SPEECH REGULATION BY ALGORITHM

Enrique Armijo*

ABSTRACT

The rapid convergence of speech and technology on social media platforms means it is likely the case that, either now or soon, more expressive activity will be regulated by Artificial Intelligence (AI) than by any legislature, regulator, or other government entity. Mark Zuckerberg has repeatedly told Congress and other audiences that AI is the key to resolving Facebook's content moderation challenges, envisioning a moderation regime where algorithms detect and take down speech infringing Facebook's Community Standards ex ante, that is, prior to its public posting and before it reaches other users. According to Zuckerberg, this would eventually replace its initial content moderation practices, which relied more on human moderators and user complaints than on automated detection and removal—a system that can be slow, inconsistently applied, and often subjects front-line moderators to harrowing emotional harms by exposing them to the worst of the Internet.

This Article argues that private parties' speech-regulation-by-algorithm is itself protected speech. Government efforts to regulate the content moderation of platforms will thus necessarily be barred by the First Amendment, even if that moderation is automated via AI. Nor would users whose speech has been regulated by AI have any better speech-related claims against AI-informed platform moderation decisions than they have had against nonautomated moderation, for the same reason. However, automated front-end filtering of user speech via AI is in serious tension with several core tenets of First Amendment doctrine. Ex ante AI-based content moderation operates in much the same way as a prior restraint; like government prepublication censorship, it gives users no notice of takedowns prior to publication, nor reasons for the takedown decision (at least reasons that a lay user would be capable of understanding). Additionally, it is already clear based on its current use, and the necessities of machine learning processes, that content-based speech regulation via AI is necessarily overinclusive,

* Associate Dean for Academic Affairs (through June 2021) and Professor, Elon University School of Law; Affiliated Fellow, Yale Law School Information Society Project; Faculty Affiliate, UNC–Chapel Hill Center for Information, Technology and Public Life. Thanks to Helen Norton, Margot Kaminski, Iria Giuffrida, the students of the *William & Mary Bill of Rights Journal*, and other participants at this Symposium for their comments and insights, and to Cameron Capp for research assistance. Thanks also to Michael Wooldridge and Kate Crawford, whose recent books on AI were both immensely helpful and are cited here throughout Part II. I hope I have characterized their work accurately and apologize in advance for any mistakes if I haven't.

which is normally sufficient to find a law or regulation unconstitutional under the First Amendment—a problem that will only grow worse as AI moderates more speech.

Given these concerns, this Article advocates for many of the same robust notice and procedural rights for platform users whose speech is regulated via AI, as the First Amendment compels of governments seeking to regulate private speakers.

INTRODUCTION

> [O]ver the long term, building AI tools is going to be the scalable way to identify and root out most of th[e] harmful content [on Facebook].
>
> —Mark Zuckerberg[1]

The "long term" is here. During his congressional testimony on the Cambridge Analytica scandal in 2018 and several times since, Facebook's Mark Zuckerberg has pointed to AI as the primary identification and enforcement tool for Facebook's Community Standards.[2] Implicit in Zuckerberg's technological faith in AI is an acknowledgment that primarily human review of user content will never be up to the task. There is simply too much content for human review to scale.

It is important to focus on not just Zuckerberg's invocation of AI—after all, he is a computer engineer—but also the role he envisions for it, in particular the point at which the platform would use it to moderate content. More precisely, Zuckerberg offers AI as the means by which offending content would *never reach other Facebook users at all*—in other words, to operate *ex ante* rather than *ex post*.[3] Under a primarily human-driven moderation regime, AI permits human content moderators to find and take down offending content more quickly than the traditional adjudication decision system. Rather, what Zuckerberg envisions is a system whereby AI would both identify the content and remove it from the platform before other users could see or interact with it at all.

The AI content moderation world that Zuckerberg has described is not as far off as one might think. According to Facebook's own statistics, machine learning is

---

[1] *Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection*, C-SPAN (Apr. 10, 2018) [hereinafter *Zuckerberg Hearing*], https://www.c-span.org/video/?443543-1/facebook-ceo-mark-zuckerberg-testifies-data-protection&start=6378 [https://perma.cc/W9KB-DU2T].

[2] Mike Schroepfer, *Update on our Progress on AI and Hate Speech Detection*, FACEBOOK (Feb. 11, 2021), https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ ("97% of hate speech taken down from Facebook was spotted by our automated systems before any human flagged it, up from 94% in the previous quarter and 80.5% in late 2019.").

[3] *See Zuckerberg Hearing*, *supra* note 1 (explaining that Facebook is "developing A.I. tools that can identify certain classes of bad activity proactively and flag it for our team").

playing an exponentially larger role in rooting out offensive user content.[4] Social media platform Parler was initially removed from the Apple Store and other hosting services for its failure to effectively moderate content; it has had to adopt AI-based content moderation tools as a condition of its availability to users.[5] In sum, as more speech goes online, more speech will be moderated by AI, and the transitional point between a predominantly ex ante versus ex post system, at least for the largest platforms, has already occurred.

Automated speech regulation has obvious consequences for free expression. To be sure, a platform's decisions as to what user speech to host or not is itself expressive, and thus speech protected by the First Amendment. But the move from primarily human-based to AI-based enforcement of content rules carries significant consequences for user speech. To begin with, AI is a blunt tool at best for identifying many forms of content that most platform terms of service might find infringing. The move from ex post to ex ante review of user content exacerbates that imprecision and its consequences. And the fact still remains that though human content moderation is necessarily imperfect, it still does a much better job at explaining its reasoning than machine learning does. Because of how machine processes learn, every incorrect decision that AI makes with respect to content moderation comes at the expense of user speech.

This Article proceeds as follows. To unpack the free speech-related issues applicable to algorithmic content moderation, Part I distinguishes between two types of AI used social media management: "Type I," which involves social media platforms' ordering and ranking of user content in order to optimize or decrease other users' engagement with content on their platforms, and "Type II," which involves the detection and removal of offensive user content from the platform. "Offensive" here means not just illegal speech, but also speech violative of the platforms' terms of service. Type II AI content moderation is this Article's focus. Part II addresses several issues concerning Type II algorithmic content regulation which implicate user expression on platforms, focusing on the bluntness of AI as a moderation tool and the lack of notice-based issues inherent to Type II moderation—issues that raise both speech and due process concerns that are exacerbated by the machine learning process. Part III proposes some remedies to those Type II–related problems, highlighting the role of Facebook's Oversight Board in proposing policies to Facebook that would, in the Board's view, give users more notice and process around Type II takedown decisions. Part III also looks to Type II content moderation as a decision-making model and proposes several interventions to ameliorate user concerns about over-moderation while still preserving for platforms the ability to act against disinformation and harmful content in an expedient fashion.

