Framing Online Speech Governance as an Algorithmic Accountability Issue

MEHTAB KHAN*

Automated tools used in online speech governance are prone to errors on a large-scale yet widely used. Legal and policy responses have largely focused on case-by-case evaluations of these errors, instead of an examination of the development process of the tools. Moreover, information on the internet is no longer simply generated by users, but also by sophisticated language tools like ChatGPT, that are going to pose a challenge to speech governance. Yet, legal and policy measures have not responded adequately to AI tools becoming more dynamic and impactful. In order to address the challenges posed by algorithmic content governance, I argue that there is a need to frame a regulatory approach that focuses on the tools used in both content moderation and content generation contexts—which can be done by viewing this technology through an algorithmic accountability lens. I provide an overview of the various aspects of the technical and normative features of these tools that help us frame the regulation of these tools as an algorithmic accountability issue. I do this in three steps: First, I discuss the lack of sufficient attention towards AI tools in current regulatory approaches. Second, I highlight the shared features of both content moderation and content generation to offer insights about the interlinked and evolving landscape of online speech and AI Governance. Third, I situate this discussion of speech governance within a broader framework of algorithmic accountability to guide future regulatory interventions.

---

* Fellow, Berkman Klein Center for Internet & Society at Harvard University, Visiting Fellow, Yale Law School, Information Society Project. The author would like to thank Jack Balkin, Robert Post, Lyrissa Lidsky, Pauline Trouillard, the participants of the Yale Information Society Project Writing Workshop, and the participants of the Freedom of Expression Scholars Conference 2023 at Yale Law School for their helpful comments and feedback.

INTRODUCTION

During the Covid-19 pandemic, social media platforms increased their reliance on AI content moderation tools due to the increase in user traffic alongside the unavailability of human moderators.[1] Platforms were facing criticism for failing to remove misinformation about vaccines and anti-vax groups, and hence had to respond more swiftly amidst the loss of life and livelihood during the global pandemic. However, increased use of AI content moderation tools also meant that there was an increase in error rates.[2] AI tools were found to be ill-equipped to detect harmful content that was previously managed by human moderators.[3]

The Covid-19 pandemic highlighted issues that have always existed in the online speech ecosystem. AI tools have always been a heavily contentious feature of content governance on the internet.[4] There have been documented instances of where AI tools have inadvertently led to censorship of protests around the world.[5] Facebook also famously faced scrutiny for repeatedly taking down "The Terror of War" photograph (the Napalm girl), mislabeling it as explicit content.[6] In many of these contestations, there have recurring questions about the responsibilities of platforms that host and curate such content, and users who express themselves on these platforms.

More recently, AI tools have become the source of content generation and creation. Information on the internet is no longer simply created by human users, but also by sophisticated language tools like ChatGPT.[7] This has led to more confusion about who may be responsible for the development of these AI tools and for the governance of content generated through them. Language and image generating tools are likely going to pose unique challenges to content governance due to the scale and accessibility of the tools. In some ways, content moderation and generation tools will exacerbate their respective issues. This is because if content generation tools make it

---

1. Marc Faddoul, *Covid-19 is Triggering a Massive Experiment in Content Moderation*, BROOKINGS (Apr. 28, 2020), https://www.brookings.edu/techstream/covid-19-is-triggering-a-massive-experiment-in-algorithmic-content-moderation/ [https://perma.cc/TR2E-AZAR].

2. "Algorithmic moderation" refers to the use of content techniques to classify content and apply an outcome of content moderation. Tarleton Gillespie, *Content Moderation, AI, and the Question of Scale*, BIG DATA & SOC'Y (Aug. 21, 2020), https://journals.sagepub.com/doi/full/10.1177/2053951720943234.

3. *Id.*

4. CAREY SHENKMAN, DHANARAJ THAKUR & EMMA LLANSÓ, CTR. DEMOCRACY & TECH., DO YOU SEE WHAT I SEE? CAPABILITIES AND LIMITS OF AUTOMATED MULTIMEDIA CONTENT ANALYSIS (2021).

5. Tomiwa Ilori, *Facebook's Censorship of the #EndSARS Protests Shows the Price of its Content Moderation Errors*, SLATE (Oct. 27, 2020, 11:38 AM) https://slate.com/technology/2020/10/facebook-instagram-endsars-protests-nigeria.html [https://perma.cc/9AFW-SDUR].

6. Sarah T. Roberts, *Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation*, FIRST MONDAY (Mar. 3, 2018), https://firstmonday.org/ojs/index.php/fm/article/view/8283.

