech, concerns a "put back" obligation.[75] 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.[76]

In April 2021, the EU Commission presented its proposal for an AI Act,[77] aimed at guaranteeing that the deployment of AI technologies conforms to "Union values, fundamental rights and principles."[78] The AI Act was finally approved in March 2024.[79] This legislation covers a wide array of topics, from the prohibition of the use of various types of AI systems[80] to the imposition of operational requirements according to the type and category of the AI system.

AI systems falling in the category of "high risk" are subject to a strict standard of requirements,[81] including implementation of a risk management system[82] and data management governance standards.[83] 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,"[84] or to provide the operator of the AI system with appropriate instructions.[85] The AI Act does not grant the end user, *meaning* the non-professional user, any individual rights, and its focus is on regulating either the producer or the operator of systems. The only obligation relating to end

---

72 DSA, Article 18 (1).
73 DSA, Articles 17, 18 (1).
74 DSA, Article 18.
75 DSA, Article 17 (3).
76 Alexander Peukert, *Five Reasons to be Skeptical About the DSA*, *in* TO BREAK UP OR REGULATE BIG TECH? AVENUES TO CONSTRAIN PRIVATE POWER IN THE DSA/DMA PACKAGE, 22, 24 (eds. Heiko Richter, Marlene Straub & Erik Tuchtfeld, Max Planck Max Planck Institute for Innovation and Competition, 2021).
77 Proposal For A Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, COM/2021/206 Final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. (AI Act)]
78 AI Act, at p. 1.
79 *See supra* note 10.
80 AI Act, Article 5.
81 AI Act, Articles 7, 8.
82 AI Act, Article 9.
83 AI Act, Article 10.
84 AI Act, Article 13 (1).
85 AI Act, Article 13 (2).

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 is disclosed up front.[86]

The 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 online platforms has become pervasive, the AI Act serves as a complementary measure in building the online digital governance regime, in particular, content moderation practices implemented by AI systems. 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 aspects of content moderation systems would serve as an additional layer, on top of the DSA provisions. But whereas the AI Act is focused on the producers' and operators' obligation to comply with adequate standards, following a model of "product liability" legislation,[87] the DSA reflects a more consumer-centric approach that also addresses the conduct of the online platforms *vis-à-vis* end users.

III. The Rise of Digital Procedural Justice (or the Digital Due Process Approach)

Content moderation practices pose a serious challenge to freedom of speech and consequently to democracies worldwide. Together with other practices deployed in the digital environment, this challenge has given rise to the digital constitutionalism approach, which stresses the importance of observing the rule of law and fundamental rights in the digital realm.[88] At the same time, some adjustments of fundamental rights to the new digital environment may be necessary.

The new digital governance regime designed by the recent EU initiatives seeks to safeguard human rights and observe the rule of law in the digital domain. Yet, a main criticism of the new regime is the absence of measures aimed at protecting *individuals* from the power of the operators of online platforms, and the failure to provide adequate *procedural* guarantees for human rights.[89] 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

---

86 AI Act, Article 52.

87 *See* Nathalie Smuha et al., *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act*, 1, 48–54 (Aug. 5, 2021), https://ssrn.com/abstract=3899991 [https://perma.cc/9DVA-X8X5].

88 *See e.g.* S. Amato, *Artificial Intelligence and Constitutional Values*, *in* ENCYCLOPEDIA OF CONTEMPORARY CONSTITUTIONALISM 10–12 (J. Cremades & C. Hermida eds., Springer 2021), https://link.springer.com/referencework/10.1007/978-3-319-31739-7 [https://perma.cc/5C5C-KX3W]; De Gregorio, *supra* note 5; Celeste, *supra* note 5; Gentile, *supra* note 6; Frosio & Geiger, *supra* note 6.