---

[4] *See infra* text accompanying notes 46–52.

[5] Kevin Randall, *Social App Parler Is Cracking Down on Hate Speech—but Only on iPhones*, WASH. POST (May 17, 2021, 4:41 PM), https://www.washingtonpost.com/technology/2021/05/17/parler-apple-app-store/ [https://perma.cc/5SRJ-6Z2H].

I. TYPE I VS. TYPE II ALGORITHMIC CONTENT MODERATION

Before addressing the issues that speech regulation by AI raises, it is first necessary to identify what kind of speech regulation by AI one means, because different AI moderation systems present different challenges. In a recent article,[6] Tim Wu distinguished between what he called platforms' "positive" (or "affirmative"; he uses both terms interchangeably) and "negative" algorithmic speech control. Positive speech control "entails choosing what is brought to the attention of the user [and] is found in the operation of search results, newsfeeds, advertisements, and other forms of promotion and is typically algorithmic."[7] By contrast, negative speech control "consists of removing and taking down disfavored, illegal, or banned content, and punishing or removing users."[8] Wu also notes that positive speech controls—including Google search results, YouTube video auto-play selections and recommendations, and Facebook news feed and Twitter timeline ordering—have been predominantly algorithmic since their inception. Conversely, negative speech controls "were originally complaint-driven" and administered by humans, until the scale of content review compelled platforms to automate more and more moderation, initially at the screening level and now at the point of takedown as well.[9]

Wu's distinction is useful, but his terminology potentially confuses. "Positive" content moderation, as he notes, involves not only bringing certain content to users' attention to attempt to drive engagement, but also downgrading, de-promoting, and disadvantaging of content that the algorithmic system deems low quality but not violative of use rules.[10] In other words, when sorting user feeds and other user-platform interactions, "positive" content moderation treats user content both positively and negatively.[11] And "negative" content moderation involves not only the removal of

---

[6] Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2014 (2019).

[7] *Id.*

[8] *Id.*

[9] *Id.* at 2015.

[10] *See, e.g.*, Nick Clegg, *You and the Algorithm: It Takes Two to Tango*, MEDIUM (Mar. 31, 2021), https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2 [https://perma.cc/KD3V-XDXD] (discussing how Facebook "reduces [ ] distribution" of problematic "but non-violating" content); Josh Constine, *Facebook Will Change Algorithm to Demote "Borderline Content" That Almost Violates Policies*, TECHCRUNCH (Nov. 15, 2018, 3:21 PM EST), https://techcrunch.com/2018/11/15/facebook-borderline-content/ [https://perma.cc/4CQD-DPUK].

[11] See Sang Ah Kim, *Social Media Algorithms: Why You See What You See*, 2 GEO. L. TECH. REV. 147 (2017), for an excellent explanation of engagement-based algorithmic content moderation and its motivations. One fundamental issue for Type I content moderation is the fact that, at least as to Facebook, the most engagement-maximizing content is also the content most likely to violate Facebook's Community Standards. In other words, Type I algorithms "that maximize engagement reward inflammatory content." *See* Karen Hao, *How Facebook*

offending content but also leaving it up after an algorithmic or human moderator determines that it does not violate a platform's terms of service.[12] Accordingly, this Article preserves the distinction but uses "Type I" to describe algorithmic ordering and ranking and "Type II" to describe the detection, assessment, and potential removal of potentially offending content.

A primary reason for disaggregating the types of algorithmic speech regulation is because the rationale and capacity for government intervention is likely to change depending on whether the practice sought to be regulated is either Type I or Type II. To start, the goals of each type of content moderation are distinct; as noted, Type I regulates user content to optimize user engagement, while Type II regulates user content to enforce rules concerning what is permissible on the platform. This distinction may be of constitutional significance. Free speech scholars, including Helen Norton in her contribution to this Symposium, have argued that Type I content moderation is less "speech-like" in a conventional First Amendment sense, either because algorithmic speech ordering is not itself speech[13] or because, as Norton argues, its primary intent is to manipulate other users,[14] and thus it should be more amenable to regulation to prevent or minimize harms associated with its use. Indeed, in light of the role that social media may have played in disseminating information related to the Capitol insurrection on January 6, 2021, legislators at the federal level have begun taking steps to try and regulate platforms' uses of algorithmic amplification.[15]

There are good reasons, however, to focus on Type II content moderation as well. First of all, though the scope of Type II regulation has increased exponentially, it has a pedigree similar to that of Type I in domains outside of moderation of content that the platform deems offensive. Platforms have long used automated processes,

---

Got Addicted to Spreading Misinformation, MIT TECH. REV. (Mar. 11, 2021), https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/ [https://perma.cc/F4U2-NPP2].

[12] Wu, *supra* note 6.

[13] *See* Tim Wu, *Machine Speech*, 161 U. PA. L. REV. 1495, 1498 (2013) [hereinafter Wu, *Machine Speech*] ("Too much protection would threaten to constitutionalize many areas of commerce and private concern without promoting the values of the First Amendment."). *But see* Stuart Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445 (2013).

[14] *See* Helen Norton, *Manipulation and the First Amendment*, 30 WM. & MARY BILL RTS. J. 221 (2021).

