ARTICLE

THE OVERSIGHT OF CONTENT MODERATION BY AI: IMPACT ASSESSMENTS AND THEIR LIMITATIONS

YIFAT NAHMIAS* & MAAYAN PEREL†

ABSTRACT

In a world in which artificial intelligence (AI) systems are increasingly shaping our environment, as well as our access to and exclusion from opportunities and resources, it is essential to ensure some form of AI oversight. Such oversight will help to maintain the rule of law, to protect individual rights, and to ensure the protection of core democratic values. Nevertheless, achieving AI oversight is challenging due to the dynamic and opaque nature of such systems. Recently, in an attempt to increase oversight and accountability for AI systems, the proposed US Algorithmic Accountability Act introduced mandatory impact assessment for private entities that deploy automated decision-making systems. Impact assessment as a means to enhance oversight was likewise recently adopted under the EU's General Data Protection Regulation. Taken together, these initiatives mark the latest development in AI oversight policy.

In this paper, we question the merits of impact assessment as a tool for promoting oversight of AI systems. Using the case of AI systems of content moderation, we highlight the strengths and weaknesses of this oversight tool and propose how to improve it. Additionally, we argue that even an improved impact assessment does not fit equally with the oversight challenge raised by different systems of AI. Especially, impact assessments might be insufficient to oversee AI systems that are deployed to achieve purposes that could be classified as public, such as making our online public sphere safer. Meaningful oversight of AI systems that impose costs on society as a whole, like AI systems of content moderation, cannot be pursued by mechanisms of self-assessment alone. Therefore, as we suggest in this paper, such systems should be additionally subjected to objective mechanisms of external oversight.

TABLE OF CONTENTS

* Senior Research Fellow, Center for Cyber Law and Policy (CCLP), University of Haifa; Researcher, The Heth Academic Center for Research of Competition and Regulation, College of Management.

†Associate Professor, Netanya Academic College, Faculty of Law; Senior Research Fellow, Haifa Center for Law & Technology, University of Haifa Faculty of Law. This research was supported by The Israel Science Foundation (grant No. 1820/17).

I. Introduction

In recent years, the world has witnessed an exponential growth in the use of artificial intelligence ("AI") and other automated decision-making systems. Government institutions increasingly rely on automated decision-making technologies in many areas, such as managing traffic,[1] conducting risk assessments,[2] screening immigrants,[3] allocating social services,[4] and

[1] *See, e.g.*, Miguel Carrasco et al., *The Citizen's Perspective on the Use of AI in Government*, Bos. Consulting Group (Mar. 1, 2019), https://www.bcg.com/en-il/publications/2019/citizen-perspective-use-artificial-intelligence-government-digital-benchmarking.aspx [https://perma.cc/W8QM-ACM3]; *Red Ninja's Smart Tech Clears the Road for Ambulance Crews*, Innovate UK (Aug. 7, 2017), https://www.gov.uk/government/case-studies/red-ninjas-smart-tech-clears-the-road-for-ambulance-crews [https://perma.cc/6NC3-GAAW].

[2] *See, e.g.*, State v. Loomis, 881 N.W.2d 749, 753–54 (Wis. 2016); Karl Manheim & Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, 21 Yale J.L. & Tech. 106, 155–56 (2019).

[3] *See, e.g.*, Margaret Hu, *Algorithmic Jim Crow*, 86 Fordham L. Rev. 633, 641 (2017).

[4] *See, e.g.*, Aaron Rieke et al., *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods: An Upturn and Omidyar Network Report*, Omidyar Network 7 (Feb. 27, 2018), https://omidyar.com/wp-content/uploads/2020/09/Public-Scrutiny-of-Automated-Decisions.pdf [https://perma.cc/6JS7-RZW7].

more.[5] Likewise, private companies have integrated AI into their hiring processes,[6] lending and loan management,[7] and other functions.[8] In fact, many decisions that were once carried out by humans are now gradually being placed into the hands of automated AI systems.[9] Content moderation, which is at the focus of this paper, is one prominent area in which decisions are increasingly governed by AI. The growing pervasiveness of AI-based systems that govern human behavior carries with it numerous advantages, but also heightens the need to ensure sufficient oversight of such automated decision-making processes.[10] A major concern is that AI systems exhibit and intensify human biases and unfair, discriminatory, and derogatory practices.[11] For example, Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS"), a case-management and decision-support system used by U.S. courts to assess the likelihood of a defendant to reoffend, which is often used to inform bail decisions,[12] was reported to underestimate the probability of white recidivism, while overestimating the

---

[5] The Computational Journalism Lab at Northwestern University curated a set of algorithms being used in the U.S. federal government. *See Algorithm Tips*, COMPUTATIONAL JOURNALISM LAB, http://algorithmtips.org/ [HTTPS://PERMA.CC/44A8-KFF4].

[6] Hilke Schellmann & Jason Bellini, *Artificial Intelligence: The Robots Are Hiring*, WALL ST. J. (Sept. 20, 2018), https://www.wsj.com/articles/artificial-intelligence-the-robots-are-now-hiring-moving-upstream-1537435820 [https://perma.cc/GMB9-CQNW].

[7] Daniel Fagella, *Artificial Intelligence Applications for Lending and Loan Management*, EMERJ (May 19, 2019), https://emerj.com/ai-sector-overviews/artificial-intelligence-applications-lending-loan-management/ [https://perma.cc/T7H5-DW77].

[8] For instance, Uber is using AI to identify and circumvent officials in cities all over the world. *See* Mike Isaac, *How Uber Deceives the Authorities Worldwide*, N.Y. TIMES (Mar. 3, 2017), https://www.nytimes.com/2017/03/03/technology/uber-greyball-program-evade-authorities.html [https://perma.cc/LZR4-86QP].

[9] Shlomit Yanisky-Ravid & Sean K. Hallisey, *"Equality and Privacy by Design": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes*, 46 FORDHAM URB. L.J. 428, 431 (2019).

[10] *See* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 111–13 (2019); John O. McGinnis, *Accelerating AI*, 104 NW. U. L. REV. 366, 379–81 (2010). The legal scholarship can be roughly divided into two opposing views: those who acknowledge the threat AI poses and those who dismiss it. *See generally* Brian S. Haney, *The Perils and Promises of Artificial General Intelligence*, 45 J. LEGIS. 151 (2018); Matthew U. Scherer, *Regulating Artificial Intelligence Systems*, 29 HARV. J.L. & TECH. 353 (2016).

[11] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 1–18 (2015); Joshua New & Daniel Castro, *How Policymakers Can Foster Algorithmic Accountability*, CTR. FOR DATA INNOVATION 1, 3 (May 21, 2018), http://www2.datainnovation.org/2018-algorithmic-accountability.pdf [https://perma.cc/S6MZ-GULM]; Karen Yeung et al., *AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing*, *in* THE OXFORD HANDBOOK OF ETHICS OF AI 77, 78 (Markus D. Dubber et al. eds., 2020) (suggesting a "human-rights-centered design, deliberation, and oversight approach" to the governance of AI); Muhammad Ali et al., *Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes*, ARXIV (Sept. 19, 2019), https://arxiv.org/pdf/1904.02095.pdf [https://perma.cc/PRK3-8CJU].

[12] Aaron M. Bornstein, *Are Algorithms Building the New Infrastructure of Racism?*, NAUTILUS (Dec. 21, 2017), http://nautil.us/issue/55/trust/are-algorithms-building-the-new-infrastructure-of-racism [https://perma.cc/HN63-64PV].

probability of black recidivism.[13] Specifically, the system "was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants" and "[w]hite defendants were mislabeled as low risk more often than black defendants."[14] When the effect of race was isolated from criminal history and recidivism, as well as from defendants' age and gender, "black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind."[15] In other words, the system generated incorrect conclusions or "bias" against African-Americans. These problems also plague private entities. For instance, although developed to help with employment recruitment, Amazon's experimental AI-based hiring system showed bias against women when it learned from the training to favor candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured."[16] How did this happen? As reported, Amazon's computer models were essentially "trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry."[17] In other instances, AI-led systems demonstrated bias in facial recognition technologies,[18] bias in online ads,[19] and bias in word association.[20] Given the widespread presence of AI-based systems, the encapsulating of prejudices and biases could have a direct impact on individuals' fundamental rights, such as freedom of speech, privacy, equality, and

---

[13] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/EUW3-NPLY].

[14] *Id.*

[15] *Id.*

[16] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/N4YF-6AMG]; *see also* Isobel Asher Hamilton, *Why It's Totally Unsurprising that Amazon's Recruitment AI Was Biased Against Women*, BUS. INSIDER (Oct. 13, 2018), https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10 [https://perma.cc/4FL7-RBHK].

[17] Dastin, *supra* note 16.

[18] *See* Larry Hardesty, *Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems*, MIT NEWS (Feb. 11, 2018), http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 [https://perma.cc/3N3X-QCLN].

[19] *See generally* Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMM. ASS'N FOR COMPUTING MACHINERY 44 (May 2013), https://privacytools.seas.harvard.edu/files/privacytools/files/p44-sweeney.pdf [https://perma.cc/3NTT-KNKW].

[20] *See* Adam Hadhazy, *Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices*, PRINCETON U. OFF. ENGINEERING COMM. (Apr. 18, 2017), https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices [https://perma.cc/A8DH-EPEN]; *see also* Nicol Turner Lee et al., *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, BROOKINGS (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ [https://perma.cc/5CYM-PTYX].

autonomy. Legislatures around the world are attuned to such risks, and several legislative initiatives have recently been introduced to address them, including a mandatory impact assessment for private entities that rely on automated decision-making systems. Nevertheless, as this paper shows, impact assessments are insufficient to achieve their intended purpose.

The U.S. Algorithmic Accountability Act of 2019 ("Algorithmic Accountability Act"),[21] for instance, seeks to require all corporations that use "automated decision systems"[22] to submit impact assessments of the accuracy, fairness, bias, discrimination, privacy, and security of their automated decision-making systems to the Federal Trade Commission ("FTC").[23]

Another prominent example is the E.U.'s broad-reaching General Data Protection Regulation ("GDPR"), which provides two important accountability-enhancing mechanisms: a requirement that regulated entities submit to impact assessment and an individual's right to explanation. These two mechanisms are intended to produce better oversight of systems that solely depend on automated decision-making.[24]

Although these initiatives aim to make AI systems accountable, they remain insufficient, given the myriad issues inherent to AI. This problem is particularly salient in the area of content moderation.

The above-mentioned initiatives—i.e., the Algorithmic Accountability Act and the GDPR—are novel in requiring entities that are implementing AI-based or other automated decision-making systems to deploy an impact assessment.

Although different jurisdictions have different impact-assessment schemes in place, each has its own specificities and objectives. Generally, an impact assessment can be defined as "the process of identifying the future consequences of current or proposed action."[25] Impact-assessment schemes carry some important advantages. They improve organizational behavior,

---

[21] *See* Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019). It is important to note this is not the first attempt to regulate automated decision-making. In 2017, the New York City Council passed legislation to establish public accountability for the city of New York's use of algorithms. *See Testimony of the New York Civil Liberties Union Before the New York City Council Committee on Technology Regarding Automated Processing of Data*, N.Y.C.L. UNION (Oct. 16, 2017) https://www.nyclu.org/en/publications/nyclu-testimony-nyc-council-committee-technology-re-automated-processing-data [https://perma.cc/26X9-G298].

[22] The term "automated decision system" is defined as "a computational process, including one derived from machine learning, statistics, or other data processing or artificial intelligence techniques, that makes a decision or facilitates human decision making, that impacts consumers." H.R. 2231 § 2(1).

[23] *Id.* § 2.

[24] *See* Commission 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), 2016 O.J. (L 119) 1, 7–10 [hereinafter Commission Regulation 2016/679].

[25] See the definition employed by the International Association for Impact Assessment. INT'L ASS'N FOR IMPACT ASSESSMENT, https://www.iaia.org [https://perma.cc/XGR6-PT7F] [hereinafter IAIA].

promote information sharing, and incentivize private entities to consider ex ante the effect of their AI-led systems on individuals as well as on the public at large.[26] But are they sufficient to maintain oversight of all AI systems? First, as our analysis suggests, the way that current initiatives structure impact assessments still falls short of facilitating sufficient accountability. In particular, impact assessments provide only limited transparency, secure due process insufficiently, and allow only limited room for public review.[27] The initiatives discussed throughout this paper do not require regulated entities to disclose any part of the self-assessment to the public, nor do they provide other means for the public to know that specific conduct took place. Further, under the Algorithmic Accountability Act, individuals are not entitled to notice or the right to be heard. Consequently, the strategy of impact assessment fails to facilitate sufficient public oversight and proper opportunities for correcting erroneous decisions. To improve their oversight potential, we thus recommend some improvements to the existing impact-assessment schemes, including periodic impact assessments, mandatory notice-and-comment procedure, and mandatory publication.

Second, we find that those impact assessments might not fit the oversight challenges raised by different forms of AI-based systems. Focusing on the case of AI-based online content moderation by online platforms, we argue that an oversight mechanism of self-assessment—such as the one discussed throughout this paper (i.e., impact assessments)—is insufficient to oversee private moderation of speech that directly and substantially affects shared public interests. The application of AI-based content-moderation systems by prominent online platforms is riddled with externalities. It directly affects people's ability to engage in certain forms of expression, communication, and sharing of thoughts and critical information. Consequently, it shapes our online public sphere and ultimately governs the free flow of information.[28] Since platforms are private actors, at first glance, mechanisms of self-assessment seem to be the most suitable way to hold them accountable. Indeed, despite concerns that "the real threat to free speech today comes from private entities such as Internet service providers, not from the Government,"[29] interfering with the editorial discretion of platforms is seen as a violation of platforms' First Amendment rights under the United States Constitution.[30] As commercial speakers, platforms might be entitled to the con-

---

[26] For further discussion, *see infra* Parts III.B, IV.

[27] For further discussion, *see infra* Parts III.B, IV.

[28] *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1622–24 (2018).

[29] U.S. Telecom Ass'n v. FCC, 855 F.3d 381, 434 (D.C. Cir. 2017) (Kavanaugh, J., dissenting).

[30] *See* Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power over Online Speech*, HOOVER INSTITUTION 1, 17–22 (2019) https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf [https://perma.cc/KB3N-DFBX].

stitutional protection of free speech.[31] Impact assessments fit well within this deeply rooted scheme because they are non-coercive and collaborative and therefore can be generally regarded as a form of self-regulation.[32] They do not force platforms to speak by demanding them to host content against their will[33] but instead require them to be more transparent about their goals and evaluate the possible implications of their systems.

Nevertheless, a deep analysis of the operation and impacts of AI-based systems of online content moderation suggests that mechanisms of self-regulation cannot sufficiently oversee them. While the impact assessments enunciated under the Algorithmic Accountability Act and the GDPR are tailored to mitigate concerns about the ways general AI-based decision-making systems affect individuals,[34] the most worrying consequences of poorly performed AI-driven content moderation concern our online public sphere. Although the removal of legitimate content affects the speaker's freedom of expression, it also affects the interest of the public in freely consuming and accessing information. Hence, the use of AI for content moderation can impose costs, not only upon the individual speaker, but especially upon society.[35]

Nonetheless, in contrast to evaluating the impact of AI-based systems on individuals (such as assessing the impact of an incorrect credit score), it is extremely difficult to evaluate the public impact of AI-based content moderation. The main reason for this difficulty is that AI-based content moderation is personalized. Even if platforms disclose how they minimize the spread of harmful content, they do not apply a common threshold of content legitimacy for all users. Indeed, in practice each individual views a different curated segment of the online discourse that meets her personal profile. The idea of a common public discourse is consequently becoming a fiction, given that AI-based content-moderation systems create personally tailored but fragmented "publics" of information.[36] As a result, it is hardly possible to detect illegitimate deprivations of information. If a user does not see a specific piece of information, it is not necessarily because this piece of content was removed, but possibly because it did not match her personal inter-

---

[31] *See* Jack Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. 1149, 1152 (2018).

[32] *See generally* Katyal, *supra* note 10. *See also* Michael Guihot et al., *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 427 (2017).

[33] *See generally* La'Tiejira v. Facebook, Inc., 272 F. Supp. 3d 981, 991–92 (S.D. Tex. 2017); Daniel Y. Zhang et al., *Crowdsourcing-Based Copyright Infringement Detection in Live Video Streams*, 2018 PROC. 2018 IEEE/ACM INT'L CONF. ON ADVANCES SOC. NETWORKS ANALYSIS & MINING 367. *See also* Search King, Inc. v. Google Tech., Inc., No. CIV-02-1457-M, 2003 WL 21464568 (W.D. Okla. May 27, 2003).

[34] *See supra* notes 16–20 and accompanying text.

[35] For further discussion, see *infra* Part IV.C.

[36] Anat Ben-David, *Counter-Archiving Facebook*, 35 EUR. J. COMM. 249, 255 (2020).

ests.[37] Yet, without having a common and accessible view of what counts as our public discourse, it is extremely challenging to determine if a platform's AI-based system of content moderation complies with what is disclosed in its impact assessment.

Given these insights, this paper provides two important contributions. First, on a general level, it highlights several shortcomings of impact assessments and proposes how to address them in order to enhance their oversight potential. Second, on a specific level, this paper shows that different contexts of AI-based decision-making systems may require different processes and levels of oversight. Specifically, to generate accountability in AI-based content-moderation systems, it is insufficient to count on self-assessment conducted by platforms, but rather it is necessary to subject them to a higher level of external and objective scrutiny. Part II begins with an introduction to the importance of accountability, transparency, and public scrutiny in the realm of AI and other automated decision-making systems and is then followed by a short review of its inherent challenges. Part III reviews the latest initiatives designed to enable some form of oversight over automated decision-making systems by using impact assessments. Next, it demonstrates how—despite having some benefits—these schemes fail to achieve meaningful accountability. As above, impact assessments provide only limited transparency, secure due process insufficiently, and allow limited room for public review. Part IV surveys key features of AI systems of content moderation that makes them a special case of AI, including contextual decision-making and public sphere attributes. In the last part, this paper will offer a novel dual mechanism of oversight for AI-based content-moderation systems, comprised of internal checks and external auditing, which may facilitate more efficient oversight of content moderation by AI.