89 For a similar stance, *see* SUZOR, *supra* note 36, at 144-145; 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).

data.[90] Yet, no similar measure was enacted regarding content moderation practices in the DSA nor under the AI Act. Consequently, there are no adequate procedural protective measures for the personal rights of end users who may be injured by content removal. Thus, the EU digital governance regime pertaining to content moderation practices lacks *procedural justice rights* that can provide adequate guarantees for individuals' rights to free speech.[91]

Procedural justice measures, referred to in US terminology as *procedural due process*, have emerged 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 and its agencies.[92] Administrative law seeks to accomplish some basic underlying goals, headed by legality, rationality, and fairness in the administrative decision-making process.[93] Legality, rationality, and fairness are perceived as fundamental values in democratic societies.[94] Administrative law has evolved over the years into a distinct discipline of law, along with its internal notion of "administrative justice."[95] This particular strain of justice is reflected in concrete obligations concerning the procedures of decision making, that is, the "justice inherent in routine day-to-day administration."[96]

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 standards of transparency, providing reasons for their decisions, that their decisions should be subject to objective external review, and that the affected individuals should have an opportunity to be heard.[97] These are key procedural rights, providing guarantees that the decision-making process of public

---

90 Regulation (Eu) 2016/679 Of the European Parliament and Of the Council of 27 April 2016 On The Protection Of Natural Persons With Regard To The Processing Of Personal Data And On The Free Movement Of Such Data, And Repealing Directive 95/46/EC (General Data Protection Regulation) (GDPR), Chapter III – Rights of the Data Subjects.

91 *See* Orit Fischman-Afori, *supra* note 1; Gentile, *supra* note 6; Frosio & Geiger, *supra* note 6.

92 Larry Alexander, *The Relationship Between Procedural Due Process and Substantive Constitutional Rights*, 39 U. FLA. L. REV. 323, 325 (1987). For a comprehensive analysis of the underlying rationale of procedural justice principles, *see* Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181 (2004).

93 PAUL DALY, UNDERSTANDING ADMINISTRATIVE LAW IN THE COMMON LAW WORLD, 6-7 (Oxford University Press, 2021).

94 *Id*.

95 Michael Adler, *Understanding and Analysing Administrative Justice*, *in* ADMINISTRATIVE JUSTICE IN CONTEXT, 129 (Michael Adler, ed. Hart, 2010); JOE TOMLINSON, JUSTICE IN THE DIGITAL STATE: ASSESSING THE NEXT REVOLUTION IN ADMINISTRATIVE JUSTICE, 1 (2019).

96 Michael Adler, *A Socio-Legal Approach to Administrative Justice*, 25 LAW & POL'Y 323, 329 (2003) (referring to JERRY L. MASHAW, BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY DISABILITY CLAIMS, 24 (YALE U. PRESS 1983).

97 STEPHEN BREYER, ADMINISTRATIVE LAW & REGULATORY POLICY (3rd ed. 1992); U.S. Administrative Procedure Act 5 U.S.C. §§ 551-559 (1946); Regulatory Accountability Act of 2017, S. 951, 115th Cong. (2017). Cary Coglianese, *Administrative Law: The U.S. and Beyond*, *in* INTERNATIONAL ENCYCLOPEDIA OF SOCIAL & BEHAVIORAL SCIENCES, (James D. Wright, ed., 2d ed. 2015). *See also* Christopher J. Walker, *Modernizing the Administrative Procedure Act*, 69 ADMIN. L. REV. 629 (2017) (describing the evolution of the US Administrative Procedure Act).

authorities is "accountable," that is, conforms with the rule of law.[98] These procedural justice measures also aim to promote trust in public authority and support the legitimacy of the public sector.[99]

The call advocating for digital constitutionalism, that is, the insertion of constitutional guarantees for individual rights in the digital realm, includes a sub-category about the adoption of digital procedural justice or digital due process safeguards into operational systems in the digital realm.[100] In particular, content moderation practices, which are based on AI technologies and challenge freedom of speech, require the adoption of procedural measures aimed at protecting individual rights and serving democratic goals.