[15] *See, e.g.*, Taylor Hatmaker, *At Social Media Hearing, Lawmakers Circle Algorithm-Focused Section 230 Reform*, TECHCRUNCH (Apr. 27, 2021, 5:14 PM), https://techcrunch.com/2021/04/27/section-230-bills-algorithms-congress-hearing/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAEHe1peg3qAj5lTYVmEjSdd71g1UCzL852QIHd8IXK6b67oOIugsJ9sB79zsn2k0fYGUgyz0HXUsWibSnaYsxB5RHVo3U17v8pdcmW4nvkVKIJwueTsFfhDI9Yc502wEVo1UyaqFiKOOHOj9EG8fzL8UYrBm5szS5qbf5PKXjI3b [https://perma.cc/V7KF-6C4Y]; Protecting Americans from Dangerous Algorithms Act, H.R. 2154, 117th Cong. (2021); Algorithmic Justice and Online Platform Transparency Act, S. 1896, 117th Cong. (2021).

such as Content ID[16] and the Digital Millennium Copyright Act[17] to enforce their IP-related terms of service. Further, as noted in the introduction above, a marked increase in the use of Type II regulation by platforms is inevitable given the scale problem: there is no way to moderate user content on Facebook or YouTube without some form of automated review. Additionally, pressures from governments in the United States and abroad are pushing platforms to move Type II regulation further back in the user experience—to the point where regulators increasingly expect platforms to, as Hannah Bloch-Wehba has shown, "prevent the dissemination of unlawful online content before it is ever seen or distributed."[18] And most importantly for present purposes, moderation of offensive user content simply looks more like speech than algorithmic engagement optimization, due to the fact that, as discussed in more detail at Section II.A below, the decisions that platforms make with respect to what content "belongs" on their platforms is inherently expressive.[19] Accordingly, as Type II content moderation moves from a predominantly human-centered form of response to user complaints and assessment of the appropriateness of that content to an algorithmic screening of user content ex ante, Type II problems look more like conventional speech problems as well. I now turn to discuss these problems in greater detail.

---

[16] Wu, *supra* note 6, at 2007 (citing *How Content ID Works*, YOUTUBE HELP, https://support.google.com/youtube/answer/2797370?hl=en [https://perma.cc/WV5R-CBUR]).

[17] JENNIFER M. URBAN ET. AL, NOTICE & TAKEDOWN IN EVERYDAY PRACTICE 8 (2016) (describing the move in DMCA's notice-and-takedown regime from predominately human to reliance on computer algorithms to both detect potential infringements and generate notices.).

[18] Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 41, 42–43 (2020); *see also* Carey Shenkman et al., *Do You See What I See?: Capabilities and Limits of Automated Multimedia Content Analysis*, CTR. FOR DEMOCRACY & TECH. (May 2021), https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf [https://perma.cc/8EEL-DCFW].

By use of the term "further back," I have in mind a three-step user experience on a platform that involves: (1) the submission of content by one user, (2) the posting of that content by the platform to which the user submitted the content, then (3) the viewing of that content by a second user. Moving content moderation "further back" would move both the assessment of whether the content violated the platform's user rules, and its takedown if it did, from after Step 3 to between Steps 1 and 2. As discussed in Part II, *infra*, this move raises due process concerns.

[19] *See* Enrique Armijo, *Reasonableness as Censorship: Section 230 Reform, Content Moderation, and the First Amendment*, 73 FLA. L. REV. 1199, 1277 (2021) [hereinafter Armijo, *Reasonableness as Censorship*]. For an argument to the contrary, namely that the ranking and ordering of content may be more constitutionally protected than the hosting of user speech, see Eugene Volokh, *Treating Social Media as Common Carriers?*, 1 J. FREE SPEECH L. 377 (2020).

II. TYPE II PROBLEMS

A. Content Moderation as Expressive (Human or Algorithmic)

I begin with a proposition that should not be closely contested: Type II content moderation decisions by platforms and websites, whether undertaken by humans or algorithms, are protected from government regulation by the First Amendment.[20] The "whether undertaken by humans or algorithms" part of that proposition requires some further argument in support.

Content moderation policies themselves—the speech rules by which users are bound as conditions for their use of the platform in question—are expressive. Platforms are motivated to moderate user speech to ensure that the platforms and the speech they host are consistent, or at least not in conflict, with their values and goals.[21] Accordingly, Facebook's Community Standards, which the platform crafts "to ensure that everyone's voice is valued" and to be "inclusive of different views and beliefs," define "Objectionable Content" to include, for example, "Hate Speech," and proceed to define that term as a "direct attack against people . . . on the basis of what [it] calls protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease."[22] Parler's Community Guidelines states the platform's goal is "to provide all community members with a welcoming, nonpartisan Public Square," and that its mission is "to create a social platform in the *spirit* of the First Amendment to the United States Constitution," and so it claims to moderate content in a "viewpoint-neutral" way.[23] Twitter's Rules state their "purpose is to serve the public conversation," that "[v]iolence, harassment and other similar types of behavior discourage people from expressing themselves, and ultimately diminish the value of global public conversation," and that the Rules are "to ensure all people can participate in the public

---

[20] It may be more accurate to say they are "covered" rather than "protected," i.e., that the First Amendment applies, but that the government could potentially overcome a First Amendment–dependent defense with a justification and sufficiently tailored means to achieve that end that would survive judicial review. *See generally* Frederick Schauer, *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 HARV. L. REV. 1765 (2004).

[21] Bloch-Wehba, *supra* note 18, at 53 (explaining that undesirable activity has driven online communities to develop rules and restrictions to avoid damage to the online community and depletion of resources).

[22] *Community Standards: Introduction*, FACEBOOK, https://www.facebook.com/communitystandards/introduction [https://perma.cc/2Q8Z-GX83] (last visited Dec. 13, 2021); *Community Standards: Objectionable Content*, FACEBOOK, https://www.facebook.com/communitystandards/objectionable_content [https://perma.cc/6JEB-32T5].

[23] *Community Guidelines*, PARLER (Feb. 14, 2021), https://legal.parler.com/documents/guidelines.pdf [https://perma.cc/R8PC-ZD9P].

conversation freely and safely."[24] The Rules then set out categories of content that are not permitted due to their conflict with the values of "safety," "privacy," and "authenticity."[25] The identification of these values, and the process of defining them, are quintessentially expressive; platforms are deciding on and defining the user expression they want to be associated with or not. This is true of any use rule that is content-based. Content-based use rules obviously express a preference for a certain kind of content.

If the values underlying a platform's use rules are expressive, then it follows that the act of banning users or removing content that fails to conform with those values and goals is expressive as well—the enforcement act expresses content preferences by disassociating the platform from user speech that expresses a conflicting view. In the non-social media context, this process is uncontroversially understood to be a form of editing. As courts have consistently found in rejecting claims brought against platforms by banned users, content moderation is an exercise of editorial discretion.[26]

Given that, a case that algorithmic content moderation is less constitutionally protected must rest on the argument that the *process* by which content moderation occurs can reduce its protection. In other words, the automation of moderation—reducing the role of a person or people in the act of assessing and removing content—changes the First Amendment analysis.[27] To be sure, the argument has some appeal; extending constitutional protection via the Speech Clause to machine processes seems to stray from the core human-based (or humanity-based[28]) expressive acts and decisions with which First Amendment doctrine has long been most concerned. But the argument relies on a misapprehension as to how algorithmic content moderation actually operates.