II. Oversight of AI-Based Decision-Making

A. Notions of Accountability

AI-based systems are increasingly applied to make decisions that dramatically affect individuals. Government institutions use such systems to conduct risk assessments,[38] to screen immigrants,[39] and to allocate social services.[40] Likewise, private companies have applied AI to assist them in making hiring decisions[41] or managing loan determinations.[42] Repeat concerns about AI systems exhibiting and intensifying both human biases and unfair,

---

[37] Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 Lewis & Clark L. Rev. 857, 876 (2020).

[38] *See supra* note 2.

[39] *See, e.g.*, Hu, *supra* note 3, at 641.

[40] *See* Rieke et al., *supra* note 4.

[41] Schellmann & Bellini, *supra* note 6.

[42] *See* Fagella, *supra* note 7.

discriminatory, and derogatory practices[43] led to statutory initiatives that aim to increase the accountability of these systems.[44] The idea behind notions of accountability is to ensure that decision makers do not abuse their power but rather exert it in a fair and effective manner for the benefit of the public. Indeed, "people with public responsibilities should be answerable to 'the people' for the performance of their duties."[45] "Such persons are expected to justify their choices to those affected by these choices, and be held responsible for their failures and wrongdoings."[46] A host of doctrines, procedures, laws, and regulations are employed in order to hold government and public officials accountable for their decision-making processes.[47] In addition, freedom-of-information laws[48] and sunshine laws[49] attempt to ensure that governmental decision-making processes are open to some form of inspection, either by requiring governmental bodies to make their records available for public scrutiny,[50] or by giving the public access to observe agency meetings.[51]

With private bodies, accountability can be enforced through legal rules and regulations, but also through informal means, such as market forces that check decision makers' discretion and promote voluntary disclosure in rela-

---

[43] *See, e.g.*, *supra* note 11 and accompanying text.

[44] For further discussion, see *infra* Part III.

[45] Michael W. Dowdle, *Public Accountability: Conceptual, Historical, and Epistemic Mappings*, *in* PUBLIC ACCOUNTABILITY: DESIGN, DILEMMAS AND EXPERIENCES 1, 3 (Michael W. Dowdle ed., 2006).

[46] Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 481 (2016).

[47] *See* Jim Rossi, *Participation Run Amok: The Costs of Mass Participation for Deliberative Agency Decisionmaking*, 92 NW. U. L. REV. 173, 175 (1997); Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L.J. 377, 399–400 (2006); Michele Estrin Gilman, *"Charitable Choice" and the Accountability Challenge: Reconciling the Need for Regulation with the First Amendment Religion Clauses*, 55 VAND. L. REV. 799, 803 (2002); Mark Bovens, *Analyzing and Assessing Public Accountability. A Conceptual Framework* 7–8, (Eurogov Euro. Governance Papers, Paper No. C-06-01, 2006), https://www.ihs.ac.at/publications/lib/ep7.pdf [https://perma.cc/UYK9-7NT3] (noting that, in contemporary scholarly discourse, the term "accountability" is often used to denotate various distinct concepts including transparency, equity, democracy, efficiency, responsiveness, responsibility, and integrity, and arguing its narrower definition can be understood as the obligation to explain and justify conduct).

[48] *See, e.g.*, Freedom of Information Act, 5 U.S.C. § 552 (2018); District of Columbia Freedom of Information Act, D.C. CODE §§ 2-531–540 (2006); Arizona Public Records Law, ARIZ. REV. STAT. §§ 39-121–128 (2009); Kentucky Open Records Act, KY. REV. STAT. ANN. §§ 61.870–884 (2009). For further discussion, see Justin Cox, *Maximizing Information's Freedom: The Nuts, Bolts, and Levers of FOIA*, 13 CUNY L. REV. 387, 413 nn.117–21 (2010).

[49] *See* Government in the Sunshine Act, 5 U.S.C. § 552b (2018).

[50] *See* 5 U.S.C. § 552; U.S. Dep't of Justice v. Reps. Comm. for Freedom of the Press, 489 U.S. 749, 772–73 (1989) (stating that the one key aim of FOIA is informing citizens "what their government is up to"). It is important to note that there are nine statutory exemptions, 5 U.S.C. § 552(b)(1)–(9), and three exclusions, *id.* § 552(c)(1)–(3), to the open records requirement. Moreover, FOIA "does not obligate agencies to create or retain documents; it only obligates them to provide access to those which it in fact has created and retained." Kissinger v. Reps. Comm. for Freedom of the Press, 445 U.S. 136, 152 (1980).

[51] *See* 5 U.S.C. § 552b.

tion to their choices and related outcomes.[52] Using market forces, members of the public can penalize private entities for unacceptable behavior[53] or even compel organizations to change their practices and alter their behavior.[54]

B. Accountability in AI-Based Governance: The Challenges

A system of governance by AI challenges these notions of accountability.[55] First, AI-based systems operate behind closed doors and are therefore considered a "black box"[56] in the sense that the public has only limited access to, and very little understanding (if any) of, how they work in practice.[57] Indeed, when AI relies on machine learning algorithms, "there is no straightforward way to map out the decision-making process of these complex networks of artificial neurons."[58] While these systems could be as complex as the human brain, they cannot be explained by legal doctrines that focus on human conduct rather than the learning capabilities of algorithms.[59] This means that members of the public have no way of knowing how the decision-making process works, what the goals are that the system was designed to carry out, or how a specific recommendation or decision was de-

---

[52] Perel & Elkin-Koren, *supra* note 46, at 482.

[53] Indeed, Schedler argues that accountability can be seen as the synthesis of two concepts: answerability and enforcement. *See* Andreas Schedler, *Conceptualizing Accountability*, *in* THE SELF-RESTRAINING STATE: POWER AND ACCOUNTABILITY IN NEW DEMOCRACIES 13, 14–15 (Andreas Schedler et al. eds., 1999). The former refers "to the right to receive information and the corresponding obligation to release all necessary details." *Id.* Thus, it can be roughly broken down into transparency and justification. The latter focuses on the idea that "accounting actors do not just 'call into question' but also 'eventually punish' improper behavior." *Id*. at 15–17. Although not directly addressed by Schedler's bipartite categorization, both market forces and public discourse can serve as a key catalyst in this so-called punishment. *See id.* at 14–17.

[54] Thomas N. Hale, *Transparency, Accountability, and Global Governance*, 14 GLOB. GOVERNANCE 73, 77–87 (2008) (discussing market pressure and its limitations); *see also* Orna Rabinovich-Einy, *Technology's Impact: The Quest for a New Paradigm for Accountability in Mediation*, 11 HARV. NEGOT. L. REV. 253, 260–61 (2006).

[55] Perel & Elkin-Koren, *supra* note 46, at 481–84.

[56] *See, e.g.*, PASQUALE, *supra* note 11; Nicholas Diakopoulos, *Algorithmic Accountability: On the Investigation of Black Boxes*, TOW CTR. FOR DIG. JOURNALISM (Dec. 3, 2013), https://www.cjr.org/tow_center_reports/algorithmic_accountability_on_the_investigation_of_black_boxes.php [https://perma.cc/9W6J-MV6F].

[57] *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1129 (2018) (noting transparency "is a particularly pronounced problem in the case of machine learning, as its value lies largely in finding patterns that go well beyond human intuition.").

[58] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 891–92 (2018).

[59] *See id.*

rived.[60] Moreover, when faced with a black box, the public has little chance of pressuring private entities into modifying their behavior.[61]

Second, even if the system is not completely closed and the public is privy to some information, AI-based decision-making systems are highly complex and constantly changing.[62] Thus, any attempt to review these decision-making processes and their results—even by a very tech-savvy individual—becomes even harder to accomplish.[63] To illustrate, consider the case of Mount Sinai Hospital in New York. When a group of researchers at Mount Sinai Hospital employed a system of deep learning to the hospital's database, the resulting program was proven to be extremely good at predicting diseases, including psychiatric disorders like schizophrenia. However, even its own designers do not know how.[64] This is a major drawback of any call to adopt a right to explanation.[65]

In addition, AI-driven systems not only implement specific rules and policies—whether originating from a private entity or the legislature—but also constantly reshape rules and policies in order to accommodate changes and new information. AI systems continuously improve their decision-making processes based on their accumulated information (e.g., via machine learning and deep learning),[66] thus rendering decision-making a continuous

---

[60] *See* Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 1 BIG DATA & SOC'Y 1, 3 (2016).

[61] *See* Roger Bickerstaff, *Does Your Machine Mind? Ethics and Potential Bias in the Law of Algorithms*, DIGITAL BUS. L. (June 19, 2017), https://digitalbusiness.law/2017/06/does-your-machine-mind-ethics-and-potential-bias-in-the-law-of-algorithms/ [https://perma.cc/W2A4-XKR7] ("Greater transparency of the principles, parameters and logic [underpinning] AI and algorithms in particular may lead to public review and scrutiny. This is likely to [be] a lot more effective in putting pressure on digital players to conform with good principles."); *Study of the Panel for the Future of Science and Technology on "A Governance Framework for Algorithmic Accountability and Transparency,"* at 5 (Apr. 2019), https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf [https://perma.cc/BUY3-JU55].

[62] *See* FILIPPO RASO ET AL., BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y, ARTIFICIAL INTELLIGENCE & HUMAN RIGHTS: OPPORTUNITIES AND RISKS 10 (2018).

[63] *See generally* Jef Ausloos et al., *Algorithmic Transparency and Accountability in Practice*, 2018 ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS. 1, https://uploads-ssl.webflow.com/5a2007a24a11ce000164d272/5ac883392c10d1baaa4358f2_Algorithmic_Transparency_and_Accountability_in_Practice_CameraReady.pdf [https://perma.cc/HK48-V9TD].

[64] Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [https://perma.cc/Z4SE-66MY].

[65] Ctr. for Data Innovation, Comment Letter to the FTC Regarding Competition and Consumer Protection in the 21st Century Hearings, Project Number P181201 (Feb. 15, 2019), http://www2.datainnovation.org/2019-ftc-competition-consumer-protection.pdf [https://perma.cc/Z7QN-34P7].

[66] *See* Raso et al., *supra* note 62, at 10; Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 90 (2014) (claiming that AI-based systems "learn from experience and thus improve their performance" over time); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 680 (2017) ("A significant concern about automated decision-making is that it may simultaneously systematize and conceal discrimination. Because it can be difficult to predict the effects of a rule in advance (especially for large, complicated rules or rules that are machine-derived from data), regulators and observers may be unable to tell that a rule has discriminatory effects.").

process.[67] The dynamic nature of AI-driven systems makes them unpredictable and difficult to monitor. In fact, even successful attempts to perform retrospective and independent oversight essentially becomes a form of "whack-a-mole," providing only partial insights into how the system works.[68]

Surely, if the public cannot understand the AI decision-making process, it is unable to identify misbehaviors, such as unfair, discriminatory, and derogatory practices that may be the result of tainted training data or biased algorithms.[69] The ability of the public to utilize market forces to penalize or otherwise affect private entities' behavior is hence limited. Furthermore, unless the public has access to significant monetary and/or legal means to dispute erroneous or unfair decisions and cause their correction, oversight cannot be meaningful.[70]

These factors, in conjunction with the fact that automated decision-making systems may produce discriminatory or biased outcomes,[71] could undermine public trust and confidence in AI,[72] thereby threatening all of its potential benefits.[73] Nevertheless, a carefully constructed accountability mechanism of public scrutiny should be able to mitigate these risks. The use

---

[67] For instance, AI-based systems do not instinctively know whether a specific content is offensive or otherwise unwanted. The system requires large amounts of training data to make such a distinction. Based on this training data, the system gradually learns to distinguish between suitable and offensive content. *See* Monika Bickert & Brian Fishman, *Hard Questions: How We Counter Terrorism*, FACEBOOK NEWSROOM (June 15, 2017), https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/ [https://perma.cc/5PHR-D26R].

[68] *See* Perel & Elkin-Koren, *supra* note 46, at 519; Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 599 (2018); *see also* Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L. 235, 246 (2011); Barb Darrow, *How Hackers Broke into U.S. Voting Machines in Less than 2 Hours,* FORTUNE (July 31, 2017), http://fortune.com/2017/07/31/defcon-hackers-us-voting-machines/ [https://perma.cc/74LZ-AGYT].

[69] *See* Dastin, *supra* note 16; New & Castro, *supra* note 11; PASQUALE, *supra* note 11; Yeung et al., *supra* note 11.

[70] *See generally* Perel & Elkin-Koren, *supra* note 46.

[71] *See* Dastin, *supra* note 16; *see also* Levendowski, *supra* note 68; Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271–72 (2008); Charles Lane, *Will Using Artificial Intelligence to Make Loans Trade One Kind of Bias for Another?*, NPR: MORNING EDITION (Mar. 31, 2017, 5:06 AM), https://www.npr.org/sections/alltechconsidered/2017/03/31/521946210/will-using-artificial-intelligence-to-make-loans-trade-one-kind-of-bias-for-anot [https://perma.cc/WCA9-34DU]; Jeff Larson et al., *Breaking the Black Box: How Machines Learn to Be Racist*, PROPUBLICA (Oct. 19, 2016), https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?word=trump [https://perma.cc/3TR6-M25U]; Ted Greenwald, *How AI Is Transforming the Workplace*, WALL ST. J. (Mar. 10, 2017), https://www.wsj.com/articles/how-ai-is-transforming-the-workplace-1489371060 [https://perma.cc/BM3M-NRV3]; Angwin et al., *supra* note 13; Cathy O'Neil, *I'll Stop Calling Algorithms Racist when You Stop Anthropomorphizing AI*, MATHBABE (Apr. 7, 2016), https://mathbabe.org/2016/04/07/ill-stopcalling-algorithms-racist-when-you-stop-anthropomorphizing-ai/ [https://perma.cc/JN5L-L4L4].

[72] *See* Russell T. Vought., Acting Dir., Office of Mgmt. & Budget, Draft Memorandum to the Heads of Executive Departments and Agencies on Guidance for Regulation of Artificial Intelligence Applications (Jan. 2020), https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf [https://perma.cc/QC69-WRH5].

[73] *See* Levendowski, *supra* note 68.

of impact assessments as a means to achieve accountability is a growing trend in AI oversight policy. In Part III, we turn to examine whether these policy tools can facilitate meaningful oversight.

III. Enhancing Accountability in AI-Based Governance: Impact Assessments

Notwithstanding the challenges discussed in Part II.B, subjecting AI-based decision-making systems to public scrutiny remains an important goal and tool for fostering trust.[74] It is no surprise, then, that initial attempts to develop new regulatory frameworks to promote public scrutiny have recently emerged. The two most high-profile recent examples of this trend are the proposed Algorithmic Accountability Act[75] and the E.U.'s broad-reaching GDPR.[76] Common to these frameworks is the demand that entities deploying AI-based judgments conduct an impact assessment.[77]

Although the Algorithmic Accountability Act has yet to become a binding law, with the rapid advances in AI, it is important to place in context the advantages and disadvantages of impact assessment as a tool to maintain oversight over AI-based systems.

A. Impact Assessments

Generally, an impact assessment can be defined as "the process of identifying the future consequences of current or proposed action."[78] A key advantage of impact assessments of AI-driven systems is their ability to influence entities' internal organizational conduct. By requiring an entity to conduct an internal inspection, impact assessments urge coders and designers to conduct a deeper form of analysis, carefully investigating plausible areas of bias, error, and uncertainty, as well as implementing the necessary

---

[74] *See* Janelle Berscheid & Francois Roewer-Despres, *Beyond Transparency: A Proposed Framework for Accountability in Decision-Making AI Systems*, 5 AI Matters 13, 15 (2019).

[75] Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019) (allowing the FTC up to two years to promulgate regulations in accordance with 5 U.S.C. § 553); *see also* Adi Robertson, *A New Bill Would Force Companies to Check Their Algorithms for Bias*, Verge (Apr. 10, 2019, 3:52 PM), https://www.theverge.com/2019/4/10/18304960/congress-algorithmic-accountability-act-wyden-clarke-booker-bill-introduced-house-senate [https://perma.cc/2G44-AQ5M].

[76] *See* Commission Regulation 2016/679, *supra* note 24.

[77] The notion of impact assessment has been promulgated in a variety of areas. *See* Katyal, *supra* note 10, at 112; Dillon Reisman Et Al., AI Now Inst., Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability 5 (2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/FNJ9-D9UY]; Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 Ga. L. Rev. 109, 119 (2017); Org. for Econ. Cooperation & Dev., What Is Impact Assessment?, https://www.oecd.org/sti/inno/What-is-impact-assessment-OECDImpact.pdf [https://perma.cc/757F-7QML].

[78] IAIA, *supra* note 25.

steps to correct them.[79] The internal and flexible nature of impact assessments shifts the regulated entity's focus away from mere compliance and towards problem solving and improvement.