7. Teirnan Ray, *ChatGPT is Not Particularly Innovative*, ZDNET (Jan. 23, 2023, 5:05 AM) https://www.zdnet.com/article/chatgpt-is-not-particularly-innovative-and-nothing-revolutionary-says-metas-chief-ai-scientist/ [https://perma.cc/SW3W-5BMS].

easier to produce harmful information, automated content tools used to detect harmful content will become more pervasive—compromising careful decision-making over quick responsiveness.

Despite the shortcomings, AI tools are seen as a necessity.[8] Moderating the staggering scale of content on the internet is impossible without assistance. And platforms face liability if they do not moderate content.[9] Google has faced lawsuits in the last decade for including copyright-infringing content in its search results and image and video repositories.[10] Every day, major platforms like Facebook, Twitter, and YouTube receive thousands of requests to review or take down content that is violative of their internal policies or an external law. Sometimes they receive requests, both from the US government and foreign governments, for information on users, or to censor specific people and accounts.[11] Facebook reports that between April and June 2021, it took action on over 31.5 million pieces of "content" that were classified as hate speech, largely flagged using automated content detection technology.[12] But regulation is limited, which has led social media platforms to take the role of legislator, executive, and judiciary when mediating online speech—with the assistance of AI tools.

Scholars and policymakers have long been studying the shortcomings of AI tools, but the focus on solutions has largely been on a case-by-case evaluation of the failures of algorithmic content moderation. The examples mentioned above show not just the failures of AI tools and how little we know about how these tools are developed and deployed but also that the importance of the people and institutions involved in the collection, training, flagging, and processing of content that gets fed into the machine learning models used by platforms. Both the unintentional errors of automated tools as well as the intentional design choices have consequences for the kinds of information we engage with and learn from. In order to examine these aspects of AI content governance tools, scholars have suggested looking at content moderation as a system,[13] and the need to understand the ex-ante processes involved in its creation. The scale of moderation will always involve errors, so we need to decide the extent to which errors are acceptable and preferred.[14]

The issue of content governance is existential for platforms and closely tied to their business models, yet there is an absence of a legal framework to assess the automated tools. This gap in law and policy becomes more evident whenever we see the consequences of AI tools. For example, Facebook has repeatedly come under fire for the disinformation campaigns allowed on its platform during the 2016 U.S.

---

8. Gillespie, *supra* note 2.
9. 47 U.S.C. § 230.
10. *E.g.*, Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).
11. *Government Requests for User Data*, META, https://transparency.fb.com/data/government-data-requests/ [https://perma.cc/E6MT-PCK9].
12. *Community Standards Enforcement Report*, META, https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/#content-actioned [https://perma.cc/P677-DQUY].
13. Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526 (2022).
14. *Id.*

elections and for amplifying disinformation and hate speech, which some contend has led to a genocide in Myanmar.[15]

Yet there is little public access and scrutiny of the processes of construction and implementation. This article conceptualizes algorithmic speech governance issues as worthy of attention within broader discussions about the accountability of AI systems and to devise regulatory approaches targeted towards the technical system designing content moderation and content generation tools. I do this in three steps: First, I discuss the lack of sufficient attention towards AI tools in current regulatory approaches. Second, I highlight features of both content moderation and content generation to offer insights about the interlinked and evolving landscape of online speech and AI governance. Both content moderation and content generation tools have commonalities when it comes to their regulating them, and we can begin to identify those commonalities by giving more systematic attention to the processes involved in the creation and use of these tools. Third, I situate this discussion of speech governance within a broader framework of AI accountability to guide future regulatory interventions. This discussion is intended to provide a starting point for present discussions and policy questions about the kinds of transparency we need from platforms, in what contexts the AI tools ought to be used, and how to respond to features of the system that limit or promote user rights and freedom of expression.[16]

This discussion is timely given how social media platforms are now facing increased legal scrutiny. A case at the Supreme Court last term term touched upon the role of algorithms in shaping what we see on social media platforms. In the oral arguments in *Gonzalez v. Google*, the Supreme Court considered whether a platform can be liable for the algorithmic tools used to recommend harmful content.[17] This case raised novel questions about where we may situate algorithmic tools within existing legal regimes. For example, the petitioners in this case tried to distinguish between content moderation tools and recommendation algorithms, arguing that the latter are not protected under Section 230 of the CDA.[18]

---

15. Paul Mozur, *A Genocide Incited on Facebook, with Posts from Myanmar's Military*, N.Y. TIMES (Oct 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html [https://perma.cc/F7G7-U5Y7].