To understand why the fact of automation does not change the First Amendment analysis with respect to content moderation, it helps to understand AI, or at least the

---

[24] *The Twitter Rules*, TWITTER, https://help.twitter.com/en/rules-and-policies/twitter-rules [https://perma.cc/2JGC-ZZ86] (last visited Dec. 13, 2021).

[25] *Id.*

[26] *See, e.g.*, Adi Robertson, *Social Media Bias Lawsuits Keep Failing in Court*, THE VERGE (May 27, 2020, 5:43 PM), https://www.theverge.com/2020/5/27/21272066/social-media-bias-laura-loomer-larry-klayman-twitter-google-facebook-loss [https://perma.cc/8TUG-CF6G].

[27] *See* Tim Wu, *Free Speech for Computers?*, N.Y. TIMES (June 19, 2012), https://www.nytimes.com/2012/06/20/opinion/free-speech-for-computers.html [https://perma.cc/U6E8-YHXV] ("To give computers the rights intended for humans is to elevate our machines above ourselves."); Wu, *Machine Speech*, *supra* note 13, at 1517–18, 1521–23 (arguing that under the First Amendment functionality doctrine, AI communication tools perform tasks unrelated to the communication of ideas and are therefore exempt from free speech protection); Niva Elkin-Koren & Maayan Perel, *Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 LEWIS & CLARK L. REV. 857, 862–63 (2020).

[28] *See, e.g.*, C. EDWIN BAKER, HUMAN LIBERTY AND FREEDOM OF SPEECH (1989); STEVEN J. HEYMAN, FREE SPEECH AND HUMAN DIGNITY (2008).

type of AI used in Type II content moderation, as an application of agency theory. As is the case in legal theory, the study of AI by computational theorists has also used the language of agency, and the term is used similarly in both contexts.

For AI theorists, an "agent-based" AI system is one that is "'complete' in the sense that it [is] a self-contained, autonomous entity, situated in some environment and carrying out some specific task *on behalf of a user.*"[29] Among other influential AI thinkers, MIT roboticist Rodney Brooks was instrumental in arguing against the knowledge-based approach that conceptualized AI systems as "disembodied . . . logical reasoners" that predominated before the 1980s.[30] Instead, Brooks argued instead in favor of a behavioral approach that ties machine learning to its instructions and environment.[31] Decades before the implementing hardware associated with AI reached the technical capacity to do so on a large scale, robotics theorists were describing agent-based algorithmic processes that would eventually automate "email management, meeting scheduling, filtering news, and music recommendation" for millions of human users.[32]

To be clear, these processes were not following programmer instructions in a rote fashion; as the decision-making context changed, the AI agent would "start to take the initiative and process [the instructed function] according to its prediction" and based on what it had learned.[33] But the processes' predictions were as to *how* to reach the result the user had chosen based on the user's preferences, not the actual result itself.[34] When faced with a decision, an automated agent "chooses an action whose outcome maximizes utility on [its user's] behalf; which is the same as saying that it chooses an action in order to bring about [the user's] most preferred outcome."[35] These processes involve applying inferences to new facts, but machine learning inferences are what Kate Crawford calls "inductive inference[s]," or a "hypothesis based on available data," rather than "deductive inference[s]," the type which "follow[] logically from a premise" that are the product of human knowledge and

---

[29] MICHAEL WOOLDRIDGE, A BRIEF HISTORY OF ARTIFICIAL INTELLIGENCE: WHAT IT IS, WHERE WE ARE, AND WHERE WE ARE GOING 93 (2021) (emphasis added). I realize that this Part uses the term "user" to alternate between (1) a "user" of a machine learning system to execute a task that the user needs done, and (2) a "user" of a social media platform. I apologize for the confusion and hope the context makes clear which "user" I mean in a sentence of mine or a quotation or paraphrase from another source.

[30] *Id.*

[31] *Id.* (discussing Brooks' work and influence).

[32] *Id.* at 96–97 (citing Pattie Maes, *Agents that Reduce Work and Information Overload*, 37 COMMC'NS OF THE ACM, 30–40 (1994)).

[33] *Id.* at 97.

[34] *See id.* at 99–100 (describing optimal decision-making theory in agent-based AI as based around "preferences" and the principle that "[i]f your agent is to act on your behalf, then it needs to know what your wishes are. You then want the agent to act in order to bring about your preferred choices as best it can.").

[35] *Id.* at 100.

intuition.[36] The sum of what the machine learning system "knows" is only what the principal has either taught it directly (in the form of a training dataset of examples compiled by the principal) or taught it to learn (instructions as to how to classify, based on the training dataset, a new example that was not part of the that dataset).[37] The behavioral grounding of AI agency theory does not reject the concept of autonomous knowledge development entirely, but even the system's accumulation of knowledge and its application of that knowledge to new facts is in service to the principal's expressive goal. That's what makes the system an agent that executes the expressive intent of another, and not a principal that executes its own.

In other words, to return to the legal context, "[t]he fact that an algorithm is involved [in effectuating a content-related decision] does not mean that a machine is doing the talking."[38] As is the case in legal agency theory, the agent (the AI system that moderates content) is effectuating decisions on behalf of the principal (the platform using the AI system), not making its own decisions.[39] The expressive decision of *what* to moderate is a protected editorial decision that rests with the principal, not the agent.[40] The fact that AI *assists* in effectuating that decision does not affect the constitutional analysis. And this is true whether the AI the platform is using for Type II content moderation is "learning" the difference between permissible and barred content through training examples before operating on the platform, or on actual user posts once the AI is in the act of moderating those posts.[41] The machine learning process also inevitably involves implementing the principal's decision incorrectly by putting a piece of user content in the wrong category, and the process is then trained to self-correct once it is instructed of its error.[42] But this is simply agency theory in operation—after learning of an implementation mistake, the AI agent is being reinstructed with respect to the principal/user's expressive goals.[43] Nothing

---

[36] KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 97 (2021).