The idea of a rigorous, standardized process in the form of an impact assessment as a tool to facilitate public accountability and oversight is not new.[80] For instance, many jurisdictions require an environmental impact assessment ("EIA") to evaluate the effects of a proposed project and its alternatives on the environment.[81] EIAs are considered powerful tools for assessing projects' environmental impacts.[82] Consequently, several scholars and policymakers have suggested adopting the impact assessment model in other contexts.[83] However, only recently has the concept of impact assessments drawn the attention of interest groups, scholars, and policymakers with regards to the use of AI in automated decision-making systems.[84] The following discussion introduces these two novel legislative initiatives—i.e., the Algorithmic Accountability Act and the GDPR—along with the central aspects of each initiative, laying the foundation for the consideration of impact assessments as a tool to provide oversight and accountability of AI use.

---

[79] *See* Katyal, *supra* note 10, at 112; Nicholas Diakopoulos et al., *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*, FAT/ML, https://www.fatml.org/resources/principles-for-accountable-algorithms [https://perma.cc/SU3H-DQ37].

[80] *See* Sibout Nooteboom & Geert Teisman, *Sustainable Development: Impact Assessment in the Age of Networking*, 5 J. Env't Pol'y & Plan. 285, 289 (2004). This is particularly true in the areas of human rights, environmental, privacy, and data protection. *See* Reisman et al., *supra* note 77; Yeung et al., *supra* note 11, at 10.

[81] *See* Leonard Ortolano & Anne Shepherd, *Environmental Impact Assessment: Challenges and Opportunities*, 13 Impact Assessment 3, 3 (1995); Erika L. Preiss, *The International Obligation to Conduct an Environmental Impact Assessment: The ICJ Case Concerning the Gabcikovo-Nagymaros Project*, 7 N.Y.U. Envtl. L.J. 307, 313 (1999); Matthew Cashmore et al., *The Interminable Issue of Effectiveness: Substantive Purposes, Outcomes and Research Challenges in the Advancement of Environmental Impact Assessment Theory*, 22 Impact Assessment & Project Appraisal 295, 295–96 (2004); Jie Zhang et al., *Critical Factors for EIA Implementation: Literature Review and Research Options,* 114 J. Envtl. Mgmt. 148, 151 (2012); Douglas C. Baker & James N. McLelland, *Evaluating the Effectiveness of British Columbia's Environmental Assessment Process for First Nations' Participation in Mining Development*, 23 Envtl. Impact Assessment Rev. 581, 582–83 (2003); Matthew J. Rowe et al., *Accountability or Merely "Good Words"? An Analysis of Tribal Consultation Under the National Environmental Policy Act and the National Historic Preservation Act*, 8 Ariz. J. Envtl. L. & Pol'y 1, 47 (2018).

[82] These assessments present information to the public and decision makers about potential negative environmental impacts. *See* Jameson Tweedie, *Transboundary Environmental Impact Assessment Under the North American Free Trade Agreement,* 63 Wash. & Lee L. Rev. 849, 860–61 (2006).

[83] *See, e.g.*, Michael Froomkin, *Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements*, 2015 U. Ill. L. Rev. 1713, 1745–57 (2015); Selbst, *supra* note 77, at 171; Alessandro Mantelero, *AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment*, 34 Computer L. & Security Rev. 754, 760 (2018); Katyal, *supra* note 10.

[84] *See, e.g.*, Reisman et al., *supra* note 77, at 6–8; Eddie Copeland, *10 Principles for Public Sector Use of Algorithmic Decision Making*, NESTA (Feb. 20, 2018) https://www.nesta.org.uk/blog/10-principles-for-public-sector-use-of-algorithmic-decision-making/ [https://perma.cc/2PH7-SX65]; Froomkin, *supra* note 83; Katyal, *supra* note 10, at 112–14.

1. *Algorithmic Accountability Act*

In 2019, several members of Congress introduced the Algorithmic Accountability Act.[85] The members introduced the bill following various reports of automated decision-making systems leading to undesirable consequences.[86] As noted by Senator Cory Booker, one of the sponsors of the proposed law, "This bill requires companies to regularly evaluate their tools for accuracy, fairness, bias, and discrimination. It's a key step toward ensuring more accountability from the entities using software to make decisions that can change lives."[87] If passed in its current form, the Algorithmic Accountability Act will require covered entities that use any automated decision-making system to conduct data protection impact assessments ("DPIAs") and automated decisions system impact assessments ("ADSIAs").[88]

Specifically, covered entities[89] could be required to submit a DPIA of existing and new high-risk information systems.[90] The evaluation is to be conducted in consultation with external third parties, if possible.[91] Publication of the assessment, however, is not mandatory and is left to the discretion of the covered entity.[92]

ADSIAs focus on a system's "development process, including the design and training data of the automated decision system, for impacts on accuracy, fairness, bias, discrimination, privacy, and security."[93] ADSIAs must contain the following parameters: a description of the system, an assessment of the relative benefits and costs of the system in light of its purpose, an assessment of the risk posed by the system, and measures that can be taken to mitigate risk.[94] Although ADSIA guidelines increase transparency, the publication of the assessment results is left to the discretion of the con-

---

[85] Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019).

[86] Press Release, Sen. Ron Wyden, Sen. Cory Booker & Rep. Yvette D. Clarke, Wyden, Booker, Clarke Introduce Bill Requiring Companies to Target Bias in Corporate Algorithms (Apr. 10, 2019), https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms [https://perma.cc/S828-GX4E].

[87] *Id.*

[88] The term "automated decision system" is defined as "a computational process, including one derived from machine learning, statistics, or other data processing or artificial intelligence techniques, that makes a decision of facilitates human decision making, that impacts consumers." H.R. 2231.

[89] Covered entities include those that (1) had greater than $50,000,000 in average annual gross receipts over the last three years; (2) possess or control information on more than one million consumers or consumer devices; and (3) collect personal information on their users. *Id.*

[90] *See id.* § 3(b)(1)(B). DPIA is defined as an evaluation of "the extent to which an information system protects the privacy and security of personal information the system processes." *Id.* § 2(6).

[91] *Id.* § 3(b)(1)(C).

[92] *Id.* § 3(b)(2).

[93] *See id.* § 2(2).

[94] *Id.*

ducting entity. This is why ADSIAs' ability to foster public review is tempered, as will be discussed in further detail in the next part of this Article.[95]

One of the most powerful and revolutionary aspects of these impact-assessment requirements is the extensive enforcement powers given to the FTC.[96] The Algorithmic Accountability Act would empower the FTC to issue and enforce regulations that would require covered entities to complete impact assessments and address the results of the impact assessments in a timely manner.[97] Additionally, the FTC would have authority under the bill to enforce compliance; specifically, the bill provides that any violations would be treated as "an unfair or deceptive act or practice under section 18(a)(1)(B) of the Federal Trade Commission Act."[98] Further, it confers on the FTC the power to enforce compliance "in the same manner, by the same means, and with the same jurisdiction, powers, and duties as though all applicable terms and provisions of the Federal Trade Commission Act . . . were incorporated into [this Bill]."[99]

Supplementing the FTC's enforcement powers, the Algorithmic Accountability Act would authorize the different state attorneys general, as *parens patriae*, to bring a civil action against an entity in violation.[100] It would not, however, allow for private enforcement. The bill also allows for actions to be brought forward by other state officials; specifically, "any other officer of a State who is authorized by the State to do so may bring a civil action [in the same manner as the State's attorney general]."[101] These mechanisms can be regarded as a means to oversee the regulated entities' decision-making processes.

2. *General Data Protection Regulation*

The objective of the EU's GDPR is to give individuals more control over their personal data, and it has come to be regarded as a global gold

---

[95] For further discussion, see *infra* Part III.B.

[96] The FTC is an independent U.S. law enforcement agency responsible for protecting consumers and competition. *See* New & Castro, *supra* note 11, at 13–18; Andrew Tutt, *An FDA for Algorithms*, 69 Admin. L. Rev. 83, 115 (2017). For a similar discussion, see Ryan Calo, Brookings, The Case for a Federal Robotics Commission 14 (2014), https://www.brookings.edu/wp-content/uploads/2014/09/RoboticsCommissionR2_Calo.pdf [https://perma.cc/QAL8-7R9F]; Jack M. Balkin, *The Path of Robotics Law*, 6 Cal. L. Rev. Cir. 45, 50 (2015). The FTC's primary legal authority comes from Section 5 of the Federal Trade Commission Act, which prohibits unfair or deceptive practices in the marketplace. *See About the FTC*, Fed. Trade Commission, https://www.ftc.gov/about-ftc [https://perma.cc/AT46-9739]; Fed. Trade Comm'n, Privacy & Data Security: Update: 2017, at 1 (2017), https://www.ftc.gov/system/files/documents/reports/privacy-data-security-update-2017-overview-commissions-enforcement-policy-initiatives-consumer/privacy_and_data_security_update_2017.pdf [https://perma.cc/ND88-Y6LZ].

[97] H.R. 2231.

[98] *Id.* § 3(d).

[99] *Id.* § 3(d)(2)(A).

[100] *Id.* § 3(e)(1) (authorizing such entities to "bring a civil action on behalf of the residents of the State in an appropriate district court of the United States to obtain appropriate relief").

[101] *Id.* § 3(e)(5)(A).

standard for privacy regulation.[102] The regulation also has several important provisions pertaining to automated decision-making.

In particular, the GDPR states that, as a rule, there is a prohibition on fully automated individual decision-making, including profiling that has a legal, or similar, effect on the individual.[103] The regulation provides for three exceptions: (1) if the decision is necessary for performing or entering into a contract; (2) if the decision is authorized by Union or Member State law to which the data controller is subject and that lays down suitable measures to safeguard the data subject's rights, freedoms, and legitimate interests; or (3) if the decision is based on the data subject's explicit consent.[104] When one of those exceptions applies, the data controller must implement suitable measures with which to safeguard the individual's (i.e., data subject's) rights, freedoms, and legitimate interests. These measures should include "*at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision*."[105] So, the GDPR provides individuals with a "right to explanation."[106] However, the right to

---

[102] *See* Commission Regulation 2016/679, *supra* note 24.

[103] *See id.* at art. 22(1) ("The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."); *see also*, Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 145, 179 (2019) ("[T]the GDPR's "right to explanation" is no mere remedial mechanism to be invoked by data subjects on an individual basis, but it implies a more general form of oversight with broad implications for the design, prototyping, field testing, and deployment of data processing systems."). *See generally* NICK WALLACE & DANIEL CASTRO, CTR. FOR DATA INNOVATION, The Impact of the EU's New General Data Protection Regulation on AI (2018).

[104] *See* Commission Regulation 2016/679, *supra* note 24, at art. 22(2). *See also The European Commission's Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, COM (2018) final (Feb. 6, 2018) [hereinafter *European Commission's Guidelines*].

[105] Commission Regulation 2016/679, *supra* note 24, at art. 22(3) (emphasis added). Recital 71 adds to this, stating, "In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision." Commission Regulation 2016/679, *supra* note 24, recital 71.

[106] *See, e.g.*, Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233, 235 (2017); Bryce Goodman & Seth Flaxman, *EU Regulations on Algorithmic Decision-Making and a "Right to Explanation,"* 38 AI MAG. 50, 51 (2017); Citizens' Rights and Constitutional Affairs, Artificial Intelligence: Potential Benefits and Ethical Considerations, EUR. PARL. DOC. PE 571.380 (2016); *Guide to the General Data Protection Regulation (GDPR)*, INFO. COMMISSIONER'S OFF. 147 (May 22, 2019), https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf [https://perma.cc/8ZV4-JB2Y]; Report with Recommendations to the Commission on Civil Law Rules on Robotics, EUR. PARL. DOC. PE 582.443v03-00 (2017). *But see* Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 624 (2020); Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 77 (2017) (claiming that the GDPR's right of access allows for a limited right to explanation of the functionality of automated decision-making systems—what they refer to as the "right to be informed").

explanation is limited to circumstances where a decision is *based solely on automated processing*.[107] An additional clause of the right to explanation provides data subjects with the right to receive notice of solely automated decision-making processes and to request access to meaningful information.[108]

However, even when an individual fails to invoke any of these rights, the GDPR will still establish and enforce accountability through an array of tools, including mandatory DPIAs.[109] The GDPR requires data controllers to carry out a DPIA on any type of processing that is likely to result in "high risk" to an individual's rights and freedoms prior to adoption.[110] This is particularly the case when the data controller uses systematic and extensive evaluation of "personal aspects relating to natural persons *which is based on automated processing*, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person."[111]

The GDPR further requires DPIAs to include the following elements: (1) a systematic description of the envisaged processing and its purpose, including the legitimate interest pursued by the data controller; (2) an assessment of the necessity and proportionality of processing in relation to the identified purposes; and (3) an assessment of the risks to the rights and freedoms of data subjects.[112] Hence, the GDPR also aims to enhance oversight by requiring regulated entities to perform a form of impact assessment.

---

[107] The European Commission's Guidelines on Automated Individual Decisions stated that Article 22 applies only where there is "no human involvement in the decision process." *European Commission's Guidelines*, *supra* note 104.

[108] *See* Commission Regulation 2016/679, *supra* note 24, at arts. 13–14, recital 60; *European Commission's Guidelines*, *supra* note 104. For further discussion of the potential effects of the GDPR on AI-based decision-making systems, see generally Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation* (Berkman Klein Ctr. for Internet & Soc'y, Working Paper, 2017); Casey et al., *supra* note 103; Selbst & Barocas, *supra* note 57. Importantly, scholars, policymakers, and industry leaders have been debating what the GDPR's new "right to explanation" entails. *See, e.g.*, sources cited *supra* note 106.

[109] Margot E. Kaminski & Gianclaudio Malgieri, *Algorithmic Impact Assessments Under the GDPR: Producing Multi-Layered Explanations* 3, 7 (Univ. Colo. Law Sch. Legal Studies, Working Paper No. 19-28, 2019) (arguing that the DPIA is best understood as a nexus between the GDPR's two approaches to algorithmic accountability: individual rights and collaborative governance).

[110] *See* Commission Regulation 2016/679, *supra* note 24, at art. 35; *Guide to the General Data Protection Regulation (GDPR)*, *supra* note 106, at 83. Accordingly, DPIAs are mandatory when, "taking into account the nature, scope, context and purposes of the processing," "a high risk to the rights and freedoms of natural persons" "is likely to result." Commission Regulation 2016/679, *supra* note 24, at art. 35; *see also Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679*, COM (2017) 5 final (Apr. 4, 2017).

[111] Commission Regulation 2016/679, *supra* note 24, at art. 35(3)(a) (emphasis added). The GDPR further states that, in order to help data controllers ascertain whether processing is likely to present high risks, supervisory authorities will maintain a list of processing operations which are subject to the requirement for DPIA, or for which no impact assessment is required. *See id.* art. 35(4)–(5).

[112] Commission Regulation 2016/679, *supra* note 24, at art. 35.

Like the Algorithmic Accountability Act, the relevant entity must prepare the DPIA before beginning processing activity. Nonetheless, a publication of the DPIA under the GDPR is optional and is generally left to the discretion of the entity conducting the impact assessment,[113] possibly hindering public access to important information.

The GDPR vests extensive enforcement powers with the relevant regulatory agencies.[114] These independent[115] and competent[116] agencies are bestowed with broad monitoring, advisory, and investigatory powers,[117] including the power (1) to issue warnings or reprimands to a data controller or processor;[118] (2) to order data controllers or processors to comply with a data subject's request to exercise his or her rights (e.g., the right to explanation);[119] and (3) to impose fines.[120] More than that, the GDPR also explicitly states that each Member State shall provide supervisory authority with the power to bring any infringements of this Regulation to the attention of judicial authorities and commence legal proceedings to enforce the provisions of the GDPR.[121]

In sum, both the Algorithmic Accountability Act and the EU's GDPR are examples of recent attempts to harness the advantages of impact assessments to achieve meaningful oversight for AI and other automated decision-making systems. These initiatives, however, are not sector-specific. Rather, they target automated decision-making and other AI-based systems across the board. The problem is that automated decision-making systems are not homogenous. Hence, given the fact that a single regulatory scheme may not be flexible enough to properly seize the multiplicity of AI systems, it might be difficult for these initiatives to properly regulate all possible sectors with one set of rules. Particularly, it may prove difficult for policymakers to consider the necessary level of oversight based on factors such as the regulated activity, public policy objectives, and externalities. Therefore, as further discussed below in certain areas such as content moderation, policymakers must adopt domain-specific legislation to ensure oversight.

---

[113] *Id.* at recital 90; *see also* REISMAN ET AL., *supra* note 77, at 7; Katyal, *supra* note 10, at 115.

[114] Chapter VI of the GDPR focuses on independent supervisory authorities. *See* Commission Regulation 2016/679, *supra* note 24, at ch. VI. Principally, Chapter VI of the GDPR provides that each Member State shall appoint independent supervisory authorities to be responsible for monitoring the application of the GDPR. *Id.* art. 51.

[115] *Id.* art. 52.

[116] *Id.* art. 53–56.

[117] *Id.* art. 57–58; *see also* Casey et al., *supra* note 103, at 165.

[118] Commission Regulation 2016/679, *supra* note 24, art. 58(2)(a)–(b).

[119] *Id.* art. 58(2)(c).

[120] *Id.* art. 58(2)(f), (i).

[121] *Id.* art. 58(5).

B. Impact Assessments and Public Oversight

To evaluate the merits of impact assessments as a tool for enhancing the public scrutiny of AI-driven systems, we will analyze them through the lens of a previously introduced accountability model.[122] The accountability model identifies three proxies for the public's ability to understand automated decision-making systems, to challenge those systems, and to correct improper decisions.[123] These proxies are transparency,[124] due process,[125] and public review.[126] First, the ability of the public to understand how AI-based decision-making is implemented depends on knowledge of the subject.[127] "Without knowing that specific conduct took place, it is impossible for the public to render judgment on the merits of such conduct."[128] Second, the ability of the public to contest decisions made by AI systems also depends on the availability of different measures of procedural due process, such as adequate notification and an opportunity to be heard.[129] Third, for the public to be able to correct errors made by AI-based systems of decision-making, it is necessary to have sufficient mechanisms for public sanctions and corrections.[130]

As we demonstrate below, although they offer many benefits, impact assessments are still insufficient in facilitating accountability under the proxy test because they (1) provide only limited transparency; (2) fall short of securing due process; and (3) provide for inadequate public scrutiny.