IV. Regulating Content Moderation by AI Systems: From the Product Liability Model to Individual Rights

Given the rise of the digital procedural justice approach and the understanding that procedural guarantees for protecting individuals' rights should become part of the various content moderation practices, the question arises of what exact measures should be proposed. Three main issues, reflecting core principles of procedural justice, should be taken into account in meeting this challenge: (a) giving understandable reasons for decisions made; (b) setting objective review mechanisms involving human discretion; and (c) giving affected individuals an opportunity to be heard in the decision-making process and to provide relevant information.

A. Giving Understandable Reasons for Decisions

When a decision is of great importance to an individual and subject to a right to appeal, procedural justice principles support a duty to give reasons for the decision made.[101] The affected individual has a right to receive an explanation in order to *understand* the decision made.[102] The more important the decision is to the individual and more closely tailored to the case at hand (in contrast to a general decision), the more detailed the explanation should be.[103] Moreover, an appeal cannot be held unless the decision is explained.[104] The basis for the duty to give reasons is associated with

98 Kevin M. Stack, *An Administrative Jurisprudence: The Rule of Law in the Administrative State*, 115 COLUM. L. REV. 1985 (2015).

99 Matthew D. Adler, *Justification, Legitimacy, and Administrative Governance*, ISSUES IN LEGAL SCHOLARSHIP (2005).

100 Orit Fischman-Afori, *supra* note 1; Gentile, *supra* note 6; Frosio & Geiger, *supra* note 6.

101 Daly, *supra* note 93, at 20-21; Kendrick Lo, *When Efficiency Calls: Rethinking the Obligation to Provide Reasons for Administrative Decisions*, 43 QUEEN'S LAW JOURNAL 325, 326 (2018).

102 Paul Daly, Jennifer Raso & Joe Tomlinson, *Administrative Law and The Digital World*, Forthcoming *in* RESEARCH HANDBOOK ON ADMINISTRATIVE LAW (Carol Harlow ed., Edward Elgar, 2021), Available at SSRN: https://ssrn.com/abstract=4008531), at p. 5.

103 Daly, Raso & Tomlinson, *supra* note 102, at p. 6.

104 Daly, *supra* note 93, at 20-21; Lo, *supra* note 101, at 326.

fostering trust in the relevant authorities, endowing them with legitimacy, and supporting values concerned with individuals' wellbeing and autonomy.[105]

As explained above, the DSA offers significant measures promoting fairness in the operation of online platforms, including attempts to open the black box of algorithmic decision making with certain transparency obligations, including an obligation to notify end users about content moderation decisions and to include in such a notification its substantive elements.[106] Therefore, the DSA appears to adopt the basic principle of providing reasons for decisions made. In practice, however, it is often impossible to provide such explanations because these are merely an AI system output, not meaningful to humans.[107] AI systems are using mathematical and statistical models for generating their outputs, based on large datasets the system was trained on. Therefore, the output is not aimed at applying "discretion" regarding the particular circumstances of the case at hand, but rather to locate its proximity to past cases.[108] These impediments lead to the conclusion that the mere requirement of an explanation for a decision made is not enough to maintain fair procedures that would enable to truly protect individual rights,[109] therefore additional principles of procedural justice should be examined.

The AI Act exacerbates the failure to ensure individuals' right to procedural justice because it regulates the activities of the manufacturer and operator of high-risk AI systems but grants no rights to individual end users.[110] The transparency and explainability obligations are imposed on the manufacturers and operators of high-risk AI systems,[111] leaving the end user with only potential general law claims *vis-à-vis* the operators.[112] The standard of transparency and disclosure adopted by the AI Act was criticized for being more akin to the standard required by consumer protection

105 Daly, *supra* note 93, at 22; Keenan Molaskey, *Black Box Artificial Intelligence And The Rule Of Law: Note: Their Brothers' Keepers: Procedural Justice In The Intermediate Appellate Courts*, 84 LAW & CONTEMP. PROB. 81, 83-84 (2021).

106 *See supra* note 71.