[37] *Id.*; *see also id.* at 134.

[38] *See* Benjamin, *supra* note 13, at 1479.

[39] RESTATEMENT (SECOND) OF AGENCY § 1 (AM. L. INST. 1958).

[40] Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (First Amendment protects platform's ability to "decid[e] [which content] to publish, withdraw, postpone or alter"); Pittsburgh Press Co. v. Pittsburgh Comm'n on Hum. Rel., 413 U.S. 376, 391 (1973) (reaffirming the First Amendment speech "protection afforded to editorial judgment"); Miami Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 258 (1974).

[41] "Training examples" in the automated content moderation context would include examples of terms-of-service-infringing content in image form, or repeating patterns of certain infringing word and phrase usages in the case of text. *See, e.g.*, TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 98–101 (2018) (explaining training examples for AI content moderation systems and how they lead to false positives).

[42] As Kate Crawford notes, error-checking and correcting maintenance in AI systems is often done by humans, not other machines—predominantly underpaid and exploited humans. *See* CRAWFORD, *supra* note 36, at 66–68.

[43] *See id.*

in the fact that the agent makes implementation mistakes makes the agent's actions more regulatable because those mistakes, just like the instances in which the principal's will is executed correctly, are not the product of the agent's expressive choice.

Accordingly, automated content moderation does not undertake any expressive decisions of its own, but rather effectuates the expressive decisions of its user. It may do so at a much faster and more replicable rate than the user, but the expressive intent underlying the decision remains the same.[44]

B. The Rise of Type II Algorithmic Content Moderation and Algorithms as Blunt Instruments

A fully realized Type II content moderation world such as the one Mark Zuckerberg described to Congress—an algorithmic system that can preemptively detect, without the need for a triggering complaint, not just existing categories of problematic content like nudity, child pornography, or material infringing copyright, but also more subjective categories like hate speech, incitement, and harmful disinformation—does not yet fully exist, even on Facebook. There is no doubt, however, that on the largest platforms the move from mostly human to Type II moderation has already begun.[45] Facebook is touting AI's improved effectiveness in proactively detecting offending content—"proactively" here meaning that the AI has detected the offending content without having to wait for a user report.[46] Even with respect to context-based content like hate speech, Type II is playing an increasingly larger role in content moderation, which is what one would expect as the technology develops.[47]

---

[44] And to the extent the decisions the algorithm is instructed to effectuate are implicitly biased, the outputs will reflect that bias as well—and bias is itself expressive. *See id.* at 135 ("Every dataset used to train machine learning systems, whether in the context of supervised or unsupervised machine learning, whether seen to be technically biased or not, contains a worldview. To create a training set . . . requires inherently political, cultural, and social choices."). *See also* Armijo, *Reasonableness as Censorship*, *supra* note 19, at 20–21; Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017); Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, 2019 ASS'N COMPUTATIONAL LINGUISTICS 1668 (establishing strong connections between use of "African American English dialect" used by "self-identifying African American users" and "toxicity annotations" by automated hate speech categorizers).

[45] This move was exacerbated by the Covid-19 pandemic, which caused platforms to have to "drastically scal[e] back human moderation and increase[ ] reliance on AI." Evelyn Douek, *Governing Online Speech: From "Posts-as-Trumps" to Proportionality & Probability*, 121 COLUM. L. REV. 759, 802 (2021).

[46] Mike Schroepfer, *How AI Is Getting Better at Detecting Hate Speech*, FACEBOOK AI BLOG (Nov. 19, 2020), https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/ [https://perma.cc/7HNP-BUDL] (noting that its AI content moderation system's proactive detection rate rose from 24% of all hate speech removed from the platform in the fourth quarter of 2017 to 95% in the third quarter of 2020).

[47] Douek, *supra* note 45, at 793–94.

But even a fully realized ex ante Type II content moderation system would remain a blunt instrument with respect to many kinds of user speech. As noted, machine learning is trained on the familiar: it identifies problematic content only to the extent its users have supplied it with examples that are substantially similar.[48] And text-based data sets alone are not well-suited for making the contextual considerations that human judgment undertakes when assessing factors like a speaker's intent or motive.[49] Innovation in toxicity online under a Type II system, in other words, has compounding returns. New forms of toxic content are much more difficult to detect and remove and are thus more likely to go viral before the platform can intervene.[50]

When considering machine learning as applied to speech problems,[51] one also sees how another concept from First Amendment doctrine, the concept of over-inclusiveness, pervades the decision-making process. To the extent a Type II system

---

[48] *See, e.g.*, Laura Hanu, James Thewlis & Sasha Haco, *How AI Is Learning to Identify Toxic Content Online*, SCI. AM. (Feb. 8, 2021), https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/ [https://perma.cc/VZ5S-8QUY] (describing how toxic speech–detecting algorithms that use text classification models work well "on examples that are similar to the data they have been trained on[, b]ut they are likely to fail if faced with unfamiliar examples of toxic language."); GILLESPIE, *supra* note 41; *see* Arcadiy Kantor, *Measuring Our Progress Combatting Hate Speech*, FACEBOOK NEWSROOM BLOG (Nov. 19, 2020) [hereinafter Kantor, *Measuring Our Progress*], https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/ [https://perma.cc/Z6KU-HD7E] ("[d]efining hate speech isn't simple, as there are many differing opinions on what constitutes hate speech. Nuance, history, language, religion, and changing cultural norms are all important factors to consider as we define our policies.").

[49] Monika Bickert, *European Court Ruling Raises Questions About Policing Speech*, FACEBOOK NEWSROOM BLOG (Oct. 14, 2019), https://about.fb.com/news/2019/10/european-court-ruling-raises-questions-about-policing-speech/ [https://perma.cc/B9DT-9MXJ] (stating that "[w]hile [Facebook's] automated tools have come a long way, they are still a blunt instrument and unable to interpret the context and intent associated with a particular piece of content," and that "[d]etermining a post's message is often complicated, requiring complex assessments around intent and an understanding of how certain words are being used."); Shenkman et al., *supra* note 18, at 15 ("An important consideration in utilizing matching-based systems is their general inability to assess context. The same pieces of content . . . in one context may have significant expressive and public interest value in a different setting, such as in art, academic or journalistic work, or human rights commentary."); Natasha Duarte et al., *Mixed Messages? The Limits of Automated Social Media Content Analysis*, CTR. FOR DEMOCRACY & TECH 1, 8 (Nov. 2017) [hereinafter Duarte et al., *Mixed Messages?*], https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf [https://perma.cc/Q8LS-45YH] (proposals calling for or requiring platforms to automate content moderation "wrongly assume that automated technology can accomplish on a large scale the kind of nuanced analysis that humans can accomplish on a small scale."); *see* Deepa Seetharaman et al., *Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts.*, WALL ST. J. (Oct. 17, 2021, 9:17 AM), https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184 [https://perma.cc/6UHH-XCQX].