1. *Transparency*

A major advantage of impact assessments—consisting of a study evaluating a system's development process, including the design and the training data thereof[131]—is that they actively promote information disclosure in key areas, therefore rendering the decision-making process less opaque.[132] With-

---

[122] Perel & Elkin-Koren, *supra* note 46, at 493–95.
[123] *Id.* at 493–96.
[124] *Id.* at 494.
[125] *Id.* at 495; *see also* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 20 (2014).
[126] Perel & Elkin-Koren, *supra* note 46, at 496.
[127] *Id.* at 495.
[128] *Id.*
[129] *Id.* at 495–96.
[130] *Id.* at 496.
[131] Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019).
[132] In order to illustrate this point, consider environmental law, where impact assessments are often used as a tool for addressing the effects of technology and the possible harm caused by it to the public. *See* Kathleen Waugh & Gary E. Marchant, *Collaborative Voluntary Programs: Lessons from Environmental Law*, in THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM 183, 184 (Gary E. Marchant et al. eds., 2011). The environmental-impact-assessment model shows that an impact assessment can guarantee that critical information will be made available to a larger audience, thus rendering the decision-making process less opaque. *See* Selbst, *supra* note 77, at 169; Robertson v. Methow Valley Citizens Council, 490 U.S. 332, 349 (1989); Kleppe v. Sierra Club, 427 U.S. 390, 409 (1976).

out impact assessments, governmental agencies—and the public—are not likely to be privy to such information. In other words, an impact assessment could lead to greater transparency.[133] Nevertheless, the levels of transparency possibly generated by the impact assessments structured under the Algorithmic Accountability Act or GDPR are still lacking. First, these mechanisms fail to respond adequately to the dynamic nature of AI-based systems. Second, they do not establish mandatory publication requirements.

As previously noted, machine learning enables the system to constantly improve and adjust in response to a user's patterns and the information that is fed back into the system.[134] This characteristic also makes it much more difficult for third parties to obtain information regarding which factors are being considered and how those factors are weighted. However, for the most part, under the initiatives surveyed throughout this paper, regulated entities are usually required to submit an impact assessment before the implementation of the decision-making system.[135]

Examining a system before it is implemented could theoretically provide meaningful insights into how some algorithms operate and could certainly help to detect faulty nodes or other inherent issues with the algorithm.[136] However, even if the algorithm is free of defects, if the data fed back into the system present a distorted or biased picture or perpetuate human biases against certain groups, then the outcome of the automated decision-making system will still be biased or discriminatory.[137] An ex ante evaluation of the system *alone* will not identify a problem.[138] To illustrate, consider the following hypothetical: a company has recently implemented a new credit scoring system that is based on a machine-learning algorithm, using previously underutilized, internal, and legitimate third-party data to perform smarter credit scoring. The company evaluated the system before implementation and submitted an impact assessment to the relevant agency, indicating its benefits, costs, and risks. During the training period, the company fed the system with data previously made by the company's human

---

[133] Michael Froomkin argues that "requiring those conducting mass surveillance in and through public spaces to disclose their plans publicly via an updated form of environmental impact statement" could be beneficial and plausibly trigger a more informed public debate about privacy and the trade-offs between privacy and other values. *See* Froomkin, *supra* note 83, at 1713.

[134] *See* Surden, *supra* note 66; Bickert & Fishman, *supra* note 67.

[135] The Algorithmic Accountability Act gives special attention to "high-risk automated decisions systems." H.R. 2231 § 2(7).

[136] *See* Huq, *supra* note 106, at 642.

[137] *See* Kroll et al., *supra* note 66, at 680–81; Curt Levey & Ryan Hagemann, *Algorithms with Minds of Their Own*, WALL ST. J. (Nov. 12, 2017) https://www.wsj.com/articles/algorithms-with-minds-of-their-own-1510521093 [https://perma.cc/S3XF-7H2M]; *see also* Ctr. for Data Innovation, *supra* note 65.

[138] Scholars commonly differentiate between ex ante and ex post approaches to oversight and accountability. The ex ante approach generally aims at determining whether the automated decision-making process works as expected. The ex post approach is designed to support review and oversight once a decision has been made. *See* Kroll et al., *supra* note 66, at 637.

employees.[139] Clearly, if the training data presents a distorted or biased picture against a certain group of individuals, then the result generated by the scoring system will be erroneous.[140] For instance, in our hypothetical example, if the vast majority of higher income individuals turn out to be male, while almost none of them were women, the overrepresentation of males in this sample may affect the learning of the algorithm. Although the algorithm itself is not faulty, discriminatory, or biased, the AI-based system might learn to discriminate against women.[141] Thus, the AI system could amplify and perpetuate bias.[142]

In other words, the use of flawed training data can prolong stereotypes, biases, and discriminatory practices[143] in a way that is very difficult to mitigate by current impact-assessment schemes. Furthermore, one of the main characteristics of an AI system is its complexity and ability to learn and evolve;[144] therefore, it is almost impossible to predict what transpires as part of the system's decision-making process ex ante.[145] Accordingly, relying predominantly on ex ante assessments is insufficient to foster transparency since it fails to adequately take into account the dynamic nature of AI-based systems.[146]

---

[139] Practically speaking, although the algorithm might initially comply with its designated aim, the ultimate results will depend on the training data and information being fed back into the system. *See* Robert D. Atkinson, *"It's Going to Kill Us!" and Other Myths About the Future of Artificial Intelligence*, INFO. TECH. & INNOVATION FOUND. 26 (June 2016), http://www2.itif.org/2016-myths-machine-learning.pdf?_ga=2.97197983.1475650685.1601830190-1439281446.1600616886 [https://perma.cc/A6D6-378F]. *See generally* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018).

[140] *See* Kroll et al., *supra* note 66, at 680–81.

[141] *See, e.g.*, Taylor Telford, *Apple Card Algorithm Sparks Gender Bias Allegations Against Goldman Sachs*, WASH. POST (Nov. 11, 2019, 10:44 AM), https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/ [https://perma.cc/AM8T-PH8M].

[142] *See* Hamilton, *supra* note 16.

[143] *See* Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 192 (2019); *Optimizing for Engagement: Understanding the Use of Persuasive Technology on Internet Platforms: Hearing Before the Subcomm. on Commc'n, Tech., Innovation and the Internet of the S. Comm. on Commerce, Sci., & Transp.*, 116th Cong. (2019) (statement of Rashida Richardson, Director of Policy Research, AI Now Institute, New York University).

[144] *See supra* notes 66–67.

[145] *See, e.g.*, Silvan Jongerius, *Artificial Intelligence and the Right to Explanation Under the GDPR*, TECH GDPR (Mar. 30, 2018), https://techgdpr.com/blog/artificial-intelligence-right-to-information-explanation [https://perma.cc/HYH8-4UGV]. Further, as above, when a group of researchers at Mount Sinai Hospital in New York employed a system of deep learning to the hospital's database, the resulting program ("Deep Patient") was proven to be extremely good at predicting diseases, including psychiatric disorders like schizophrenia. However, even its own designers do not know how. *See* Knight, *supra* note 64.

[146] It is not clear what "prior to implementation" entails. Does it require assessment of the algorithm alone? Or, does it perhaps require an assessment of the system after the training data has been fed into the system, but before it has been implemented into a real-world scenario? No matter which approach one takes, requiring the assessment to take place before implementation suggests that the analysis does not account for the real-world impact of the system. This is, without a doubt, a fundamental shortcoming, given the dynamic nature of AI systems.

Furthermore, while impact assessments are presumably meant to provide information about the impacts of possible actions with the aim of improving decision-making about these actions, the Algorithmic Accountability Act and the GDPR do not require regulated entities to publicize the results of their impact assessments.[147] Hence, at least with respect to members of the general public, a genuine transparency problem is embedded within the framework of these mechanisms of impact assessment. This lack of genuine transparency deprives the public of an opportunity to play a meaningful role in the decision-making process, to share information, to give the moderating entity feedback, and to comment about possible ramifications of the system.[148]

2. *Due Process*

Another important aspect of accountability is procedural due process.[149] Automated decision-making systems often combine individual rulemaking with adjudication,[150] which means that an erroneous decision can be the result of an invalid policy, incorrect adjudication, or both.[151]

This is why the literature discussing automated decision-making systems suggests incorporating due process safeguards, such as notice and the right to be heard, into the use of those systems.[152] To illustrate, let us get back to our credit-scoring example. Following the introduction of the new credit-scoring system, a patron of one of the leading banks has been informed that his loan application has been approved but with higher interest rates. He believes that his credit score is inaccurate and might be the result of erroneous data, embedded biases, or even an incorrect rule encoded in the system. To maintain due process, the individual should be provided with information pertaining to his credit score and the data used to calculate it, and he should also have the right to contest incomplete or inaccurate infor-

---

[147] *See also* Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1608 (2019) (claiming that "[t]he most striking gap in public and third-party accountability in the GDPR is its approach to releasing—or not releasing—algorithmic impact assessments. While the GDPR's impact assessments have been heralded as a model for algorithmic accountability, the process does not in fact involve releasing information to the public. A company is merely encouraged, not required, to publicly release its impact assessments."). Although it is important to note that the GDPR requires companies to consult with third parties in conducting impact assessments, it only has to do so where appropriate and while taking into account commercial secrets. *See* Commission Regulation 2016/679, *supra* note 24, art. 35(9).

[148] Selbst, *supra* note 77, at 179–81 ("In the benefits column, they can force government agencies to both think hard about the collateral effects of the proposed policy and justify the policy to the public."); Scherer, *supra* note 10.

[149] Perel & Elkin-Koren, *supra* note 46, at 495.

[150] Citron, *supra* note 71, at 1253.

[151] Citron, *supra* note 71, at 1279.

[152] *See, e.g.*, Kaminski, *supra* note 147, at 1554–55; Citron & Pasquale, *supra* note 125, at 20, 28; Citron, *supra* note 71, at 1305–08; Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 190–200 (2017).

mation.[153] Without those rights, and due to the opacity of the system, an individual would find it hard to determine exactly why the system assigned him a low credit score. More importantly, he would not be able to contest the score or correct erroneous information.

While the GDPR combines a systematic regulatory approach in the form of impact assessments with several individual rights, which are designed to secure due process and the right to be heard,[154] the Algorithmic Accountability Act does not provide members of the public with such a process. Specifically, the Algorithmic Accountability Act secures neither meaningful notice nor an opportunity to be heard.[155] Thus, it fails to adequately safeguard due process. In fact, the black box nature of AI systems may preclude traditional forms of due process, namely notice and hearing.[156] Crawford and Schultz consider the role of a neutral data arbiter that could file complaints and investigate allegations of bias or financial interest that might render the adjudication unfair, under the theme of due process.[157] Accordingly, one could argue that bestowing a government agency with broad monitoring, advisory, and investigatory powers[158] designed to address systematic errors, unfairness, bias, and discrimination is an element of due process. However, vesting oversight powers with a regulatory body, without supplementary commitment to genuine transparency and public scrutiny, as demonstrated throughout the previous section, does not seem to offer adequate challenging opportunities. As stated by Kaminski, "A system of governance through third-party audits, expert boards, government inspection and enforcement, and performance reports might produce better and more legitimate algorithms, but it might still not produce a justificatory system that would be acceptable from the perspective of an individual affected by a particular decision."[159]

---

[153] Citron & Pasquale, *supra* note 125 at 16–17; Kaminski, *supra* note 147, at 1554–55.

[154] *See, e.g.*, Commission Regulation 2016/679, *supra* note 24, art. 13–25 (providing data subjects with the right to receive notice of solely automated decision-making processes and to request access to meaningful information); *see also* European Comm'n, Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (Feb. 6, 2018). For further discussion of the potential effects of the GDPR on AI-based decision-making systems, see Doshi-Velez et al., *supra* note 108. *See also* Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 145 (2019); Selbst & Barocas, *supra* note 57, at 1085.

[155] *See* Citron, *supra* note 71, at 1305–08 (identifying notice and opportunity to be heard as key components of technological due process); Kaminski, *supra* note 147, at 1554; Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 120–21 (2014).

[156] Kaminski, *supra* note 147, at 1592 (arguing that "[t]hese may not be rights to a hearing in a traditional sense, but they give individuals the ability to intervene in data processing—and not just solely automated processing—in ways familiar to those steeped in the algorithmic due process literature").

[157] *See* Crawford & Schultz, *supra* note 155, at 127.

[158] Commission Regulation 2016/679, *supra* note 24, art. 51; Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. § 3(b)(1)(D) (2019).

[159] Kaminski, *supra* note 147, at 1578.

Further, a single, centralized regulatory agency specializing in AI might not have the context-specific knowledge necessary for oversight of the types of decisions that an AI-based system makes.[160] It would require enormous resources to develop this kind of subject-matter expertise. Moreover, even if an agency were able to retain the necessary technological knowledge and expertise, it might not utilize the full extent of its power due to political and practical reasons.[161] Given the public interest in overseeing automated decision-making systems, this is troublesome. For example, the Algorithmic Accountability Act, as noted earlier, vests the FTC with the power to take action when a platform fails to adequately employ its AI-driven content-moderation system. This failure is treated as an unfair practice—that is, an activity that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition."[162] This injury includes harm to one's privacy, as well as more general injury to consumers.

MacCarthy claims that in the content-moderation context, an unfair practice could also be understood as an "unwanted exposure to material distributed by third parties on systems operated by digital platform companies" or a failure to establish and maintain an adequate content-moderation scheme.[163] In those instances, he argues, it is unlikely that the FTC will use its authority because doing so would require the Commission (1) to show that the injury is outweighed by consumer benefits;[164] and (2) to test the limits of its unfairness authority.[165] In light of how it can affect a consumer's conduct, "a platform's failure to disclose key elements of its content moderation program" is a deceptive practice.[166] However, MacCarthy argues that the Commission is not likely to use its authority in connection with a platform's content moderation disclosure practices, mainly because doing so would require the Commission to exercise its power in an area removed

---

[160] Ctr. for Data Innovation, *supra* note 65, at 14 ("If it would be ill-advised to have one government agency regulate all human decision-making, then it would be equally ill-advised to have one agency regulate all algorithmic decision-making."); *see also* Karen Kornbluh & Ellen P. Goodman, *Bringing Truth to the Internet*, 53 DEMOCRACY (2019) ("The Federal Trade Commission (FTC), with its duty to protect consumers and prevent anti-competitive practices, lacks the tools and authorities. Many of today's online challenges—like disinformation—are not only threats to individual consumers, but also systemic threats to the economy and democracy. . . . Staffing is another shortcoming. . . . This staff shortage is compounded by a lack of substantive technical know-how: The FTC currently has just five full-time technologists on staff, and no chief technologist.").

[161] *See* Ctr. for Data Innovation, *supra* note 65, at 15–16.

[162] 15 U.S.C. § 45 (2018).

[163] Mark MacCarthy, *A Consumer Protection Approach to Platform Content Moderation*, *in* FUNDAMENTAL RIGHTS PROTECTION ONLINE: THE FUTURE REGULATION OF INTERMEDIARIES (Bilyana Petkova & Tuomas Ojanen eds., 2020).

[164] *Id.* at 10.

[165] *Id.*

[166] *Id.* at 13.

from its normal deception authority and in a way that is very close to content regulation.[167]

Finally, it is unclear how appropriate it is to cabin the knowledge necessary for oversight of the types of decisions that an AI-based system makes within a single, centralized regulatory agency (e.g., the FTC) because other agencies might require that information.

3. *Public Review*

Another key failure of impact-assessment initiatives is the lack of meaningful public review. As stated earlier, accountability with private bodies can be achieved through informal means, such as market forces that check decision makers' discretion and promote voluntary disclosure in relation to their choices and related outcomes.[168] Indeed, a key advantage of impact assessments is their ability to influence platforms' internal organizational conduct. By requiring a platform to conduct an internal inspection, impact assessments urge coders and designers to conduct a deeper form of analysis, in which they carefully investigate plausible areas of error and uncertainty and implement the necessary steps to correct them.[169] The internal and flexible nature of impact assessments shifts the regulated entity's focus away from mere compliance and towards problem solving and improvement. Moreover, an internal inspection can deter private companies from developing automated decision-making systems that would not withstand public scrutiny and debate.[170] All of this would contribute to the development and implementation of ameliorated AI systems.

In addition, at the most basic level, genuine transparency could plausibly prompt a broader public discussion and greater collaboration between private entities, government officials, and citizens. [171] For instance, government officials and citizens' groups could highlight certain weaknesses of the automated decision-making processes that the private entity may not have considered. This would ultimately lead to improved AI systems, enable the correction of erroneous decisions, and reduce deleterious effects to consumers. However, as shown above, current impact-assessment mechanisms provide only very limited transparency[172] and fall short of securing due process.

---

[167] *See id.* at 15–16.

[168] *See* Elkin-Koren & Perel, *supra* note 37, at 482.

[169] *See supra* notes 79–80.

[170] *See, e.g.*, Robert G. Dreher, *NEPA Under Siege: The Political Assault on the National Environmental Policy Act*, GEO. ENVTL. L. & POL'Y INST. 6 (2005) (discussing the National Environmental Policy Act), https://www.sierraforestlegacy.org/Resources/Conservation/LawsPoliciesRegulation/ForestPlanningRegulations/NEPA/NEPA-UnderSiege.pdf [https://perma.cc/TMH5-AGP7].