16. *See* Daphne Keller, *Amplification and its Discontents*, KNIGHT FIRST AMENDMENT INST. (June 08, 2021), https://knightcolumbia.org/content/amplification-and-its-discontents [https://perma.cc/K2ZL-3VD7] (discussion on why it is difficult to regulate the social media algorithms responsible for recommendation and amplification is difficult); Benjamin Laufer & Helen Nissenbaum, *Algorithmic Displacement of Social Trust*, KNIGHT FIRST AMENDMENT INST. (Nov. 29, 2021), https://knightcolumbia.org/content/algorithmic-displacement-of-social-trust [https://perma.cc/56V9-FLCU] (Nissenbaum argues that algorithmic amplification is a symptom of the actual problem, which is the loss of processes that allow us to determine trustworthy content).

17. Transcript of Oral Argument, Gonzalez v. Google, 598 U.S. 617 (2023) (No. 21–1333).

18. Although there was discussion about the role of algorithms on social media in the oral arguments for *Gonzalez v. Google*, 598 U.S. 617 (2023), the Supreme Court did not ultimately consider the question of whether Section 230 applies to recommendation algorithms. The Court remanded *Gonzalez* to the Ninth Circuit in light of its decision in *Twitter v. Taamneh*, 598 U.S. 471 (2023) to not hold social media companies responsible for aiding

This case raises salient points about the broader regulation of AI tools. First, there is increased legal scrutiny of these tools, but it is not clear when and how existing laws apply. Second, there is a lack of clarity about what tools are in use, for what purposes, what the tools do, where platform liability lies, and no clear mapping out of the various scenarios under which these tools may lead to liability. This is also an issue when it comes to content generation tools where we are seeing lawsuits coming up against platforms creating generative tools for copyright infringement, and questions about how publicly available information may be used to create AI tools in the first place. Third, there is a lack of consensus on what is actionable conduct by platforms—is the operationalization of the tool, the creation, the intention behind it—or something else that may be uncovered by looking at these tools as a system. Platforms have thus far evaded heavy regulation due to the First Amendment implications of such measures.[19] In this uncertain legal and policy landscape, it is worth focusing on the development of the underlying technologies and examining online speech tools as a system since they have thus far proven to be consequential to individual rights and freedoms on the internet.

I. PLACING AI TOOLS WITHIN EXISTING LEGAL REGIMES

This section discusses why we need systemic attention towards AI tools. It also uses the Content ID system developed by YouTube to illustrate why current legal frameworks do not adequately address the accountability of AI content governance tools, and how we may think through the technical aspects of a tool as a matter in need of regulatory attention.

A. Automated Systems used in Online Speech

The governance of the content moderation ecosystem we see today is shaped by a multitude of distinct systems, such as law, discourse, community standards, design, automation, and people.[20] With the addition of technical systems, the regulatory environment in which online speech is governed can now be referred to as "New School Speech Regulation." These regulatory responses are aimed at infrastructure

---

and abetting terrorism simply by showing and recommending terrorist content on their platforms.

19. *See generally*, Kate Klonick, *The New Governers: The People, Rules, And Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1609 (As Klonick notes, the debate over how to moderate online content and also protect user speech is ongoing, with no exact analogy to whether social media platforms are broadcasters, editors, or public squares); Daphne Keller, *One Law, Six Hurdles: Congress's First Attempt to Regulate Speech Amplification in Padaa*, CTR. FOR INTERNET & SOC'Y (Feb. 01, 2021, 5:00 AM), https://cyberlaw.stanford.edu/blog/2021/02/one-law-six-hurdles-congresss-first-attempt-regulate-speech-amplification-padaa [https://perma.cc/LKY3-WE7F] (Keller summarizes the constitutional issues with trying to regulate speech on the internet. Laws that restrict speech invite strict First Amendment scrutiny, and on the internet, laws that restrict distribution also risk the suppression of lawful speech.).

20. Ari Ezra Waldman, *Disorderly Content*, 97 WASH. L. REV. 907, 916 (2022).

that includes social media platforms and search engines, as opposed to "Old School Speech Regulation," which targeted speakers and publishers of content.[21] Rules and exceptions now target the owners of digital infrastructures, which are platforms.[22] This is because platforms make decisions and tradeoffs about what kinds of speech stays online and accessible and what does not. And how that power is exercised through the development of automated tools needs more interrogation because we do not know enough about the tradeoffs and decisions and the stakeholders that influence them.