[50] *See* Kantor, *Measuring Our Progress*, *supra* note 48 ("Language continues to evolve, and a word that was not a slur yesterday may become one tomorrow.").

[51] *See* Armijo, *Reasonableness as Censorship*, *supra* note 19, at 20–21.

learns from its mistakes, the cost of those mistakes, in the form of false positives, results, at least initially, in less speech. For example, take an algorithmic system that has learned how to screen for and take down livestreams of gun violence. That system would have prevented the Christchurch murders in March 2019 from being posted, reposted, and viewed by thousands of Facebook and Facebook Live users. But it would likely also detect and preemptively block streams showing police shootings or their immediate aftermaths, such as the shootings of Philando Castile in St. Paul and Sean Reed in Indianapolis.[52] In such a scenario, assume the system initially blocks both posts, but the content moderator overseeing the system wants users to be able to livestream police brutality, and so a human intervention reorients the machine learning system to permit the second type of case (assuming, of course, that such a distinguishing intervention is actually possible). So, every modification of the machine learning system in favor of more speech requires an initial speech harm. In other words, the Type II system only makes the "both barred" mistake once, learns from it, and then learns to leave the second category of speech up. But the cost of that learning case is that user speech that the platform values does not reach other users, at least until the error is recognized as such and corrected.

In addition to those substantive problems, Type II systems present procedural challenges as well. The traditional role of automation in content moderation has been to flag potentially offending content on the platform for a human moderator's review, at which point the human decides whether the content stays or is removed.[53] But if a true ex ante Type II system deems user speech as contrary to a platform's terms of service, the consequence of that classification is removal from the platform before that content is publicly posted.

Indeed, it might invite imprecision to use the term "removal" in this circumstance at all. In the analogous First Amendment context, we would consider this conduct by the decision-maker a prior restraint; the speech in question is prevented from being disseminated because a government decision maker has concluded the speech offends the relevant standard, and the primary harm is procedural in nature

---

[52] *See* Fred Ritchin, *In the Livestream Era, "the Trauma Is Widespread*," TIME (July 11, 2016, 12:47 PM), https://time.com/4400930/philando-castile/ [https://perma.cc/ZWX5-KXKE]; Ryan Martin & Tony Cook, *Indianapolis Police Fatally Shoot Man After a Chase Possibly Broadcast on Facebook Live*, INDIANAPOLIS STAR (May 6, 2020, 8:08 PM), https://www.indystar.com/story/news/crime/2020/05/06/police-shooting-reported-on-the-north-side/5180240002/ [https://perma.cc/5M3G-D6HP]. Additionally, some evidence indicates that the production of false positives is itself also biased—i.e., that algorithms can have higher false positive rates for under-represented speakers than for white ones. *See* Duarte et al., *Mixed Messages?*, *supra* note 49, at 19 (citing 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/GSA5-UMV5]). I have illustrated the overinclusiveness problem with respect to content moderation using these two examples before. *See* Armijo, *Reasonableness as Censorship*, *supra* note 19, at 20–21.

[53] *See* GILLESPIE, *supra* note 41, at 97.

because the speaker is deprived of the opportunity to argue that the decision-maker's interpretation of the standard and classification of the speech under it is wrong. We do not think of the decision-maker's post-classification action in the traditional prior restraint context as a "removal," since the speech has not been disseminated (that's what "prior" in prior restraint means—prior to dissemination of the speech in question). Likewise here, the algorithm has prevented the speech from reaching the platform at all, not "removed" it once it has been posted and classified as offending. Removal is an *ex post* intervention by a speech regulator, not an *ex ante* one.

In fact, if one puts aside the state action question and focuses solely on the harm to speech under each system, this form of digital prior restraint is even worse for the speaker than the conventional one; in the latter example, at least a speaker in most cases knows their speech is being restrained. In an ex ante Type II system, by contrast, the restriction happens behind the speaker's screen, so absent some operational constraint rule requiring notice, they may not know their speech is being restricted at all.[54] Additionally, unlike in a conventional prior restraint system, where the party imposing the classification is (at least nominally) applying some standard for harmful speech promulgated by another party, here the platform has the sole authority to decide both (1) what constitutes a violation, and (2) the manner by which violations are penalized and their implementation.[55]

And all these procedural problems are compounded by the fact that a computational decision-maker is good at explaining *how* it makes decisions, but not *why*. As AI theorist Michael Wooldridge writes, "one feature of the current wave of AI systems is that they are black boxes: they cannot *explain* or *rationalize* the decisions they make in the way that a person can."[56] Embedded throughout free speech doctrine, as well as the emerging commentary around Type II social media content moderation and decision-making more generally,[57] is a user's right to explanation—social media

---

[54] The digital prior restraint problem is also spreading further down the Internet protocol stack, from the application level to the Internet access level. See Mark Lemley, *The Splinternet*, 70 DUKE L.J. 1297, 1315–16 (2021), stating:

> There are increasing moves by companies and internet service providers ("ISPs") to filter malicious sites at the domain-name system ("DNS") level so that they are never accessible at all, even on your server system. Not that you just don't see them on your device. Your corporate server never sees them either. The DNS routing system pretends that site on the internet simply doesn't exist. If you try to send a message to it, you will not get a response.

[55] *See, e.g.*, *How Technology Detects Violations*, FACEBOOK TRANSPARENCY CTR. (June 28, 2021), https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/ [https://perma.cc/5HPF-PECD] ("We remove millions of violating posts and accounts every day on the Facebook app and Instagram. Most of this happens automatically, with technology working behind the scenes to remove violating content—often before anyone sees it.").

[56] *See* WOOLDRIDGE, *supra* note 29, at 199 (emphasis in original).