107 Daly, Raso & Tomlinson, *supra* note 102, at p. 7; Jennifer Cobbe & Jatinder Singh, *Reviewable Automated Decision-Making*, 39 COMPUTER LAW AND SECURITY REVIEW 105475, 105478 (2020); Seth Katsuya Endo, *Technological Opacity & Procedural Injustice*, 59 BOSTON COLLEGE LAW REVIEW, 821 (2018).

108 Cobbe & Singh, *id;* Katsuya Endo, *id;* Stanley Greenstein, *Preserving The Rule of Law In The Era Of Artificial Intelligence (AI)*, 30 ARTIFICIAL INTELLIGENCE AND LAW, 291, 310 (2022). *See also in the context of privacy:* Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 54-55 (2017) (explaining the type of explanations that an automated system, based on AI technologies, can provide, which is largely referring to the algorithm model).

109 For similar stance in the context of privacy *see*: Edwards & Veale, *id.* at 43 (stressing the "transparency fallacy" of algorithmic decision making).

110 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.

111 AI Act Article 13.

112 For the absence of individuals' rights in the proposed AI Act, *see* Nathalie A. Smuha, Emma Ahmed-Rengers, Adam Harkens, Wenlong Li, James MacLaren, Riccardo Piselli, & Karen Yeung, *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act* (August 5, 2021), pp. 48-54, available at SSRN: https://ssrn.com/abstract=3899991.

laws, which imposes product liability standards, and for not setting a higher threshold in protecting individuals' rights, similar to the standard imposed by the GDPR.[113] The transparency problem concerning the use of AI systems, particularly for content moderation purposes, raises serious concerns regarding freedom of speech and the maintenance of democratic civil society that far exceed questions of product liability and consumer protection.[114] Considering the challenge of unexplainable automated decision making, possibly an entitlement for reassessment should be implemented based on an appellate body, capable of giving a meaningful reason for a decision in appropriate cases, when affected individuals request it.

B. Objective Review Involving Human Discretion

Another fundamental principle of procedural justice, possibly the most important one, concerns judicial review of administrative decisions,[115] which is the final check on illegal, inappropriate, or arbitrary action by government agencies.[116] This principle has been particularly developed since the 1980s by courts in common law countries, acknowledging widely that "any governmental decision affecting the rights, interests, property, privileges or liberties of any person was reviewable by the courts to ensure the legality and rationality of the decision and the fairness of the decision-making process."[117] Yet, this understanding applies to any organization seeking to meet procedural justice standards and to promote fairness in its decision-making process. Implementing a mechanism for objective review of decisions is therefore a key element for fair procedures in the public as well as in the private sectors.[118]

The DSA requires the establishment of a user complaint mechanism and of a "certified" out-of-court dispute settlement body that would meet basic standards of independence and apply "clear and fair rules of procedure."[119] Therefore, the DSA appears to adopt this basic principle of procedural justice. First, as noted, the two basic procedural justice principles, giving reasons for decisions and objective review,

---

113 Smuha et al., *supra* note 21, at 48–54 (explaining why the proposed AI Act fails to ensure meaningful transparency, accountability, and rights to public participation and why the proposed AI Act lacks meaningful substantive rights for individuals); Hannah Bloch-Wehba, *Transparency's AI Problem*, KNIGHT FIRST AMEND. INST. (June 17, 2021), https://knightcolumbia.org/content/transparencys-ai-problem [https://perma.cc/6QKN-Z4YR]; Sümeyye Elif Biber, *Machines Learning the Rule of Law: EU Proposes the World's First Artificial Intelligence Act*, VERFBLOG (July 13, 2021), https://ssrn.com/abstract=3951908 [https://perma.cc/MUF6-AWVQ]; Gianclaudio Malgieri, & Frank Pasquale, *From Transparency to Justification: Toward Ex Ante Accountability for AI* 1 (Brook. L. Sch., Legal Studies Working Paper No. 712, Brussels Privacy Hub Working Paper, No. 33, 2022), https://ssrn.com/abstract=4099657 [https://perma.cc/NR8G-J55A].