[171] Froomkin, *supra* note 83.

[172] Since regulated entities are required to evaluate how an automated system is designed and used, the risks it poses, as well as other factors. Nevertheless, they are not required to disclose these impact assessments. *See* Perel & Elkin-Koren, *supra* note 46, at 482. *See generally* PASQUALE, *supra* note 11; Diakopoulos, *supra* note 56.

Consequently, they fail to facilitate sufficient public oversight and proper opportunities for correcting erroneous decisions.

To summarize, while impact assessments demonstrate an important and laudable attempt to generate better oversight of systems that deploy AI-based decision-making, they still suffer from several flaws. To improve their oversight potential, we thus recommend a number of improvements to the existing impact-assessment schemes, including periodic impact assessments, mandatory notice-and-comment procedure, and mandatory publication. In Part V, we describe in more detail our recommendations for addressing these flaws. Next, though, we turn to examine the merits of impact assessments for generating public oversight of AI-based content-moderation systems.

IV. The Case of Content Moderation by AI

Content moderation can be defined as the organized practice of screening online content based on the characteristics of the website, its targeted audience, and jurisdictions of user-generated content to determine whether such content is appropriate.[173] In the past, human moderators mostly performed content moderation. The human moderator had to screen each and every post and determine whether or not it was compliant with the company's guidelines in order to decide whether or not it should be removed.[174] The reviewer could be a moderator working for the platform tasked to review uploaded content, a user who flags specific content as inappropriate,[175] or any other trusted notifier.[176]

With the growth in the amount of content posted online, as well as the public and regulatory pressure on platforms to protect users[177] and expedi-

---

[173] Sarah T. Roberts, *Content Moderation, in* Encyclopedia of Big Data (Laurie A. Schintler & Connie L. McNeely eds., 2018). *But see* James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech. 42, 47 (2015) (defining "moderation" as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse").

[174] Brittan Heller, *What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech*, WIRED (May 2, 2018, 8:00 AM), https://www.wired.com/story/what-mark-zuckerberg-gets-wrongand-rightabout-hate-speech/ [https://perma.cc/7VDY-KRBG].

[175] A user may be motivated to engage in moderation activities either to maintain status or prestige or for monetary compensation in the form of reduced fees. *See* Hector Postigo, *Emerging Sources of Labor on the Internet: The Case of America Online Volunteers*, 48 Int'l Rev. Soc. Hist. 205, 207 (2003); Alistair Barr & Lisa Fleisher, *YouTube Enlists 'Trusted Flaggers' to Police Videos*, Wall St. J. (Mar. 17, 2014), http://blogs.wsj.com/digits/2014/03/17/youtube-enlists-trusted-flaggers-to-police-videos [https://perma.cc/HR5C-Z59V]; Roberts, *supra* note 173; *Report Inappropriate Content*, YouTube, https://support.google.com/youtube/answer/2802027?co=GENIE.Platform%3DAndroid&hl=EN-GB [https://perma.cc/K63S-HR32].

[176] *See generally* Sebastian Felix Schwemer, *Trusted Notifiers and the Privatization of Online Enforcement*, 35 Comput. L. & Sec. Rev., Nov. 2019, at 1.

[177] One of the most often discussed issues is the problem of "fake news" and misinformation. There is, however, also vast criticism regarding how platforms are dealing with other unwanted content, such as hate or abusive content, extremism and terrorist propaganda, and unauthorized copyrighted and otherwise illegal content. *See generally* Niva Elkin-Koren & Maayan Perel, *Guarding the Guardians: Content Moderation by Online Intermediaries and*

tiously remove illicit content, it has become almost impossible for online platforms to rely exclusively on human reviewers.[178] In particular, platforms were recently called to fight misinformation in relation to the COVID-19 health crisis[179] and address hate speech that ignites racial tensions.[180] In relation to terrorist content, Facebook recently admitted that ninety-nine percent of the terrorist content they remove is flagged by their AI-based systems before anyone on their services reports it. Likewise, YouTube has announced that it is using AI to spot extremist content and that more than eighty-three percent of the videos it deleted (three-quarters of which were deleted before they received any views) were flagged by AI.[181] Recently, during the COVID-19 pandemic, major social media platforms, including Facebook,[182] YouTube,[183] and Twitter[184] announced they would shift their content modera-

---

the Rule of Law, *in* OXFORD HANDBOOK OF INTERMEDIARY LIABILITY ONLINE (Giancarlo Frosio ed., 2020). Examples include recent calls for Twitter, Facebook and Google to do more to fight disinformation on the internet. *See* Stephanie Bodoni & Marie Mawad, *EU Renews Calls to Facebook, Twitter to Fight Fake News*, BLOOMBERG (Mar. 20, 2019, 8:10 AM), https://www.bloomberg.com/news/articles/2019-03-20/eu-calls-on-facebook-twitter-to-step-up-fake-news-fight-again [https://perma.cc/3QD2-YUKG]. Regarding other countries, see Amanda Meade, *Facebook Fake News Inquiry: The Countries Demanding Answers*, GUARDIAN (Nov. 27, 2018, 7:58 AM), https://www.theguardian.com/technology/2018/nov/27/facebook-fake-news-inquiry-the-countries-demanding-answers [https://perma.cc/8A8P-BKUG].

[178] It is important to note, however, that ultimately almost all moderation decisions originate from a human who is designing the algorithm. When machine learning and artificial intelligence are involved, the human involvement may be very low.

[179] Joan Donovan, *Here's How Social Media Can Combat the Coronavirus 'Infodemic*,' MIT TECH. REV. (Mar. 17, 2020), https://www.technologyreview.com/2020/03/17/905279/facebook-twitter-social-media-infodemic-misinformation/ [https://perma.cc/M7LZ-FVLU]; Nikolaj Nielsen, *Tech Giants Must Stop Covid-19 'Infodemic', Say Doctors*, EUOBSERVER (May 7, 2020, 12:00 PM), https://euobserver.com/coronavirus/148281 [https://perma.cc/R3NR-C5NK]; Rebecca Bellan, *Americans Don't Trust Tech Platforms to Prevent Misuse in the 2020 Elections*, FORBES (Feb. 26, 2020, 3:28 PM), https://www.forbes.com/sites/rebeccabellan/2020/02/26/americans-dont-trust-tech-platforms-to-prevent-misuse-in-the-2020-elections/ [https://perma.cc/WR4C-NHV3].

[180] Associated Press, *Social Media Platforms Face Reckoning Over Hate Speech*, VOICE AMERICA (June 30, 2020), https://www.voanews.com/silicon-valley-technology/social-media-platforms-face-reckoning-over-hate-speech [https://perma.cc/84WG-CZB2].

[181] Kate O'Flaherty, *YouTube Keeps Deleting Evidence of Syrian Chemical Weapon Attacks*, WIRED (June 26, 2018), https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video [https://perma.cc/CZ6X-6LPW]; David Meyer, *AI Is Now YouTube's Biggest Weapon Against the Spread of Offensive Videos*, FORTUNE (Apr. 24, 2018), https://fortune.com/2018/04/24/youtube-machine-learning-content-removal/ [https://perma.cc/DRX5-VM3D].

[182] Kang-Xing Jin, *Keeping Our People and Our Platforms Safe*, FACEBOOK NEWSROOM (Mar. 16, 2020, 8:46 PM), https://about.fb.com/news/2020/04/coronavirus/#keeping-our-teams-safe [https://perma.cc/V3RX-6N3V].

[183] The YouTube Team, *Protecting Our Extended Workforce and the Community*, YOUTUBE (Mar. 16, 2020), https://youtube-creators.googleblog.com/2020/03/protecting-our-extended-workforce-and.html [https://perma.cc/ZB7J-XKQ6].

[184] Vijaya Gadde & Matt Derella, *An Update on Our Continuity Strategy During COVID-19*, TWITTER (Apr. 1, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html [https://perma.cc/P6ED-LJ7T].

tion to AI, since their human reviewers were absent due to mandatory lockdowns.[185]

Content moderation by platforms unquestionably makes our public sphere a safer place. At the same time, however, it can be over-protective, silencing legitimate or marginalized speech.[186] False positives are frequent,[187] but so are false negatives.[188] This is particularly true when disagreements exist as to the underlying values. Hence, overseeing how content moderation shapes the online sphere is an important public goal.[189] As explained below, it could also coincide with the private interests of platforms. This Part surveys key features of AI systems of content moderation that make them a special case of AI, including contextual decision-making and public sphere attributes. It proceeds to conclude that the impact-assessment schemes enunciated under the Algorithmic Accountability Act and the GDPR are not tailored to mitigate concerns pertaining to AI for content-moderation purposes.

A. Content Moderation by AI and Impact Assessments

The idea of using tools of self-assessment to oversee how platforms deploy AI for content-moderation purposes raises some important legal and political issues in the United States. These issues primarily stem from the fact that the Free Speech Clause of the First Amendment of the U.S. Consti-

---

[185] Jack Goldsmith & Andrew Keane Woods, *Internet Speech Will Never Go Back to Normal*, ATLANTIC (Apr. 27, 2020, 3:15 PM), https://www.theatlantic.com/ideas/archive/2020/04/what-covid-revealed-about-internet/610549/ [https://perma.cc/S34E-E7K5].

[186] *See generally* TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018); Corynne McSherry, *Platform Censorship: Lessons from the Copyright Wars*, ELECTRONIC FRONTIER FOUND. (Sept. 26, 2018), https://www.eff.org/deeplinks/2018/09/platform-censorship-lessons-copyright-wars [https://perma.cc/MBZ2-S62K]; Queenie Wong, *Is Facebook Censoring Conservatives or Is Moderating Just Too Hard?*, CNET (Oct. 29, 2019), https://www.cnet.com/features/is-facebook-censoring-conservatives-or-is-moderating-just-too-hard/ [https://perma.cc/N5QL-CR2R].

[187] Ben Depoorter & Robert Kirk Walker, *Copyright False Positives*, 89 NOTRE DAME L. REV. 3 19, 320–21 (2013); Daphne Keller, *Empirical Evidence of "Over-Removal" by Internet Companies Under Intermediary Liability Laws*, CTR. FOR INTERNET & SOC'Y (May 8, 2020), http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws [https://perma.cc/ZK6C-PR2V].

[188] *See* Derek E. Bambauer, *From Platforms to Springboards*, 2 GEO. L. TECH. REV. 417, 421 (2018); Maayan Perel, *Digital Remedies*, 35 BERKELEY TECH. L.J. 1, 26 (2020).

[189] *See generally* Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011 (2018); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 66 (2019); Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035 (2018); Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353 (2018); Mark Zuckerberg, *Mark Zuckerberg: The Internet Needs New Rules. Let's Start in These Four Areas*, WASH. POST (Mar. 30, 2019), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html [https://perma.cc/6SX5-KSHK] (calling for increased government regulation and greater oversight over content moderation on social media).

tution generally restricts government regulation of private speech.[190] Indeed, governmental attempts to limit platforms' ability to carry out content moderation might be considered a violation of First Amendment rights.[191] This is why various attempts to regulate the ways platforms moderate online content are largely resisted in the United States.[192]

One might argue that requiring platforms to submit an impact assessment could indirectly impact the content of speech, insofar as the speech available online after the requirement for impact assessment is implemented could be significantly different from that found before. Yet, impact assessments do not force platforms to host content against their will.[193] Rather, they subject them to general requirements of internal review and external check. Impact assessments do not purport to actively interfere with the way platforms manage the content on their services, but instead require them to be more transparent about their goals and evaluate the possible implications of their systems.

In fact, the idea of conducting impact assessments could also serve the interests of the platforms, which are striving to build brand recognition and increase users' trust. Most online platforms depend on network effects to achieve and maintain their success.[194] However, illicit, hateful, illegal, or otherwise unwanted or objectionable content might lead to brand degradation and may drive away users and advertisers,[195] while effective content moderation enables more efficient user engagement and traffic.[196] To this end, preparing and submitting accessible impact assessments could allow platforms to signal their goals and ambitions to build and preserve trust. Doing so could assist platforms in improving their content-moderation systems, making them more effective in detecting and removing objectionable content. Hence, it should be in the interest of platforms to conduct impact

---

[190] *See* Tim Wu, *Is the First Amendment Obsolete?*, 117 MICH. L. REV. 547, 568 (2018); Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1926 (2019) (holding that that the Free Speech Clause of the First Amendment of the U.S. Constitution prohibits only governmental, not private, abridgment of speech).

[191] *See* Keller, *supra* note 30; Langvardt, *supra* note 189, at 1364; Kyle Langvardt, *The Doctrinal Toll of Information as Speech*, 47 LOY. U. CHI. L.J. 761, 769–75 (2016).

[192] *See, e.g.*, Packingham v. North Carolina, 137 S. Ct. 1730, 1737–38 (2017); Klonick, *supra* note 28, at 1611–12, 1626–27.

[193] *See supra* note 33 and accompanying text.

[194] *See* Spencer Weber Waller, *Antitrust and Social Networking*, 90 N.C. L. REV. 1771, 1787 (2012) ("Network effects refer to the well-known phenomenon that systems may quickly increase in value as the number of users grow, and similarly, that the network may have little, or no, value without large scale adoption."). *See generally* DENNIS C. KINLAW, CONTINUOUS IMPROVEMENT AND MEASUREMENT FOR TOTAL QUALITY (1992) (discussing self-regulation as means to maintain customers satisfaction).

[195] Balkin, *supra* note 189, at 2022; Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435, 1454–55 (2011); Klonick, *supra* note 28, at 1627–28.

[196] Kornbluh & Goodman, *supra* note 160 ("[D]igital platforms are not neutral, tamper-safe pipes. They are ad-delivery platforms constructed to reward engagement.").

assessments that indicate their goals and ambitions, making them accessible to the public to preserve their users' trust.

It is not only the desire to avoid public outrage and maintain brand recognition (i.e., direct business interests) that incentivizes platforms to assure the effectiveness and fairness of their AI-driven content-moderation systems. Increased liability risks also encourage platforms to check on their privately designed systems of moderation. Indeed, recent regulatory efforts have expanded the potential liability of online platforms for potentially harmful content on their websites.[197] For instance, the German government has introduced the Network Enforcement Act, which requires major social network providers to delete unlawful content within a short timeframe after a complaint has been filed.[198] Similarly, outside the United States, platforms can be seriously fined if they fail to remove illegal content.[199] Other reforms use copyright law to motivate content moderation. For instance, the EU's new Copyright in the Digital Single Market Directive assigns greater responsibility to platforms to monitor and screen user content uploads.[200] Hence, by preparing and submitting transparent impact assessments, platforms could arguably better convey to governments how they intend to abide by applicable laws and the expected shortcomings of their systems.

However, even though it seems like platforms should have sufficient internal incentives to engage in self-assessment, a closer look at the unique attributes of AI-based systems of content moderation reveals that impact assessments are ill suited to subject these systems to adequate public oversight. This will be further elaborated below.

---

[197] A recent report prepared by the Poynter Institute shows there are numerous anti-misinformation initiatives around the world. *See, e.g.*, Daniel Funke & Daniela Flamini, *A Guide to Anti-Misinformation Actions Around the World* (Aug. 13, 2020), https://www.poynter.org/ifcn/anti-misinformation-actions/ [https://perma.cc/49GA-X9MY].

[198] Gesetz zur Verbesserung der Rechtsdurchsetzung in Sozialen Netzwerken [Netzwerkdurchsetzungsgesetz] [Network Enforcement Act], June 30, 2017, BUNDESGESETZBLATT, TEIL I [BGBL I] at 3352 (Ger.), https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=PublicationFile&v=2 [https://perma.cc/L3FE-YZF3] [hereinafter Network Enforcement Act]. For a detailed critique, see MacKenzie F. Common, *Fear the Reaper: How Content Moderation Rules Are Enforced on Social Media*, 34 INT'L REV. L. COMPUTERS & TECH. 126 (2020).

[199] Platforms that fail to comply with the Network Enforcement Act risk fines of up to €50 million. *See* Network Enforcement Act, *supra* note 198, §4; Heidi Tworek & Paddy Leerssen, *An Analysis of Germany's NetzDG Law*, TRANSATLANTIC WORKING GROUP (Apr. 15, 2019), https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf [https://perma.cc/88QC-96C9].

[200] 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 139). The text of the Directive was adopted by the European Parliament on March 26, 2019, with 338 votes in favor, 283 against, 36 abstentions, and 93 Members not attending the session. The Directive was subsequently ratified by the European Council. *See Press Release: Censorship Machine Takes over EU's Internet*, EDRI (Mar. 26, 2019), https://edri.org/censorship-machine-takes-over-eu-internet/ [https://perma.cc/MX4K-3X5S]; Martin Husovec, *How Europe Wants to Redefine Global Online Copyright Enforcement*, *in* PLURALISM OR UNIVERSALISM IN INTERNATIONAL COPYRIGHT LAW (Tatiana Eleni Synodinou ed., 2019).

B. Contextual Decision-Making

As we explained, for public oversight mechanisms to succeed in their goals, meaningful transparency is crucial. Further, it is necessary to inform the public about the ways in which a given decision-making system is expected to affect them. Nevertheless, there is often no clear line between what does and does not violate a platform's rules since the application of content-moderation policies is context specific, time sensitive, and locality dependent.[201] This means that requiring regulated entities to conduct impact assessments primarily before implementation fails to take into account not only the dynamic nature of the AI system, but also the ever-evolving meaning of content. Unfortunately, the end result is often a further reduction in the ability of impact assessments to serve as effective oversight mechanisms when it comes to AI-based systems for content moderation.