Automated content-moderation tools are pervasive because they attend to the problem of scale in content moderation.[23] It is not humanly possible to filter and moderate millions of posts, videos, and media content on the internet every day. These tools are particularly appealing to use in filtering egregious forms of harmful speech such as terrorist content and hate speech.[24] AI tools are instrumental in high stakes situations, such as during elections. Recently, Meta stated that its detection technology would be used to detect and remove hate speech during state elections in India.[25] These tools are continuously being updated and increasingly used, as Meta's press release indicated.[26]

Researchers have pointed out that most automated content detection uses "a mixture of natural language processing, image processing, and social network analysis."[27] Facebook has been under fire for its inability to manage disinformation in Myanmar.[28] Mark Zuckerberg acknowledged that one of the shortcomings in Facebook's response was the company's lack of linguistic capability to moderate content in the country.[29] Facebook was also accused of overstating the capability of its automated tools when internal documents revealed that **"**more than 95 percent of

---

21. Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV., 1149, 1174 (2018).

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

23. Gillespie, *supra* note 2.

24. ROBYN KAPLAN, DATA & SOCIETY, CONTENT OR CONTEXT MODERATION? ARTISANAL, COMMUNITY-RELIANT, AND INDUSTRIAL APPROACHES (2018).

25. *How Meta is Prepared to Protect the Upcoming State Elections in India*, META (Feb. 10, 2022), https://about.fb.com/news/2022/02/how-meta-is-prepared-to-protect-the-upcoming-state-elections-in-india/ [https://perma.cc/9A8U-MSWQ].

26. Mike Schroepfer, *Update on Our Progress on AI and Hate Speech Detection*, META (Feb. 11, 2021), https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/ [https://perma.cc/YW8M-AZCN].

27. Devin Soni, *Machine Learning for Content Moderation—Introduction*, TOWARDS DATA SCI. (July 22, 2019), https://towardsdatascience.com/machine-learning-for-content-moderation-introduction-4e9353c47ae5 [https://perma.cc/7BL5-W2QY].

28. Anthony Kuhn, *Activists in Myanmar Say Facebook Needs to do More to Quell Hate Speech*, NPR (June 14, 2018, 1:34 PM), https://www.npr.org/2018/06/14/619488792/activists-in-myanmar-say-facebook-needs-to-do-more-to-quell-hate-speech [https://perma.cc/8UQZ-KVSS].

29. Steve Stecklow, *Why Facebook is Losing the War on Hate Speech in Myanmar*, REUTERS (Aug. 15, 2018, 3:00 PM), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

hate speech shared on Facebook stays on Facebook."[30] Furthermore, an experiment on a test user in India found that the user's newsfeed became "a near constant barrage of polarizing nationalist content, misinformation, and violence and gore."[31]

Part of the issue is in how the tools are designed. Machine learning systems are not value-neutral, and so there are distinct entities and processes that shape speech rights on the internet. Managing the scale of content moderation invariably sacrifices context and localized attention to the needs on the ground.[32] For example, AI tools fall short of the needs on the ground because they attempt to water down "complex concepts like harassment and hate speech" in the interest of efficiency at scale.[33] A feature as basic as identifying duplicates on a platform may perform badly when it fails to identify content used in a different context such as when "terrorist propaganda" is reposted in a journalistic context.[34] Yet, scholars have argued that content governance approaches have been slow to respond to these systemic issues, and the focus remains on individual failures.[35]

A systemic approach is necessary because the effective governance of online content is an ongoing balance between three parties: platforms, governments, and users who all have their respective interests.[36] AI tools are now essential to meeting public and regulatory expectations of content governance.[37] Ultimately, upholding principles like freedom of expression will depend on the design of the technical infrastructure that allows democratic participation in the first place.[38] Since technology itself also exemplifies relationships of power between one set of human beings and another,[39] this means automated content moderation and generation tools also control an individual's ability to participate and express themselves. It has costs for their reputations and safety in online environments. Technology that is built by platforms collecting, processing, and using user generated content has consequences for what people get to access, the risks they are exposed to online, and categories they are placed in. And the law needs to respond to this technological change.

---

30. Noah Giansiracusa, *Facebook Uses Deceptive Math to Hide its Hate Speech Problem*, WIRED (Oct. 15, 2021, 7:00 AM), https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/ [https://perma.cc/Y2YM-LVA9].

31. Sheera Frenkel & Davey Alba, *In India, Facebook Grapples with an Amplified Version of its Problems*, N.Y. TIMES (Oct. 23, 2021), https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html [https://perma.cc/2CXZ-5U3H].

32. KAPLAN, *supra* note 24, at 25.

33. TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 29 (2018).

34. Gillespie, *supra* note 2, at 3 (citing Emma Llansó, *Platforms Want Centralized Censorship. That Should Scare You.*, WIRED, (Apr. 18, 2019, 9:00 AM), https://www.wired.com/story/platforms-centralized-censorship/ [https://perma.cc/4QBQ-K9MK].