114 Charlotte Tschider, *Legal Opacity: Artificial Intelligence's Sticky Wicket*, 106 IOWA L. REV. 126, 160–64 (2021).

115 Daly, *supra* note 93, at 2.

116 Jacob A. Stein, Basil J. Mezines & Glenn A. Mitchell, ADMINISTRATIVE LAW, (MATTHEW BENDER 2023), at §1.01.

117 Daly, *supra* note 93, at 5.

118 Various reasons may support the application of public law norms in the private sector, *see* Orit Fischman Afori, *Online Rulers as Hybrid Bodies: The Case of Infringing Content Monitoring*, *supra* note 1.

119 DSA, Article 18.

are interrelated because there can be no poignant review of a decision if it is not explained and the underlying reasons are not disclosed.[120] Yet, the DSA does not specify what is the exact independent nature of a such review body and what are its concrete operative procedures. For instance, under the DSA, the user complaint mechanism may be operated by AI systems as well, and the insertion of a human in the loop is not required, despite concerns expressed on the topic.[121] The DSA requires, however, that in case an internal complaint handling is operated by an automated system, it must be ensured that the decision is made "under the supervision of appropriately qualified staff, and not solely on the basis of automated means."[122] Thus, although it is permitted to deploy automated review mechanisms, human involvement should be included in the oversight of the system. This standard does not require the application of human discretion in any decision made, only at the supervisory level. Again, the standard of oversight required by the AI Act is more akin to the standard required by product liability frameworks, and it is not setting a higher standard for protecting affected individuals' rights, as it is required, for example, in administrative law. Therefore, both the DSA and the AI Act fail to provide guarantees for individual rights because decisions made regarding limiting access to content might be non-explainable and the immediate appellate instance (the internal complaint system), which should provide a "second chance" for reviewing the decision and reassessing its outcome, might replicate the same flaws if operated again by an AI system. In other words, an AI system serving as a review body might fail to meet the objective oversight standard.

Likewise, the out-of-court dispute settlement bodies that should be available to affected users according to the DSA do not provide a remedy for the lack of human involvement in the reviewing process because they do not "have the power to impose a binding settlement of the dispute on the parties."[123] In other words, the out-of-court dispute settlement, which may meet higher standards of procedural justice principles, is voluntary.

Yet another accepted principle of administrative law pertains to the discretion that should be applied by a decision maker, known as the rule against fettering. According to this rule, the decision-making process must be free from *a priori* assumptions and conclusions, and each individual case should be inspected according to its circumstances.[124] The rule of "not fettering" reflects the importance of allowing flexibility in the process of operating discretion, enabling the decision maker to apply

---

120 Adler, *supra note* 96, at 325.

121 Schwemer & Schovsbo, *supra* note 51.

122 DSA Article 20 (6).

123 DSA Article 21 (2).

124 Adam Perry, *The Flexibility Rule in Administrative Law*, 76 CAMBRIDGE L.J. 375, 375 (2017) ("*An official who has a discretionary power may adopt a policy as to its exercise, but that policy must be flexible, not rigid. This rule is a branch of the principle against fettering discretion, which I will call the "flexibility rule*. *The flexibility rule is now nearly a century old");* Daly, Raso & Tomlinson, *supra* note 102, at 3-4.

exceptions and other discretionary measures that justify deviation from the general rigid rule.[125] Discretion enhances the reasonableness and rationality of a decision.[126]