While in some instances, such as in the case of child pornography, defining what constitutes illicit content might be straightforward, other cases are more problematic. For instance, although a platform can and should seek to ban offensive language, the definition of offensive language is not always clear.[202] However, platforms have had to turn context into a set of objective rules, or "laws of flagging."[203] These laws of flagging are generally hidden

---

[201] *See supra* Section III.B.1.

[202] *See* Timothy Jay, *Do Offensive Words Harm People?*, 15 PSYCHOL. PUB. POL'Y & L. 81, 89 (2009) ("Linguistic research makes it clear that no universal statements can be made about what speech will be regarded as offensive. The meaning and impact of speech is entirely determined by the contextual factors, such as the relationship between the speaker and listener and the topic of discussion." (citations omitted)).

[203] Klonick, *supra* note 28, at 1635. For instance, BandCamp's Terms of Service explicitly state that it is the platform's policy to "block access to or remove material that it believes in good faith to be the intellectual property of a third party (e.g., copyrights, trademarks, trade secrets, etc.) that has been illegally copied and distributed by any of our advertisers, affiliates, content providers, members or users." *See Intellectual Property Policy*, BANDCAMP, https://bandcamp.com/copyright [https://perma.cc/ZD34-RD5X]. BandCamp is not alone in engaging in such content moderation. To name just a few, YouTube, SoundCloud, and Vimeo all have similar policies. *See Terms of Service*, YOUTUBE https://www.youtube.com/static?template=terms [https://perma.cc/V3NT-ZZ64] ("On becoming aware of any potential violation of these Terms, YouTube reserves the right (but shall have no obligation) to decide whether Content complies with the content requirements set out in these Terms and may remove such Content and/or terminate a User's access for uploading Content which is in violation of these Terms at any time, without prior notice and at its sole discretion."); *Terms of Use*, SOUNDCLOUD, https://soundcloud.com/terms-of-use [https://perma.cc/ZGQ2-B9YF] ("SoundCloud reserves the right to block, remove or delete any content at any time, and to limit or restrict access to any content, for any reason and without liability, including without limitation, if we have reason to believe that such content does or might infringe the rights of any third party, has been uploaded or posted in breach of these Terms of Use, our Community Guidelines or applicable law, or is otherwise unacceptable to SoundCloud."); *Vimeo Copyright Policy*, VIMEO, https://vimeo.com/dmca [https://perma.cc/ZGQ2-B9YF] ("Each user must ensure that the materials they upload do not infringe any third-party copyright. Vimeo will promptly remove materials in accordance with the Digital Millennium Copyright Act ('DMCA') when properly notified that the materials infringe a third party's copyright. Vimeo will also, in appropriate circumstances, terminate the accounts of repeat copyright infringers.").

and not open to public scrutiny.[204] Even when companies voluntarily publish some information pertaining to their content-moderation practices, they normally do so in a manner that allows for very little, if any, meaningful public review.

Indeed, Facebook has kept its content-moderation guidelines secret for many years, until a former employee leaked a copy of the company's operating guidelines back in 2012, which revealed how moderators determined whether flagged content violated Facebook's Community Standards.[205] A subsequent leak of over 100 documents detailing Facebook's internal content moderation guidelines occurred in 2017.[206] Along with giving the public a glimpse into Facebook's internal rulebook on sex, terrorism, and violence,[207] these files also revealed many oddities that caused public outrage.[208] After the leak, Facebook published an expanded version of its community standards, making them available for public input.[209] These guidelines supposedly defined "what is and isn't allowed on Facebook,"[210] but even these community standards were vague and left much room for secrecy. In fact, the standards give us little to no real understanding of how the platform decides what is acceptable[211] or how the implementation of these guidelines can be assessed and monitored in practice. Within the sphere of AI oversight, impact-assessment schemes are ill suited to ratify this deficiency, as they do not require regulated entities to make any information pertaining to their decision-making process publicly available.

Moreover, even if the platform attempts to evaluate both the "rules" it follows when flagging questionable content and the means, methods, and

---

[204] Catherine Buni & Soraya Chemaly, *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, VERGE (Apr. 13, 2016), https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-redditcensorship-free-speech [https://perma.cc/TZ4Y-UCA6].

[205] *See* Adrian Chen, *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads,'* GAWKER (Feb. 16, 2012), https://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads [https://perma.cc/2W3D-RYCM]; Tarleton Gillespie, *The Dirty Job of Keeping Facebook Clean*, SOC. MEDIA COLLECTIVE (Feb. 22, 2012), https://socialmediacollective.org/2012/02/22/ [https://perma.cc/X7FE-SY8Q].

[206] Jon Fingas, *Facebook Defends Content Policy After Guidelines Leak*, ENGADGET (May 23, 2017) https://www.engadget.com/2017/05/23/facebook-defends-content-guidelines/ [https://perma.cc/5PAC-269E].

[207] *See Facebook's Manual on Credible Threats of Violence*, GUARDIAN (May 21, 2017), https://www.theguardian.com/news/gallery/2017/may/21/facebooks-manual-on-credible-threats-of-violence [https://perma.cc/USB5-BW3U].

[208] For instance, while the company's internal guidelines stipulated that the phrase "[s]omeone shoot Trump" should be deleted, as the U.S. president was categorized as part of a "protected category," the sentence "[t]o snap a bitch's neck, make sure to apply all your pressure to the middle of her throat" would not be seen as a credible threat. *Id.*

[209] *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/ [https://perma.cc/P8B9-2EDS].

[210] Nabiha Syed, *Real Talk About Fake News: Towards a Better Theory for Platform Governance*, 127 YALE L.J. 337, 344–45 (2017).

[211] *Community Standards Enforcement Report*, FACEBOOK, https://transparency.facebook.com/community-standards-enforcement [https://perma.cc/J3WY-MH2N].

processes used to implement those "rules of flagging" as part of its impact assessment, it might fail to anticipate certain risks posed by the automated decision-making systems to marginalized groups. One famous example is the way Facebook has been incorrectly deleting posts containing the word "dyke."[212] The use of the word "dyke" may be hate speech when directed as an attack on someone; however, if one posted a photo of herself with #dyke to denounce homophobia and reclaim the word, removing the content would mean restricting that person's ability to use a derogatory speech in a self-referential, non-derogatory context.[213]

Finally, of significant importance is the fact that content moderation is context specific, time sensitive and locality dependent,[214] which also accentuates the importance of due process (e.g., notice and right to be heard). As stated previously, these are not adequately secured under the GDPR and the Algorithmic Accountability Act analyzed throughout this paper.

C. Embedded Externalities

Online platforms and social media are rapidly becoming the most important spaces for people to come together and share their thoughts, ideas, and opinions. These platforms are increasingly being treated as the modern public square,[215] yet platforms such as Google, Facebook, YouTube, and Twitter are not merely a passive and neutral infrastructure for these public gatherings. These platforms play an active role in this virtual public square by mediating content, thereby controlling what content is available to which audience, for how long, and under what conditions. Thus, platforms hold the power to manipulate public discourse, not only by banning or removing cer-

---

[212] *See Facebook: Stop Discriminating Against Lesbians*, CHANGE.ORG https://www.change.org/p/facebook-stop-discriminating-against-lesbians [https://perma.cc/6T9V-VEGH]; Lisa A. Mallett & Liz Waterhouse, *Facebook Has a Problem With Dykes* (June 24, 2017), https://listening2lesbians.com/2017/06/24/facebook-has-a-problem-with-dykes/ [https://perma.cc/77XP-ZVH4]; Kenny Sharpe, *Users Face Consequences as Facebook Struggles to Filter Hate Speech*, GLOBE & MAIL (July 27, 2017), https://www.theglobeandmail.com/life/facebook-faces-pitfalls-in-quest-to-filter-hate-speech/article35819000/ [https://perma.cc/C4CN-5VGP].

[213] Annabel Thompson, *The Controversy Around Facebook Banning Lesbians from Using the Word 'Dyke,'* THINK PROGRESS (July 12, 2017), https://thinkprogress.org/is-facebook-banning-the-word-dyke-3720433451ed/ [https://perma.cc/BGQ4-XM5M].

[214] ROXANA RADU, NEGOTIATING INTERNET GOVERNANCE 179 (2019) ("Local values representation is the second point of contention towards Facebook community. The unilateral definition of what is and what is not acceptable online by a company headquartered in the United States is harder to sustain as more than 2 billion people use the platform. Facebook's largest user base at the moment is India, but little of the social and cultural norms there appear to transpire in the global policy of the company.").

[215] *See* Packingham, *supra* note 192, at 1737; William Perrin & Lorna Woods, *Reducing Harm in Social Media Through a Duty of Care*, CARNEGIE UK TRUST (May 8, 2018), https://www.carnegieuktrust.org.uk/blog/reducing-harm-social-media-duty-care/ [https://perma.cc/7SHA-V3JL].

tain illegal or unwanted speech, but also by subtly limiting the reach and exposure of speech.[216]

From an economic perspective, speech regularly generates certain costs or benefits—depending on the content and context—realized by parties other than the speaker (i.e., externalities).[217] For instance, hate speech and terrorist propaganda generate costs (i.e., negative externalities), while scientific progress as a result of profound theoretical discussions generates benefits (i.e., positive externalities). For the most part, the individual speaker will not consider these costs or benefits to third parties when deciding whether to exercise his freedom to speak.[218] But when a platform decides whether to filter, block, remove, or limit the distribution of certain speech, it could affect the externalities produced. This explains why content moderation—*if performed poorly*—can impose costs, not only on the individual (i.e., the speaker) but also on society as a whole, much like global warming or pollution.

To illustrate, consider Jane, a young activist who shares her thoughts and ideas about life with her friends, colleagues, and followers online. In particular, she has been dedicating her time and efforts to raising awareness on the issue of plastic waste. Jane's Facebook posts and Twitter tweets have

---

[216] *See* Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/HP3E-JFCV] ("This is a basic incentive problem that we can address by penalizing borderline content so it gets less distribution and engagement. By making the distribution curve look like the graph below where distribution declines as content gets more sensational, people are disincentivized from creating provocative content that is as close to the line as possible."); Evelyn Douek, *Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility*, 21 N.C. J.L. & TECH. 1, 11 (2019).

[217] Brett M. Frischmann, *Speech, Spillovers, and the First Amendment*, U. CHI. LEGAL F. 301, 317 (2008); Richard A. Posner, *Free Speech in an Economic Perspective*, 20 SUFFOLK UNIV. L. REV. 1 (1986) (discussing regulation of the freedom of speech from an economic perspective); Daniel A. Farber, *Free Speech Without Romance: Public Choice and the First Amendment*, 105 HARV. L. REV. 554 (1991) (discussing the idea of speech as a public good); *see also* Ronald H. Coase, *Advertising and Free Speech*, 6 J. LEGAL STUD. 1, 1-5 (1977); Ronald H. Coase, *The Economics of the First Amendment: The Market for Goods and the Market for Ideas*, 64 AM. ECON. REV. 384 (1974). *But see* Peter J. Hammer, *Free Speech and the "Acid Bath": An Evaluation and Critique of Judge Richard Posner's Economic Interpretation of the First Amendment*, 87 MICH. L. REV. 499 (1988) (criticizing Posner's approach); Anne Barron, *Copyright Infringement, 'Free Riding', and the Lifeworld* (London Sch. of Econ. & Political Sci., Law, Soc. & Econ. Working Paper No. 17, 2008).

[218] Such externalities are usually associated with public goods. Speech is often regarded as a public good. *See* Frischmann, *supra* note 217, at 318 (arguing that "speech involves the communication of ideas and that it thus involves the sharing of public good"). Public goods are generally characterized as nonrival and nonexclusive, signifying that, once it has been provided to one person, its benefits cannot be restricted and are inevitably spread. Also, it regularly generates costs or benefits (i.e., externalities). *See* Farber, *supra* note 217; RICHARD CORNES & TODD SANDLER, THE THEORY OF EXTERNALITIES, PUBLIC GOODS, AND CLUB GOODS 9 (2d ed. 1996) ("The benefits of private goods are fully rival and excludable, whereas the benefits of pure public goods are nonrival and nonexcludable. From the foregoing examples, we see that food and fuel are private, whereas strategic weapons and pollution control are purely public goods.").

been shared repeatedly. They have helped inform people on ways in which they can reduce their plastic consumption and increase the proportion of recycled or biodegradable plastic within their communities. Furthermore, Jane's posts have ignited a lively debate on the issue of plastic waste worldwide. Jane's posts have, no doubt, generated positive external effects.[219]

Now, assume that, for some reason, one of the major platforms deletes some of Jane's posts, limits their visibility, or prevents their ability to be shared. This would surely harm Jane, the individual speaker; but more importantly, it would slightly shift the supply of content, which in turn could harm people's ability to be informed, form their own opinion, and engage in meaningful public discourse.[220]

Although the aforementioned negative effect might seem relatively small, the aggregate effects associated with platforms' content-moderation practices are plausibly very high.[221] Impact assessments essentially require regulated entities to evaluate automated decision-making systems for "impacts on the accuracy, fairness, bias, crimination, privacy, and security" as well as the "relative benefits and costs of the automated decision system in light of its purpose."[222] Additionally, they must address the "risks that the automated decision system may result in or contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers."[223] Hence, impact assessments should make platforms take account of the collateral impact of their content-moderation practices on the individual speaker as well as on the public. This can be compared to how environmental impact-assessment mechanisms make private and governmental entities accountable for the impacts the construction of a factory, a road, or a dam might have on members

---

[219] This is not to diminish or overlook the effect of intellectual property laws on the way in which people could share those ideas, or the way intellectual property laws may cure certain market failures pertaining to the public-good nature of speech. *See generally* Rebecca Tushnet, *Copy This Essay: How Fair Use Doctrine Harms Free Speech and How Copying Serves It*, 114 YALE L.J. 535 (2004).

[220] In reality, of course, content-moderation systems can also generate positive externalities. Assume that Jane does not disseminate publicly beneficial content, but rather terrorist propaganda (a negative externality). The classical approach to negative externalities is to impose its cost on the producer. Although the idea of platforms charging Jane a fee corresponding to the costs that her speech imposes on others is theoretically possible, it is unlikely. However, by means of content moderation (e.g., blocking or removing terrorist propaganda), online platforms can reduce the number of negative externalities generated by individual speakers, such as Jane. Dhammika Dharmapala & Richard H. McAdams, *Words That Kill? An Economic Model of the Influence of Speech on Behavior (with Particular Reference to Hate Speech)*, 34 J. LEGAL STUD. 93, 93 (2005). *But see* Larry Alexander, *Banning Hate Speech and the Sticks and Stones Defense*, 13 CONST. COMMENT. 71, 98 (1996). However, the First Amendment constrains the government's ability to force speakers to internalize externalities associated with their speech. *See* Frischmann, *supra* note 217, at 337.

[221] Frischmann, *supra* note 217, at 320-24; Brett Frischmann, *Spillovers Theory and Its Conceptual Boundaries*, 51 WM. & MARY L. REV. 801, 818 (2009).

[222] *See* Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019).

[223] *Id.*

of the public.[224] However, impact-assessment schemes are largely ineffective in this regard.

Specifically, as explained next, impact-assessment schemes are unable to eliminate information gaps or provide individuals or the public at large with sufficient incentives to challenge platforms' erroneous decision-making practices.

1. *Information Gaps*

As stated earlier, due to the opaque nature of AI systems, it is extremely difficult for members of the public to understand how these systems are utilized and to identify what the potential risks embedded within them are. Impact assessments are meant to bridge this information gap. Nevertheless, since there are no mandatory publication requirements, impact-assessment schemes cannot actually solve the information gap problem.

The information gap problem is also reflected in the content-moderation context, where users are generally left in the dark not only with regard to the way the AI-based content moderation systems operate, but also in terms of whether a specific removal decision has affected them. The origin of this problem stems from the fact that platforms are simultaneously engaged in at least two forms of speech regulation.[225] The first function optimizes users' engagement with online content and with each other, while the second function aims to enable the rapid detection and removal of illicit, harmful, or otherwise unwanted content.[226] While the latter function might lead to the removal of certain content, the first aims to reward it.[227] Outrageous content might seem inappropriate but could increase user engagement. For instance, a study at Harvard's Berkman Klein Center for Internet and Society found that YouTube's recommendation system curated a list of recommended videos for users that displayed partially clothed children, sometimes after those users watched sexually explicit content.[228] Each family home video on its own is perfectly innocent, but when grouped together in a

---

[224] For a discussion in the environmental law context and the ability of impact assessment to force government agencies to consider collateral effect on the public, see Paul J. Culhane, *NEPA's Impacts on Federal Agencies, Anticipated and Unanticipated*, 20 ENVTL. L. 681, 690 (1990); Stephen Jay et al., *Environmental Impact Assessment: Retrospect and Prospect*, 27 ENVTL. IMPACT ASSESSMENT REV. 287 (2007).

[225] Elkin-Koren & Perel, *supra* note 37.

[226] Zuckerberg, *supra* note 216; Michael Kan, *Facebook Taps Next-Gen AI to Help It Detect Hate Speech*, PCMAG (May 1, 2019), https://www.pcmag.com/news/facebook-tap-snext-gen-ai-to-help-it-detect-hate-speech [https://perma.cc/Q37T-J7SZ].

[227] Kornbluh & Goodman, *supra* note 160 (noting that hateful and outrageous content drew heavy engagement); *see also* Elkin-Koren & Perel, *supra* note 37.

[228] Jonas Kaiser & Yasodara Córdova, *On YouTube's Digital Playground*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y (June 3, 2019), https://cyber.harvard.edu/story/2019-06/youtubes-digital-playground [https://perma.cc/AZ2H-F8WS].

particular path of user consumption following sexually explicit materials, their meaning could change.[229]

The integration of these two forms of speech regulation into a single system of AI-based content moderation suggests that each individual views a different fraction of the overall public discourse that is personally tailored but fragmented.[230] This personalized-but-fragmented view makes it much harder for speakers to notice and evaluate content removals or blockings. Thus, even if platforms provide compatible impact assessments, users could hardly use them to oversee the complicated manner in which their personalized views of the public discourse are shaped.