[57] *See, e.g.*, Rory Van Loo, *Federal Rules of Platform Procedure*, 88 U. CHI. L. REV. 829,

platforms whose speech is banned or limited have a process-based right to understand the basis on which the platform found their content to be infringing. To the extent AI is used to effectuate those decisions, those explanations will be increasingly difficult to provide.[58]

III. POSSIBLE REMEDIES FOR TYPE II PROBLEMS

A. Facebook's Oversight Board as Maker of Adjudicatory Policy in Type II Decisions

The best way to mitigate against the problems inherent to Type II AI content moderation is to develop a set of obligations for platforms using the system that would remedy those problems, and to conceptualize where in the system they might be placed. I refer here to information-forcing side constraints that would necessarily cut some against the efficiency of AI, but would improve system procedure and results for users in exchange for the efficiency loss.[59] In the case of Facebook, the platform's Oversight Board is already developing a role for itself in fashioning these constraints.[60] One of the Board's first cases involved Type II content moderation, and the Board took an information-forcing approach to try and resolve broader Type II–related challenges that the case raised.[61]

867–68 (2020) [Section III.B.1] (arguing that platforms should exercise their adjudicatory discretion in accordance with common law decision-making processes, which would give users the ability to "predict likely outcomes and argue their cases based on prior decisions").

[58] *See* Shenkman et al., *supra* note 18, at 33–34 (AI decision-making processes "are complex and non-linear, and do not necessarily 'show their work,' which makes it very difficult to understand how they operate, what features they use to make decisions, and how various decisions are weighted and why.") (citing Gabriel Eilertsen et al., *Classifying the Classifier: Dissecting the Weight Space of Neural Networks*, ARXIV (2020)).

[59] Thomas Nachbar has discussed side constraints, or a "rule that ignores the goals of the system on which it operates because it is in the service of some other goal," in the context of algorithmic decision-making. Thomas Nachbar, *Algorithmic Fairness, Algorithmic Discrimination*, 48 FLA. ST. L. REV. 509, 525 (2021) (citing ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 28–29 (1974)). For Nachbar, contested, open-ended concepts like fairness or transparency can only operate as side constraints to the primary goal of the computational process, that is, the problem it is designed or optimized to solve. *See id.* at 527 ("We can constrain the operation of lending algorithms in certain ways to satisfy the demands of fairness (at least the ones we can agree on) even if we can't program them to produce optimally fair outcomes without converting them into fairness algorithms instead of lending algorithms."). I do not disagree that concepts like notice and explainability are contested, but I use the concept of side constraints to illustrate how serving those values might operate in both tandem and in tension with automated content moderation processes.

[60] Facebook Oversight Board Charter, art. 3 §§ 1–7, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [https://perma.cc/U3VM-L3W2].

[61] *Case Decision 2020-004-IG-UA*, OVERSIGHT BD. (Jan. 28, 2021), https://oversightboard.com/decision/IG-7THR3SI1/ [https://perma.cc/6PYE-KSSQ].

The formation of Facebook's Oversight Board has been discussed in detail elsewhere.[62] For present purposes, the most important facet of its remit is that while the Board's decisions with respect to individual moderation decisions are binding, its recommendations with respect to policy, which are called "Policy Advisory Statements," are, as the name implies, not binding, though Facebook must respond to its recommendations within a set period of time.[63] It is in this latter role that the Board is attempting to develop information-forcing side constraints around AI content moderation.

In October 2020, Facebook's automated AI image screening system took down an image that included visible female nipples that was posted by a Brazilian user as part of a breast cancer awareness campaign on Instagram.[64] The system did so because it deemed the post in violation of the platform's Community Standard on Adult Nudity and Sexual Activity. The Board noted that "the detection and removal of this post was entirely automated."[65] In the Policy Advisory Statement accompanying its opinion overruling the platform's decision, the Board stated that Facebook should "[i]nform users when automation is used to take enforcement action against their content, including accessible descriptions of what this means"[66]—i.e., to tell users whose content is removed by AI the "why," not just the "how." The Board said the platform should also "[i]mplement an internal audit procedure to continuously analyze a statistically representative sample of automated content removal decisions to reverse and learn from enforcement mistakes," and "[e]xpand transparency reporting to disclose data on the number of automated removal decisions per Community Standard, and the proportion of those decisions subsequently reversed following human review."[67] In other words, and in response to the explainability

---

[62] *See, e.g.*, Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418 (2020).

[63] *See* Facebook Oversight Board Charter, Introduction, art. 1, §§ 4, 6, art. 3, § 4, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [https://perma.cc/PRU4-X9KN] (Board's purpose is, *inter alia*, to "issu[e] policy advisory opinions on Facebook content policies" and it will issue "policy recommendations" "specific to a case decision or upon Facebook's request," which "will be taken into consideration by Facebook to guide its future policy development."); Oversight Board Bylaws, 2.3.2, https://www.oversightboard.com/sr/governance/bylaws [https://perma.cc/3M5N-Z2HG] ("When the board chooses to issue a policy advisory statement, . . . Facebook will provide a public response regarding [same] and any follow-on action within thirty (30) days of the recommendation being received.").

[64] *Case Decision 2020-004-IG-UA*, OVERSIGHT BD. (Jan. 28, 2021), https://oversightboard.com/decision/IG-7THR3SI1/ [https://perma.cc/9U4B-SXM8].

[65] *Id.*

[66] *Id.*

[67] *Id.* The audit recommendation probably does not do too much work, since what the Board has done here, albeit likely unknowingly, is simply describe the process of machine learning. Any "internal audit" data based on machine learning moderation decisions therefore likely already exists. *See supra* Part II.

issue discussed above, the Board is advising Facebook that it should give its users what adjudications literature calls "a right to a human decision,"[68] what my fellow Symposium contributor Margot Kaminski and other law-and-technology scholars have called "explainable AI,"[69] and what the European Union has begun to compel of platforms through its General Data Protection Regulation.[70] To repeat, Facebook does not have to adopt the Board's policy recommendations, but it must consider and respond to them; in its response to these recommendations concerning AI-based content moderation, the platform committed to "assessing the feasibility" of more transparency around automated content moderation and greater notice to users around whether their content was removed or otherwise found to violate terms of service by an algorithm rather than a human reviewer.[71]

The precise nature of the information-forcing side restraints that Facebook chooses or is forced though regulation to develop and implement for AI-based content moderation remains to be seen. What is notable about this colloquy between Facebook and the Oversight Board for present purposes is its direction. It moves AI content moderation further from, not toward, the ex ante Type II reality that Mark Zuckerberg asked Congress to imagine in 2018.[72] The need for human intervention in algorithmic content moderation at several points in the decision-making model seems more pressing, not less.