A critical question is whether the rule of not fettering discretion could be observed by an automated system.[127] This question is gaining attention given the proliferation of automated systems in the service of many administrative bodies worldwide fulfilling various tasks, including decision making. AI technologies have had a pervasive effect on the operation of administrative bodies, seeking to streamline their functioning, performing what is known as "artificial administration."[128] The deployment of such technologies by administrative bodies raised the troubling question whether the use of automated decision-making systems was contrary to the principle against fettering discretion.[129] Decision making based on statistical correlations, which is at the heart of AI technologies, may fail to meet this standard of discretion.[130] Therefore, in some countries, service standard criteria have been defined for all digital government units.[131] Nevertheless, these standards do not obviate the need to preserve free discretion. The main challenge of the new artificial administration is coping with the loss of human discretion, including all its attributes, in the operation of administrative bodies.[132] In some cases, it may be argued that injecting human discretion into the decision-making process is necessary and inevitable if we want to maintain procedural justice principles.[133] Policymakers address the question of the need to insert a "human-into-the-loop" in many aspects and areas involving automated systems, beyond administrative law.[134] Procedural justice theory and the principle of not fettering may justify inserting a human into to the loop, especially where fair procedures are required by law.

The lessons derived from the operation of automated decision-making processes by administrative bodies could be implemented in the case of online content moderation systems. Given the importance of content moderation practices for the preservation of individual rights, headed by freedom of speech, and the need to ensure full-fledged procedural justice guarantees when limiting access to content, a genuine

---

125 Perry, *supra* note 124, at 376; FREDERICK SCHAUER, PLAYING BY THE RULES: A PHILOSOPHICAL EXAMINATION OF RULE-BASED DECISION-MAKING IN LAW AND LIFE, 108 (Oxford 1991).

126 Jerry L. Mashaw, *Explaining Administrative Process: Normative, Positive, and Critical Stories of Legal Development*, 6 J.L. ECON & ORG. 267, 267-268 (1990).

127 Paul Daly, *Artificial Administration: Administrative Law, Administrative Justice and Accountability in the Age of Machines*, *Forthcoming* AUSTRALIAN JOURNAL OF ADMINISTRATIVE LAW & PRACTICE, p. 5-6 (2023), Available at SSRN: https://ssrn.com/abstract=4434238.

128 Daly, *supra* note 127.

129 Daly, *supra* note 127, at p.3-4.

130 Daly, *supra* note 127, at p. 6.

131 JOE TOMLINSON, JUSTICE IN THE DIGITAL STATE: ASSESSING THE NEXT REVOLUTION IN ADMINISTRATIVE JUSTICE, 69-70 (2019).

132 Daly, *supra* note 127, at p. 2.

133 Daly, *supra* note 127, at p. 21-22.

134 *See e.g.* Eduardo Mosqueira-Rey et al, *Human-In-The-Loop Machine Learning: A State Of The Art*, 56 ARTIFICIAL INTELLIGENCE REVIEW, 3005 (2023); Therese Enarsson, Lena Enqvist & Markus Naarttijärvi, *Approaching The Human In The Loop – Legal Perspectives On Hybrid Human/Algorithmic Decision-Making In Three Contexts*, 31 INFORMATION & COMMUNICATIONS TECHNOLOGY LAW, 123 (2022).

review and oversight body should be established, based on human discretion. Such an appellate body may also serve to explain decisions and provide reasons for them. This procedural standard is appropriate at least when a take-down decision affects an important interest of the user, for example, propagating political content explicitly protected under freedom of speech.[135]

C. Hearings for Affected Individuals

The importance of appellate instances for observing the principles of procedural justice is also reflected by the requirement to provide affected individuals an opportunity to be heard, voice their positions, and give evidence essential for the decision-making process. In the US, this principle also underpins the due process clause enshrined in the Fifth Amendment to the Constitution, which requires a hearing when property or liberty is at stake, where the affected individual can be heard in administrative adjudications.[136] Allowing affected individuals to be heard serves the principle of basing decisions on objective criteria and accurate information and evidence.[137] Thorough evidence-based decisions are reliable and form the basis for objective oversight and review so that in appropriate cases, mistakes can be reassessed and corrected.[138] Another solid justification for the requirement of a hearing concerns the rule against fettering of discretion discussed above, because giving affected individuals a fair and good-faith opportunity to present their position enables the decision maker to exercise informed discretion after relevant information and evidence were presented.[139]