2. *Insufficient Stakes*

Even when information gaps can be mitigated and the speaker learns that legitimate speech has been blocked or removed, the perceived harm might seem too small for the speaker to act upon and challenge the removal.[231] Further, as noted, when a platform improperly blocks, removes, or limits a piece of content, it not only causes harm to the individual speaker, but also to members of the public who are now deprived of parts of the public discourse. Absent some cooperative mechanism or a regulatory intervention, individual members of the public seem to have insufficient incentives to maintain the public discourse or resources or abilities to demand the correction of wrong content removals.[232]

In order to illustrate this point, let us return to our activist Jane. Assuming the harm suffered by Jane due to the platform's decision to remove or block her speech is very low, she is unlikely to challenge the platform's decision. Even when the platform allows her to contest the removal of her content, doing so will require her to invest time and resources, which, due to the vague content-moderation policies of most platforms, could potentially result in their decisions being upheld. The lack of sufficient due process under the Algorithmic Accountability Act only exacerbates the problem. Moreover, even under the GDPR, where impact assessments are accompanied by a right to explanation, the content might be time sensitive, and, therefore, Jane will be discouraged from contesting the platform's decision.

---

[229] *See* Elkin-Koren & Perel, *supra* note 37, at 889–90.

[230] *See* Elkin-Koren & Perel, *supra* note 37, at 889–90.

[231] For instance, research conducted in the context of copyright notice-and-takedown policies indicates that most users do not employ the counter-notice procedure to challenge the removal of content, which may qualify as fair use. *See* Elkin-Koren & Perel, *supra* note 46, at 501–02; Jennifer M. Urban et al., *Notice and Takedown: Online Service Provider and Rightsholder Accounts of Everyday Practice*, 64 J. Copyright Soc'y U.S. 371, 393 (2017).

[232] This is somewhat akin to the "anticommons" problem in the sense that users have insufficient incentive to invest in maintaining and protecting the commons (e.g., the public discourse). *See* William W. Buzbee, *Recognizing the Regulatory Commons: A Theory of Regulatory Gaps*, 89 Iowa L. Rev. 1, 5–6 (2003). *See generally* Michael A. Heller, *The Tragedy of the Anticommons: Property in the Transition from Marx to Markets*, 111 Harv. L. Rev. 621 (1998).

Instead, she might decide it is better to redirect her efforts and attempt to repost the content or even abandon it altogether. The public as a whole, on the other hand, will suffer great costs on account of the platform's misconduct, since the possible aggregated positive externalities generated by such content could be high. However, due to the information gaps and the fact that the platform is the only entity with the ability to see the full picture, individual users are likely to be unaware of this harm. Even if members of the public were somehow to learn that Jane's post had been removed, it is doubtful whether users who individually have insufficient stakes would invest the time and effort necessary to contest the platform's automated decision-making systems.[233]

In summary, the impact-assessment schemes delineated in the Algorithmic Accountability Act and the GDPR are not tailored to mitigate AI content-moderation concerns. Such concerns include notifying users when content is restricted, providing the public with information about how restrictions are implemented and what expression is being restricted, and ensuring that users have access to due process. While requiring impact assessments could impose important reporting obligations on platforms, such assessments are still insufficient to facilitate meaningful oversight and accountability in the case of content moderation by AI. The absence of a common, comparable threshold that members of the public could use to oversee how AI-based content moderation affects it, together with the insufficient stakes of affected individuals, suggest that relying on private mechanisms of self-assessment is insufficient for AI-based content moderation. In these instances, combining the notion of self-assessment with an external and independent oversight mechanism may be the better solution.

V. A Dual Mechanism of Oversight for AI-Based Content Moderation

The efforts of the United States and the European Union to implement mandatory algorithmic impact assessments mark a historic step towards trying to achieve public transparency and oversight over automated decision-making systems. These schemes, however, provide only limited transparency, fall short of securing due process, and are unable to secure adequate public scrutiny. Notwithstanding the above, we do not argue that policymakers should abandon the idea of impact assessment, even in the area of content moderation by AI. Instead, we contend that any harmful impact caused by moderation needs to be weighed against the benefit that it would achieve.

---

[233] This is similar to the problem of insufficient stake in small claims litigation. *See, e.g.*, Jonathon R. Macey & Geoffrey P. Miller, *The Plaintiffs' Attorney's Role in Class Action and Derivative Litigation: Economic Analysis and Recommendations for Reform*, 58 U. Chi. L. Rev. 1, 8 (1991) ("In the absence of the class action device, such injuries would often go unremedied because most individual plaintiffs would not themselves have a sufficient economic stake in the litigation to incur the litigation costs.").

Traditionally, the tradeoff would attempt to balance the speaker's individual rights against harm caused to others. However, as has been stressed throughout this paper, in the moderation world, it should be equally weighted against the public interest in disseminating certain speech. Accordingly, in this last Part, we propose a dual mechanism of oversight for content moderation by AI that integrates an improved model of self-assessment with an external, public mechanism of review.

A. Internal Checks

At the first stage, we suggest a framework for improving impact assessment. Accordingly, regulated entities would be required to submit an assessment of automated decision-making and AI-led systems that are part of the content-moderation process. This assessment framework must be tailored to better advance accountability.

1. *Periodic Impact Assessment*

As stated earlier, a major flaw in the above-discussed initiatives is that assessments—either ADSIAs or DPIAs—are performed mainly ex ante, whereas the dynamic nature of AI-based systems naturally means that decision-making processes are likely to change over time in response to new data fed back into the system. Furthermore, ex ante assessment is unable to account for the context-specific and time-sensitive attributes of content moderation.[234] To better address these issues, we suggest mandatory periodic impact assessments. This means that an impact assessment would be submitted ex ante (before and after training data has been fed into the system) as well as ex post (routinely after the AI system has been implemented),[235] which would ultimately enhance transparency and public review.

Subsequent or complimentary assessments would be conducted every few months or annually in order to reflect the constant changes in the data being fed into the system as well as the dynamism of speech. Therefore, the firm would be required to assess its own automated decision-making system more frequently and regularly. When these subsequent assessments would occur could vary based on the industry; different industries might present different levels of dynamism and thus require different time frames between assessments.

---

[234] *See supra* notes 203–214 and accompanying text.

[235] *See* Anthony J. Casey & Anthony Niblett, *A Framework for the New Personalization of Law*, 86 U. Chi. L. Rev. 333, 356 (2019) ("The relevant information to test the validity of an algorithm will be what objective it was given (and how that objective was developed), how the algorithm was programmed to achieve that objective, how the data was selected, and audit data on the algorithm's performance.").

2. *Mandatory Notice-and-Comment Procedure*

Periodic assessments might fit better to the ever-changing nature of AI systems and could partly address the problems that arise due to the context-specific and time-sensitive nature of AI-based content moderation. However, in order for an impact assessment to be genuinely effective, a few additional factors must be addressed, which include the lack of adequate due process.[236] Therefore, we suggest that policymakers implement a mandatory notice-and-comment procedure within the impact-assessment scheme.[237]

In fact, in the area of environmental law, public comments are a key element in the assessment process. Such comments are considered a crucial part of overall transparency. They are a vital way in which acceptance and understanding is achieved by the public.[238] As part of the assessment process, environmental-law regulations typically provide stakeholders, interest groups, and the public with an opportunity to give comments and feedback about the possible environmental ramifications of a particular action. The underlying rationale is that, in many circumstances, these groups hold valuable knowledge about the affected environment and ecological interactions. Therefore, neither the developer nor the government can afford to miss out on this information.[239]

For instance, the U.S. National Environmental Policy Act ("NEPA"),[240] one of the leading pieces of legislation that adopts the notion of impact assessments in the environmental field,[241] requires the federal government to incorporate environmental considerations into the review of major projects. It does this by requiring an environmental assessment ("EA") and, if the EA indicates that the proposal would likely significantly impact the environment, an environmental impact statement ("EIS").[242] With regard to the EA, courts are divided as to whether or not agencies must provide a draft of the

---

[236] *See supra* note 149–161 and accompanying text.

[237] One prominent model of notice and comment is that used by U.S. administrative agencies. *See* TODD GARVEY, CONG. RESEARCH SERV., R41546, A BRIEF OVERVIEW OF RULEMAKING AND JUDICIAL REVIEW 2–3 (Mar. 27, 2017), https://fas.org/sgp/crs/misc/R41546.pdf [https://perma.cc/S6VA-W8DL].

[238] *See* Jonathan Poisner, *A Civic Republican Perspective on the National Environmental Policy Act's Process for Citizen Participation*, 26 ENVTL. L. 53, 55 (1996); Joachim Hartlik, *Requirements on EIA Quality Management*, *in* 3 STANDARDS AND THRESHOLDS FOR IMPACT ASSESSMENT 89, 92 (Michael Schmidt et al. eds., 2008); Selbst, *supra* note 77, at 178 (discussing the importance of public comments in the context of his suggestion for the implementation of algorithmic impact assessment).

[239] Hartlik, *supra* note 238, at 90.

[240] National Environmental Policy Act of 1969, 42 U.S.C. §§ 4321–70 (2018).

[241] *See, e.g.*, Richard Lazarus, *The National Environmental Policy Act in the U.S. Supreme Court: A Reappraisal and a Peek Behind the Curtains*, 100 GEO. L.J. 1507, 1520 (2012); Nicholas A. Fromherz, *From Consultation to Consent: Community Approval as a Prerequisite to Environmentally Significant Projects*, 116 W. VA. L. REV. 109, 110 (2013).

[242] Ted Boling, *Making the Connection: NEPA Processes for National Environmental Policy*, 32 WASH. U. J.L. & POL'Y 313, 318–19 (2010).

assessment to the public and solicit public comments,[243] but when it comes to the EIS, the relevant agency must give the public an opportunity to participate.[244]

Public participation in the EIS comes in two stages: first, when the agency determines the scope of the EIS; and second, when the agency prepares a draft EIS before the final version is adopted.[245] Comments made by members of civil society could eventually influence the design of the project and whether it will be implemented at all.[246] As previously argued, public participation is an important part of the impact-assessment process; therefore, we suggest that policymakers mandate some form of notice and comment during the impact-assessment period.[247]

This mandatory comment process, if implemented within the impact-assessment framework, would allow for individuals and civil-society groups to take part in the decision-making process, thus making the black box of content moderation by AI a little more transparent.[248] Apart from that, it could have several other advantages. First, it could strengthen the legitimacy of the decision-making process.[249] Second, it could make people more accepting of the outcome of the content moderation, even when that outcome is inconsistent with their own preferences.[250] Third, it could engender public involvement in the decision-making process that could increase the number of viewpoints heard, expanding the range of issues considered by the platform using AI to moderate the content.[251] Fourth, it could elicit information

---

[243] Fromherz, *supra* note 241, at 110 (citing Greater Yellowstone Coal. v. Flowers, 359 F.3d 1257, 1279 (10th Cir. 2004); Citizens for Better Forestry v. U.S. Dep't of Agric., 341 F.3d 961, 970–71 (9th Cir. 2003); Ohio Valley Envtl. Coal. v. U.S. Army Corps of Eng'rs, 674 F. Supp. 2d 783 (S.D. W. Va. 2009); Montrose Parkway Alts. Coal. v. U.S. Army Corps of Eng'rs, 405 F. Supp. 2d 587, 596 (D. Md. 2005); Natural Res. Def. Council, Inc. v. Forest Serv., 634 F. Supp. 2d 1045 (E.D. Cal. 2007); Natural Res. Def. Council v. Kempthorne, 525 F. Supp. 2d 115 (D.D.C. 2007)).

[244] 40 C.F.R. § 1501.7 (2020).

[245] *Id.*

[246] *See* Fromherz, *supra* note 241, at 125.

[247] *See, e.g.*, 40 C.F.R. §§ 1501.7, 1503.1 (establishing two mandatory notice-and-comment periods under NEPA).

[248] Albert Louis Chollet III, *Enabling the Gaze: Public Access and the Withdrawal of Tennessee's Proposed Rule of Civil Procedure 1A*, 36 U. MEM. L. REV. 695, 716 (2006).

[249] Beth Simone Noveck, *The Electronic Revolution in Rulemaking*, 53 EMORY L.J. 433, 459 (2004).

[250] Fromherz, *supra* note 241, at 149. *But see* Daniel P. Selmi, *Themes in the Evolution of the State Environmental Policy Acts*, 38 URB. LAW. 949, 975 (2006) ("[I]t seems equally likely, if not more likely, that the opposite reaction will occur. If members of the public seriously dispute the sufficiency of the project's environmental analysis, that dispute can easily escalate into accusations that the public agency is not acting in good faith, or that a private applicant is hiding the project's impacts. Alternatively, if the environmental analysis is sufficient, opponents may cite that analysis as a reason why the public agency should disapprove the project. In either case, the SEPA process will not lead to acceptance of the outcome.").

[251] *See* Rossi, *supra* note 47, at 186; Robert G. Healy & William Ascher, *Knowledge in the Policy Process: Incorporating New Environmental Information in Natural Resources Policy Making*, 28 POL'Y SCI. 1, 2 (1995).

exchange,[252] which could improve the quality of content moderation and decrease the chances of erroneous decisions.[253] Fifth, it could foster a greater public understanding of the content-moderation process and a greater appreciation of the stakes involved, thereby affecting the public's preferences and attitudes towards the agency decision.[254] Finally, it could potentially mitigate the risk that platforms will use content-moderation processes to advance the interests of specific interest groups over the interests of the public at large.[255] This is particularly important in the area of content moderation due to the information gaps discussed earlier. There have recently been some allegations that online platforms selectively draft and enforce their community guidelines to advance the interests of certain groups.[256] Due to the information gaps discussed earlier and the fact that each user views a personally tailored but fragmented segment of the public discourse, these allegations are hard to prove or disprove. Mandatory notice-and-comment procedures could diminish the risk of platforms using their content-moderation system to advance the interests of a specific group. Seeing notice would allow for a cross-political discussion.

3. *Mandatory Publication*

Given the public attributes of content moderation, we suggest that policymakers make the publication of impact assessments mandatory. Public review is, by and large, essential to effective impact assessment.[257] This is even more true in the case of content moderation, partly due to the informa-

---

[252] Noveck, *supra* note 249, at 458.

[253] *See* Rossi, *supra* note 47, at 185–86.

[254] *See* Rossi, *supra* note 47, at 187.

[255] *See* Rossi, *supra* note 47, at 184–85; Thomas W. Merrill, *Capture Theory and the Courts: 1967–1983*, 72 CHI.-KENT L. REV. 1039, 1052 (1997); Bamberger, *supra* note 47, at 468.

[256] Joseph Cox & Jason Koebler, *Why Won't Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too.*, VICE: MOTHERBOARD (Apr. 25, 2019, 12:21 PM), https://motherboard.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-likeisis-because-it-would-mean-banning-some-republican-politicians-too [https://perma.cc/G3GP-RQ5L]. See also the Executive Order entitled "Preventing Online Censorship," recently signed by the President. Exec. Order No. 13,925, Preventing Online Censorship, 85 Fed. Reg. 34,079 (May 28, 2020). The order accuses online platforms of engaging in "selective censorship," which is harming public discourse, *id.* at , 34,079, and instructs federal agencies to take action to protect against such alleged censorship, *id.* at 34,081–82. Specifically, it directs the Commerce Department to petition the FCC to generate rulemaking implementing a narrower interpretation of Section 230 of the Communications Decent Act, directs the Attorney General to prepare alternative legislation, and instructs federal agencies to review and report their spending in social media advertising. *Id.* Legal scholars have raised serious doubts as to the effective legal power of the Executive Order, arguing that the FCC, which is an independent federal agency, holds no jurisdiction over rulemaking authority under Section 230. *See, e.g.*, Michael Cheah, *Section 230 and the Twitter Presidency*, NW. U. L. REV. ONLINE 192, 204–10 (2020), https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1296&context=nulr_online [https://perma.cc/5AV5-X6FF].

[257] Selbst, *supra* note 77, at 179.

tion gaps and the lack of sufficient stakes discussed earlier.[258] Although the initiatives discussed throughout this paper allow dedicated agencies or the Attorney General to initiate some form of legal proceedings, they leave the subject of publication to the discretion of the private entity. However, we argue that, to maintain meaningful public scrutiny and support the idea of due process, policymakers should require platforms to make parts of their published impact assessments available to the public.

While this mandatory publication would not entirely eliminate the information gap, it would plausibly decrease it. For instance, by informing the public of the system's objectives, its decision-making practices, its possible risks, and even the number of posts removed and accounts suspended due to violation of content-moderation guidelines, the information gap would be significantly decreased.[259] Moreover, although individual users do not have a sufficient stake in the process, it is possible that the publication of certain aspects of the impact assessment would steer a public debate or even incentivize civil society and citizens' organizations to act. Further, once a platform discloses a policy or procedure as part of its impact assessment, it would then be obligated to do what the impact assessment states.[260] Therefore, platforms could be held accountable by government officials or by members of the public for the way that they deploy their AI systems based on the details disclosed within the impact assessment.[261]

Of course, one could object to the idea of mandatory publication, arguing that forcing platforms to publish information pertaining to their automated decision-making systems would entail disclosure of trade secrets[262] or otherwise harm their proprietary interests.[263] Interestingly, the initial text of the GDPR proposal required the impact-assessment process to involve consultation with data subjects.[264] This text was later deleted because "[t]o ac-

---

[258] *See supra* Part IV.C.