B. The Inevitable Role of Humans in Type II Decisions

Since some human role in speech regulation by algorithm seems inevitable, the last question for this Article to raise is where the human intervention in a Type II system might best go. Sensitivity to the scale problem likely leads at least part of the way to Mark Zuckerberg's position: there must be automated moderation at the

---

[68] *See generally* Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611 (2020). While citing it for the general proposition articulated in its title, I should note that Aziz Huq argues in this piece that the flaws inherent to algorithmic decision-making (or at least the resolvable flaws) call more for a right to a *better* machine learning decision than a right to require that humans decide. *See id.* at 686–88.

[69] *See generally* Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189 (2019) (discussing the need for algorithmic accountability through explainable AI).

[70] Council Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, 2016 O.J. (L 119) (algorithmic "processing [by platforms or other data holders] 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.").

[71] *Facebook's Response to the Oversight Board's First Decisions*, https://about.fb.com/wp-content/uploads/2021/02/OB_First-Decision_Detailed_.pdf [https://perma.cc/H3EE-GYHE] (last visited Dec. 13, 2021).

[72] *See Zuckerberg Hearing*, *supra* note 1 and accompanying text.

front-line decision of whether a given piece of content should go up or stay up. But for a user to both understand why their content was taken down and to have a meaningful right to challenge that decision if they choose to, there needs to be human-level explainability on the front line. Whether a human or AI does the explaining, the reason for the decision must be capable of being understood by a user. This points to the need for a resolution of the "how vs. why" problem discussed above. It also calls for a side constraint that forces the system to provide the user with notice of the automated action. Without a meaningful right to notice and an explanation at the initial review level, the digital prior restraint problem is unsolvable.[73]

Nor does notice and a right to explanation solve the overinclusiveness problem embedded in the process of machine learning. If a human sitting in review of an automated content review decision concludes that the decision is wrong, their role in overruling the system's decision has two next steps: (1) have the content put back up, and (2) churn the false positive back into the system to prevent the same mistake from happening, both with respect to the content in question and identical content elsewhere on the platform. What they cannot do is prevent the mistake from happening in the first instance.

To be fair, human-based moderation systems are not perfect either, and mistakes can and do occur under those systems as well. False positives are the inevitable by-product of any content moderation system, indeed any speech regulation system—even First Amendment doctrine—that has the capacity to evolve. But the same scale problem that calls for a prominent role for machine learning in content moderation also multiplies the number of errors the system makes.

CONCLUSION

On May 19, 2021, an article titled "Sharing Learnings About Our Image Cropping Algorithm" was posted to Twitter's Engineering Insights blog.[74] In the post, Rumman Chowdhury, the head of Twitter's Machine Learning, Ethics, and Algorithmic Transparency team, shared the results of an internal study concerning Twitter's image cropping algorithm, in particular how the algorithm chose points-of-focus in deciding how to crop pictures, and whether the algorithm's cropping decisions were biased depending on the gender and/or skin color of the people in the photos to be

---

[73] Of course, "[e]xplainability may mean different things in different contexts," and thus "different settings may require different types of explanations" depending on the user or other stakeholder to whom the explanation is owed. *See* Shenkman et al.*, supra* note 18, at 34 (citing P. Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence*, NAT'L INST. FOR STANDARDS & TECH. (Aug. 2020)).

[74] Rumman Chowdhury, *Sharing Learnings About Our Image Cropping Algorithm*, TWITTER INSIGHTS BLOG (May 19, 2021), https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.html [https://perma.cc/PHN3-PJTK].

cropped.[75] The algorithm was trained to crop photographs based on its prediction as to the part of a photo that "a human might want to see first," so as to improve consistency and fast viewability of photos in Twitter user timelines.[76] Reviewing randomly linked images depicting individuals of different races and genders, the study found that the algorithm cropped images in favor of white individuals over black individuals at a statistically significant rate.[77] The rate at which the algorithm favored white women over black women was even greater.[78]

In light of those results, the team imposed a side constraint on the algorithm.[79] Instead of automatic cropping of vertically oriented photographs, Twitter's picture-posting function would provide users a preview of the cropped image before it was posted.[80] The change, Chowdhury wrote, "reduces [Twitter's] dependency on [machine learning] for a function that [it] agree[s] is best performed by people using our products."[81]

Twitter's side constraint choice—greater user control over an otherwise fully automated decision—is not a one-size-fits-all approach to resolving a Type II moderation problem. Indeed, it may not even be a "moderation" problem that this particular side constraint is attempting to solve. But it does show that platforms are capable of making trade-offs between speed-and-scale-based solutions and resolving the problems that the use of those solutions necessarily entail. If social media platforms are as dedicated to the values of transparency, free speech, and user agency as they claim to be, then these trade-offs, and the human interventions in moderation decision that effectuate them, will be both welcome and inevitable in the age of speech-regulation-by-algorithm.

---

[75] Twitter undertook the study in response to user claims of bias in the cropping algorithm, including critiques published on Twitter itself. *See* Kyra Yee et al., *Image Cropping on Twitter: Fairness Metrics, Their Limitations, and the Importance of Representation, Design, and Agency*, ARXIV (May 18, 2021), https://arxiv.org/pdf/2105.08667.pdf [https://perma.cc/X4MD-U52R]; *see also* Anna Kramer, *Twitter's Image Cropping Was Biased, so it Dumped the Algorithm*, PROTOCOL (May 19, 2021), https://www.protocol.com/twitter-image-cropping-algorithm-biased [https://perma.cc/MYD6-A4AM].

[76] *See* Yee et al., *supra* note 75, at 2–3. Twitter crops user pictures to make their dimensions fit within the platform's standard aspect ratio. It also minimizes alternatives that degrade or distort the original image, such as shrinking the entire photo to fit the platform aspect ratio. *See id.* at 1.

[77] *See id.* at 7.

[78] *Id.* at 7–12.

[79] Chowdhury, *supra* note 74.

[80] *Id.*

[81] *Id.*