A growing body of studies about procedural justice stresses the important role that the participation of involved parties in appellate instance litigation plays in building trust in the legal system and its legitimacy.[140] Fair procedures, allowing affected parties to bring their case for judicial review that includes their participation in the process encourage the parties' willingness to view the decision as legitimate even in the face of a negative outcome.[141] Thus, appellate mechanisms based on fair

---

135 For example, the minutes of a political meeting have been found to stand at the heart of freedom of expression, overriding other interests that may prevent its publication; *see*: Ashdown v. Telegraph Group [2001] R.P.C. 34, [2002] QB 546; Kevin Garnett, *The Impact of the Human Rights Act 1998 on UK Copyright Law*, *in* COPYRIGHT AND FREE SPEECH: COMPARATIVE AND INTERNATIONAL ANALYSES, 171 (Jonathan Griffiths & Uma Suthersanen eds., 2005).

136 Stein, Mezines & Mitchell, *supra* note 116, §31.02; Solum, *supra* note 92, at 264-264.

137 Martin H. Redish & Lawrence C. Marshall, *Adjudicatory Independence and the Values of Procedural Due Process*, 95 YALE L. J. 455, 483 (1986) (discussing the principle of accuracy).

138 Jerry L. Mashaw, *Judicial Review of Administrative Action: Reflections on Balancing Political, Managerial and Legal Accountability*, DIREITO GV L. REV. 153 (2005); Stein, Mezines & Mitchell, *supra* note 116, §31.02.

139 Perry, *supra* note 124, at 376.

140 Rebecca Hollander-Blumoff, *The Psychology of Procedural Justice in the Federal Courts*, 63 HASTINGS L.J., 127 (2011); Molaskey, *supra* note 105..

141 Molaskey, *supra* note 105, at 83-84; Hollander-Blumoff, *supra* note 140; Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUST. 283, 298-299 (2003) (explaining that models of procedural justice focus on two factors: "the quality of decision making and the quality of interpersonal treatment," *id.* at 299).

procedures that include hearings play an important societal role in supporting the rule of law and obedience to it.[142]

Both the DSA and the AI Act fail to address the requirement to allow a hearing of individuals affected by decisions to limit access to content. The DSA, as noted, requires only the establishment of "fair procedures" by the dispute settlement body but does not clarify what such procedures should be. Based on the other two principles discussed above, giving reasons for decisions and establishing an objective review mechanism, and acknowledging the fact that the appellate instance can provide the first opportunity for issuing an explainable decision, affected individuals should have access to an appellate instance and be heard at least on appeal. Allowing hearings in cases of content moderation decisions is particularly important because with regard to "awful but lawful" content, such as false information, the decision is not clear cut.[143] Occasionally, a better understanding of the true meaning of the content, which may be subject to a range of perceptions and interpretations, such as "hate speech," may require an in-depth inquiry.[144] Likewise, a decision to take down allegedly illegal content may involve complex assessment of the relevant legal framework. For example, a decision to take down allegedly copyright-infringing content may involve assessment of the US fair use doctrine that permits unauthorized use of copyrighted works but is known for its uncertainties.[145] The US fair use doctrine is applied on a case-by-case basis and leans heavily on the facts relevant to the particular case.[146] Thus, an informed discussion and evidence-based assessment of the case at hand, before a decision is made is needed, at least on appeal. To that end, a hearing can provide the relevant information and evidence the decision maker needs to make an accurate decision. Accuracy is also one of the building blocks of fair procedures because it reflects the rule of law.[147] The accuracy of the decision is important for every outcome, whether to take down harmful content or allow the spread of legitimate content.