[259] For instance, the Santa Clara Principles on Transparency and Accountability in Content Moderation suggest that information regarding removed posts and suspended accounts should be broken down and provided in a regular report in an openly licensed, machine-readable format. *See* ACLU Found. of N. Cal. et al., *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, SANTACLARAPRINCIPLES.ORG, https://santaclaraprinciples.org/ [https://perma.cc/HL2Q-929U].

[260] MacCarthy, *supra* note 163.

[261] Kaminski, *supra* note 147, at 1608–09 ("Publicly disclosed impact assessments are used as a soft form of regulation to trigger market mechanisms and other forms of third-party oversight and feedback.").

[262] Rebecca Wexler, *Life Liberty, and Trade Secrets: Intellectual Property Rights in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1350 (2018). *But see* Huq, *supra* note 106, at 641 (claiming that this secrecy "does not plainly distinguish machine from human decisions").

[263] J.M. Porup, *What Does the GDPR and the "Right to Explanation" Mean for AI?*, CSO, (Feb. 9, 2018), https://www.csoonline.com/article/3254130/what-does-the-gdpr-and-the-right-to-explanation-mean-for-ai.html [https://perma.cc/6APG-H4FG].

[264] *See Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, art. 33(4), COM (2012) 0011 final (Apr. 5, 2016), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52012PC0011 [https://perma.cc/H5PS-Z2KS] ("The controller shall seek the views of data sub-

tively seek the views of data subjects represents a disproportionate burden on data controllers."[265] Therefore, in the final text of the GDPR, consultation with data subjects or their representatives is only part of the process "where appropriate."[266] Even when public participation is required, it must be "without prejudice to the protection of commercial or public interests or the security of processing operations."[267]

Notwithstanding the above, given the different ways an AI-based content-moderation system can affect public discourse, it would be incorrect to strictly follow the public/private divide in this instance.[268] Instead, we suggest balancing the interests of the platform against the interests of the public by adopting a default-disclosure rule, while at the same time allowing private entities to file for an exception, which would allow them to keep proprietary information secret under special circumstances. Under the suggested framework, mere prejudice to the protection of commercial interests of a platform may provide some grounds for an exception. Yet the entity filing for an exception would need to submit detailed information to support its request. The burden to provide a sufficient basis for approval of an exception request would potentially favor more disclosure than one under the GDPR's framework. Finally, one might argue that pre-decision public participation is less important, given the dynamic nature of content-moderation systems and the fact that users and citizens' interest groups have the opportunity to attack the results of the system after the fact. In reality, however, public involvement, both ex ante and ex post, is important. Therefore, the idea of allowing the public adequate opportunity to raise questions and comments is not without merit. This could be achieved through a mandatory notice-and-comment process, with a requirement to make available for the public published parts of the impact assessment. Ultimately, a mandatory notice-and-comment procedure in conjunction with mandatory publication requirements could strengthen transparency, due process, and public review.

B. External Auditing of Content Removal

While impact-assessment frameworks can be customized to better meet the needs and attributes of content moderation, they could not stand alone as

jects or their representatives on the intended processing, without prejudice to the protection of commercial or public interests or the security of the processing operations.").

[265] Report on the Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)), PARL. EUR. DOC. PE501.927v05-00 (2013), http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A7-2013-0402+0+DOC+XML+V0//EN [https://perma.cc/6257-WT7E]; *see also* Reuben Binns, *Data Protection Impact Assessments: A Meta-Regulatory Approach*, 7 INT'L DATA PRIVACY L. 22, 28 (2017).

[266] Commission Regulation 2016/679, *supra* note 24, art. 35(9).

[267] Commission Regulation 2016/679, *supra* note 24, art. 35(9).

[268] *See generally* Elkin-Koren & Perel, *supra* note 37.

means to advance accountability. Rather, they should be incorporated into a broader accountability program that ensures both internal checks and external oversight. As stated above, the advantage of impact assessments is that they do not interfere with the editorial discretion of platforms, but rather rely on corporate organizations' internal processes and desires to maintain users' engagement and brand recognition. Self-assessment mechanisms are particularly suitable to environments characterized by rapid technological changes.[269] On the other hand, they depend largely on those private institutions' own reporting and assessments, which can often be biased or misleading. Further, these self-assessment mechanisms could de facto entrust online platforms with public functions, thus strengthening their power and control over the public discourse.[270] This is a major concern in the context of content moderation due to the fact that AI-based content-moderation systems create personally tailored but fragmented "publics" of information.[271] As a result, it is hardly possible to detect illegitimate deprivations of information.

One way for policymakers to enhance the potency of impact assessments is by combining them with some individual rights,[272] which is the approach adopted by the GDPR. Indeed, in addition to impact-assessment schemes, the GDPR incorporated the following safeguarding measures: the right to be informed, the right to obtain human intervention, and the right to challenge the decision of an automated decision-making system.[273] These measures vest individuals with rights that have since come to be collectively referred to as the "right to explanation."[274] These measures, as noted before, are mainly designed to remedy harm to the individual's rights.[275] Such rights can complement the impact-assessment scheme by "addressing individualized error, bias, and discrimination."[276] Moreover, they can create pressure for administrative changes to improve a platform's compliance.[277] Finally, they can ensure individual review of rule setting.[278]

They are, however, poorly fitted to mitigate concerns about the ways AI-based decision-making systems affect the interest of the public in freely

---

[269] Teresa Quintel & Carsten Ullrich, *Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond*, *in* FUNDAMENTAL RIGHTS PROTECTION ONLINE: THE FUTURE REGULATION OF INTERMEDIARIES (Bilyana Petkova & Tuomas Ojanen eds., forthcoming Dec. 2020).

[270] *See* Kaminski, *supra* note 147, at 1581. *See generally* Christina Angelopoulos et al., *Study of Fundamental Rights Limitations for Online Enforcement Through Self-Regulation*, INST. FOR INFO. LAW (2015), https://pure.uva.nl/ws/files/8763808/IVIR_Study_Online_enforcement_through_self_regulation.pdf [https://perma.cc/4YQY-26DU].

[271] David, *supra* note 36.

[272] Kaminski, *supra* note 147, at 1578.

[273] *See, e.g.*, sources cited *supra* note 108.

[274] It is important to note that scholars, policymakers, and industry leaders have been debating what the GDPR's new "right to explanation" entails. *See supra* note 106 and accompanying text.

[275] *See supra* note 106 and accompanying text.

[276] Kaminski, *supra* note 147, at 1578.

[277] Kaminski, *supra* note 147, at 1578–79.

[278] Kaminski, *supra* note 147, at 1579.

consuming and accessing information,[279] since a system of individual rights is dependent on individuals exercising their rights.[280] However, as noted before, many users lack the necessary incentives or resources to exercise their rights.[281] Further, the individual and public objectives might conflict with one another.[282] Lastly, individual explanations are unlikely to trigger market mechanisms or public oversight.[283] Therefore, it might be time for policymakers to subject platforms to additional higher levels of external and objective scrutiny,[284] which could include third party audits.[285]

Kroll et al. define auditing as a means to independently evaluate whether computer systems conform "to applicable regulations, standards, guidelines, plans, specifications, and procedures."[286] External auditing may not guarantee complete transparency regarding the AI-based content-moderation system, nor completely bridge the information gaps discussed earlier, but it still may address the insufficient-stakes problem and serve as an additional oversight mechanism.[287] Ultimately, auditing may serve to "verify al-

---

[279] Kaminski, *supra* note 147, at 1607 (arguing that the GDPR "for all its coregulatory and collaborative measures, does not establish adequate public-facing or even expert-facing accountability").

[280] *See* Kaminski, *supra* note 147, at 1581 ("Even just opting out of a system without actively introducing inaccuracies can affect system bias.").

[281] *See supra* Part IV.C.

[282] *See* Kaminski, *supra* note 147, at 1580–81 (arguing that in a collaborative governance regime, which encompasses both impact assessment and individual accountability, "there is also a danger of confusing one kind of accountability for another and crafting a system that is accountable along only one axis," and that "[a] system of individual rights can conflict with system-wide accuracy and system-wide concerns about bias"); Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1, 11–12 (2018).

[283] *See* Kaminski, *supra* note 147, at 1610.

[284] It is important to note that, for high-risk activities, the GDPR requires consultation with the government. Commission Regulation 2016/679, *supra* note 24, at 54 ("The controller shall consult the supervisory authority prior to processing where a data protection impact assessment under Article 35 indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk."). However, it is not entirely clear what activity requires prior governmental consultation.

[285] For similar suggestions in the context of the GDPR, see *Report of the Working Party on the Protection of Individual with Regard to the Processing of Personal Data on "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679,"* No. WP251rev.01, at 32 (Feb. 6, 2018), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053 [https://perma.cc/2FFR-9PRH] (suggesting algorithmic auditing, third party auditing "where decision-making based on profiling has a high impact on individuals," and "ethical review boards to assess the potential harms and benefits to society of particular applications for profiling"); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017).

[286] *See* Kroll et al., *supra* note 66, at 660–61 (quoting Inst. of Elec. & Elecs. Eng'rs Comput. Soc'y, *IEEE Std 1028-2008: IEEE Standard for Software Reviews and Audits*, INST. ELECTRICAL & ELECTRONICS ENGINEERS § 8.1 (Aug. 15, 2008), http://ieeexplore.ieee.org/document/4601584 [https://perma.cc/WLD6-VPUN]) (expressing simultaneous skepticism about the complete sufficiency of auditing).

[287] *See* Ben Shneiderman, *Opinion: The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight*, 113 PROC. NAT'L ACAD. SCI. 13,538, 13,538–40 (2016); Sigal Samuel, *10 Things We Should All Demand from Big Tech Right Now,* VOX (May 29, 2019, 9:30 AM), https://www.vox.com/the-highlight/2019/5/22/18273284/ai-algorithmic-bill-of-rights-accountability-transparency-consent-bias [https://perma.cc/NR9F-43YY].

gorithmic decision-making in order to prevent improper use, discrimination, and negative impacts on society."[288]

This external and objective scrutiny would not harm platforms' freedom to design their own AI systems or laws of flagging, nor should it affect their ability to maintain absolute discretion with regard to the way their algorithms maximize users' engagement. It should, however, ensure an independent, continuous overview of the AI system. The system of auditing would ultimately better protect the shared public interests, particularly when ex ante removal and blockage of content is involved.

Certainly, government targeting of platforms' decisions pertaining to removal and blockage of content could raise constitutional questions.[289] In particular, online platforms could argue that these efforts to regulate their decision to remove content infringes on their constitutional free-speech rights, especially in the United States.[290] A profound discussion of the state-action requirement is beyond the scope of this paper.[291] Nevertheless, it is important to note that First Amendment protection does not apply the same way in every case.[292] The extent of free-speech protection depends to a large degree on the medium and the specific action being regulated.[293] For instance, laws that target specific conduct and only incidentally burden platforms' speech may be permissible.[294] Even in instances where the law does regulate speech, courts afford different degrees of protection to different categories of speech.[295] In particular, speech concerning illegal activities or advocating violence generally does not receive the same level of protection as other constitutionally guaranteed expression.[296] Likewise, political speech

---

[288] Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* 30 (Berkman Klein Ctr. for Internet & Soc'y, 2020), https://cyber.harvard.edu/publication/2020/principled-ai [https://perma.cc/D8MT-MK4H] (quoting GERMAN FED. MINISTRY OF EDUC. & RESEARCH, FED. MINISTRY FOR ECON. AFFAIRS & ENERGY & FED. MINISTRY OF LABOUR & SOC. AFFAIRS, ARTIFICIAL INTELLIGENCE STRATEGY 38 (2018), https://ec.europa.eu/knowledge4policy/publication/germany-artificial-intelligence-strategy_en [https://perma.cc/P2ZY-F5ZD]).

[289] *See, e.g.*, Reno v. ACLU, 521 U.S. 844, 882 (1997); Heather Whitney, *Search Engines, Social Media, and the Editorial Analogy*, KNIGHT FIRST AMEND. INST. (Feb. 27, 2018), https://knightcolumbia.org/content/search-engines-social-media-and-editorial-analogy [https://perma.cc/E7TM-TJ7G]. In contrast, scholars such as Wu argue that First Amendment jurisprudence is no longer relevant due to technological changes and their effect on our speech environment. *See generally* Wu, *supra* note 190.

[290] *See generally* VALERIE C. BRANNON, CONG. RESEARCH SERV., R45650, FREE SPEECH AND THE REGULATION OF SOCIAL MEDIA CONTENT (2019), https://fas.org/sgp/crs/misc/R45650.pdf [https://perma.cc/R93V-BQF2].

[291] *See, e.g.*, Daniel J. Hemel, *Executive Action and the First Amendment's First Word*, 40 PEPP. L. REV. 601 (2013).

[292] BRANNON, *supra* note 290, at 16–17.

[293] BRANNON, *supra* note 290, at 4 (citing Stuart Minor Benjamin, *Determining What "The Freedom of Speech" Encompasses*, 60 DUKE L.J. 1673, 1680 (2011)).

[294] BRANNON, *supra* note 290, at 17 (citing Ashcroft v. Free Speech Coal., 535 U.S. 234, 253 (2002); Junger v. Daley, 209 F.3d 481, 484–85 (6th Cir. 2000)).

[295] BRANNON, *supra* note 290, at 17–18.

[296] BRANNON, *supra* note 290, at 18; *see also* Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) (per curiam).

tends to receive stronger protection than commercial speech.[297] Further, in describing the boundaries of First Amendment protection, courts generally look at whether the regulation is content neutral.[298] In addition, the special characteristics of the medium being regulated (e.g., online platforms) might entail greater regulation.[299] Therefore, the public functions of platforms as "enablers of speech and gatekeepers of information"[300] may justify some form of regulatory intervention.[301]

Overall, combining internal and external oversight mechanisms might not be sufficient to completely overcome the challenges presented throughout this paper, but it could contribute to the improved oversight of AI-based content-moderation systems.

VI. Conclusion

As AI systems proliferate in places where automated decision-making processes have never existed in the past, calls for public oversight and accountability rise out of concerns of discriminatory practices and biases. To that end, the idea of mandatory impact assessments as a tool to promote oversight and public scrutiny of AI and other automatic decision-making systems is the latest legislative trend. Recent examples include the U.S. Algorithmic Accountability Act and the EU's GDPR.

Although impact-assessment schemes carry some important advantages, our analysis suggests that they fall short of facilitating sufficient accountability. In particular, impact assessments provide only limited transparency, insufficient due process, and limited room for public review. Moreover, we find that those impact assessments might not fit the oversight challenges raised by different forms of AI-based systems, especially in the case of AI-based online content moderation by online platforms.

Content moderation by AI is rapidly becoming standard for most major platforms, consequently posing new regulatory challenges in many aspects.

---

[297] *See* Brannon, *supra* note 290, at 17–18; *see also* Citizens United v. FEC, 558 U.S. 310, 340 (2010).

[298] *See, e.g.*, Reed v. Town of Gilbert, 576 U.S. 155, 170–71 (2015).

[299] *See, e.g.*, Red Lion Broad. Co. v. FCC, 395 U.S. 367, 386 (1969) ("[D]ifferences in the characteristics of new media justify differences in the First Amendment standards applied to them."); *see also* Brannon, *supra* note 290, at 20.

[300] Quintel & Ullrich, *supra* note 269, at 20.

[301] *See* Klonick, *supra* note 28, at 1658–60; Benjamin F. Jackson, *Censorship and Freedom of Expression in the Age of Facebook*, 44 N.M. L. Rev. 121, 146 (2014); Marsh v. Alabama, 326 U.S. 502, 509 (1946) (treating a company-owned town like a state actor); Amalgamated Food Emps. Union v. Logan Valley Plaza, 391 U.S. 308, 319 (1968) (holding that a privately owned shopping center could not prevent individuals from picketing on the premises). *But see* Prager Univ. v. Google LLC, No. 17-CV-06064-LHK, 2018 U.S. Dist. LEXIS 51000, at *25–26 (N.D. Cal. Mar. 26, 2018) (holding that, by operating YouTube and restricting users' access to certain videos, Google did not engage in one of the "functions that were traditionally 'exclusively reserved to the States' " for First Amendment purposes); Shulman v. Facebook.com, No. 17-764 (JMV), 2017 WL 5129885, at *4 (D.N.J. Nov. 6, 2017) (holding that Facebook is not a state actor for First Amendment purposes).

Creating an effective and feasible tool to oversee AI systems, particularly those designed to moderate content, is a challenging task. If done poorly, the oversight might not only be deemed futile, but could also lead to serious negative consequences. This is particularly so since content moderation directly and substantially affects shared public interests. These systems create a personally tailored but fragmented segment of the public discourse, making it extremely challenging to identify and oversee content moderation by AI. Consequently, current initiatives are ill equipped to oversee AI-based content-moderation systems.

To this end, we do not argue that legislatures should discard the idea of impact assessments as a means to achieve oversight altogether. Rather, this paper provides two important contributions. First, on a general level, it highlights several shortcomings of impact assessments and proposes how to address them in order to enhance their oversight potential. Second, on a specific level, this paper shows that different contexts of AI-based decision-making systems may require different processes and levels of oversight. Specifically, to generate accountability in AI-based content-moderation systems, it is insufficient to count on self-assessment conducted by platforms, but rather it is necessary to subject them to a higher level of external and objective scrutiny.

To that end, we suggest incorporating impact assessments into a broader accountability program that ensures both internal checks and external oversight in AI-based content-moderation systems. A robust, proactive, and transparent content moderation process may comfort both firms and users and prevent lapses of trust.[302]

[302] See, for example, the outrage suffered by Facebook following a leak of over 100 documents detailing Facebook's internal content-moderation guidelines in 2017. *See supra* notes 208–09 and accompanying text.