---

142 Molaskey, *supra* note 105, at 84-85.

143 A question extensively discussed in the literature is how to differentiate between the various types of false information. An accepted classification differentiates between *misinformation*, "information that is false, but spread unintentionally and without intent to cause harm," and *disinformation*, "false information that is deliberately created or disseminated with the express purpose to cause harm or make profit.", *see* Ben Epstein, *Why It Is So Difficult to Regulate Disinformation Online*, *in* THE DISINFORMATION AGE POLITICS, TECHNOLOGY, AND DISRUPTIVE COMMUNICATION IN THE UNITED STATES, 190, 192 (W. Lance Bennett & Steven Livingston eds. 2020). Disinformation is the category that is usually viewed as dangerous to democracies and therefore should be handled and regulated; see Epstein, *id*, at 193-194.

144 For example, it is not necessarily easy to identify certain speech as meeting the threshold of "hate speech," therefore it may be controversial whether it is justified to limit access to it; *see*: Andrew F. Sellars, *Defining Hate Speech* (December 1, 2016). Berkman Klein Center Research Publication No. 2016-20, Boston Univ. School of Law, Public Law Research Paper No. 16-48, Available at SSRN: https://ssrn.com/abstract=2882244;

145 Niva Elkin-Koren & Orit Fischman Afori, *Rulifying Fair Use*, 59 ARIZONA LAW REVIEW, 161 (2017).

146 *See e.g.* Cambridge Univ. Press v. Becker, 863 F. Supp. 2d 1190 (N.D. Ga. 2012), *rev'd sub nom.* Cambridge Univ. Press v. Patton, 769 F.3d 1232 (11th Cir. 2014).

147 Redish & Marshall, *supra* note 137, at 483.

V. Concluding Remarks

The deployment of AI systems in decision-making processes aims to streamline the operation of such frameworks, particularly when massive amounts of decisions need to be made. Therefore, the use of AI systems is pervasive in conducting administrative functions and handling tasks by online platforms. Our proposals for strengthening the fairness of these procedures for the sake of democratizing the digital environment reveal the need for greater involvement of human discretion in the process.

These proposals may be criticized for not being realistic regarding the time and effort invested in the application of content moderation practices. Human involvement inevitably slows down the pace of decision making, requires the investment of resources by the online platforms, and therefore may not be efficient. Another common argument against our proposals is that human involvement does not necessarily produce "better" outcomes. For example, many online platforms employ human fact-checkers and flaggers to handle false information. These tasks are carried out both by paid employees and independent users, that is, "community members."[148] Human inspection of this type also involves questions of reliability and potential conflict with freedom of speech and individual rights, as noted in the EU Voluntary Code of Practice on Disinformation.[149]

All this may be true.

Nevertheless, the underlying rationale of the human rights school of thought, which includes the protection of individual rights, among others, using procedural justice measures, is that efficiency in its pure economic understanding is not necessarily a paramount value. At times, the adoption of certain measures that necessitate investment of greater time and effort is inevitable for the accomplishment of goals other than short-term efficiency. Promoting a digital environment that better guarantees individual rights may be a costly endeavor. The new phase of digital constitutionalism requires that we move from a "liberal economic perspective to a constitution-oriented approach."[150] And from a long-term perspective, even if inspected in terms of economic efficiency, the effort to democratize the control of content moderation practices may result in a Pareto equilibrium for humanity.

---

148 Mark Leiser, *Reimagining Digital Governance: The EU's Digital Service Act and the Fight Against Disinformation* (April 24, 2023). Available at SSRN: https://ssrn.com/abstract=4427493.

149 To address the ill consequences of such practices, the EU has initiated in 2018 a Code of Practice on Disinformation, that is aimed at empowering industry to adhere to self-regulatory standards to combat disinformation. In 2022, the EU has published a strengthened version of the code, that includes guidelines concerning maintaining principles of freedom of expression along with content moderation practices exercised by humans, *see*: file:///C:/New%20Folder/Dropbox/My%20PC%20(DESKTOP-CGQADNQ)/Downloads/2022_Strengthened_Code_of_Practice_Disinformation_TeAETn7bUPXR57PU2FsTqU8rMA_87585.pdf

150 *See* De Gregorio, *supra* note 5, at 41.