35. Douek, *supra* note 13.

36. Jack Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011 (2018)

37. Douek, *supra* note 13.

38. Balkin, *supra* note 36.

39. Balkin, *supra* note 21, at 1158.

B. Content Generation

Recently, researchers have cautioned that large language models can generate falsehoods fluently and efficiently. This is a risk of online speech and how harmful content may continue to be a feature of social media platforms. Disinformation is a particularly urgent concern because automated language makes it easier to generate vast amounts of content.[40] This is because it would cost less to hire someone to edit the content generated by the model than to have someone write content from scratch. The need to map the wide-ranging harms and risks associated with content generating tools is especially urgent.[41] We are now seeing new examples of how ChatGPT and other tools are raising legal and ethical issues which need to be understood using a cohesive analytical framework.[42]

Deciding on intervention strategies is an ongoing challenge and will depend on the impact of which we are concerned. When deciding on an appropriate response, policymakers and platforms will have to grapple with weighing what constitutes a more urgent concern. For example, when it comes to the generative risks, we need to identify who the potential bad actors are. Not all uses of generative tools will be harmful and could be beneficial for research and educational purposes.[43]

We currently lack consensus on what these harms are. Some uses of content generation tools are clearly undesirable, but the degree of harm may differ. For example, content generation tools could be used to generate spam, or there could be state-sponsored actors using them to generate propaganda campaigns at a large scale.[44] Each of these instances will require a different response.[45]

Another emerging trend in content generation is that the content itself may not be harmful or offensive, but may trigger incidental liability, such as copyright

---

40. RISHI BOMMASANI, ET AL., CTR. FOR RES. ON FOUND. MODELS, ON THE OPPORTUNITIES AND RISKS OF FOUNDATIONAL MODELS 136 (2021).

41. Khari Johnson, *Chatbots Got Big—and Their Ethical Red Flags Got Bigger*, WIRED (Feb. 16, 2023, 7:00 AM), https://www.wired.com/story/chatbots-got-big-and-their-ethical-red-flags-got-bigger/ [https://perma.cc/KM6T-HCJ9].

42. *See* Andrew M. Perlman, *The Implications of ChatGPT for Legal Services and Society* (Suffolk U. L. Sch., Research Paper No. 22-14, 2023) (some ethical issues include ensuring that the information produced by the AI tool is accurate, including citations. Another issue is deciding when a lawyer should disclose that an automated tool was used to assist in preparing documents or give legal advice); *see also* Marco Marcelline, *Cybercriminals Using ChatGPT to Build Hacking Tools, Write Code*, PC MAG. (Jan. 08, 2023), https://www.pcmag.com/news/cybercriminals-using-chatgpt-to-build-hacking-tools-write-code [https://perma.cc/W4Y6-CQZX]; Mike Pearl, *The ChatGPT Chatbot from OpenAI is Amazing, Creative, and Totally Wrong*, MASHABLE (Dec. 03, 2022), https://mashable.com/article/chatgpt-amazing-wrong [https://perma.cc/93VX-YE5K].

43. *GitHub CoPilot*, GITHUB, https://github.com/features/copilot.

44. JOSH A. GOLDSTEIN, GIRISH SASTRY, MICAH MUSSER, RENÉE DIRESTA, MATTHEW GENTZEL, & KATERINA SEDOVA, GEORGETOWN UNIV. CTR. SEC. & EMERGING TECH., GENERATIVE LANGUAGE MODELS AND AUTOMATED INFLUENCE OPERATIONS: EMERGING THREATS AND POTENTIAL MITIGATIONS (2023).

45. ALEX TAMKIN, MILES BRUNDAGE, JACK CLARK & DEEP GANGULI, UNDERSTANDING THE CAPABILITIES, LIMITATIONS, AND SOCIETAL IMPACT OF LARGE LANGUAGE MODELS (2021).

infringement. For example, GitHub has recently been sued for copyright infringement for using code in its Copilot code generation tool.[46] DeviantArt has come under fire for scraping artists' images without permission to train its AI art generation tool.[47] Getty Images has sued an AI art generator for using its images without permission.[48] Individual case determinations will focus on copyright questions, but they will also inevitably involve considerations of how these content generation systems are designed, where the training data comes from, what data scraping practices are widespread, and the law and ethics of how content generation tools are used.

When it comes to understanding how language technologies work, compelling disclosure about the process may be a double-edged sword because authoritarian regimes may attempt to use disclosure of data as an attempt to control the use of the data, and implementation of applications.[49] It may also mean that they could undermine security features that some of these applications employ, such as end-to-end encryption on WhatsApp.[50] How do we prevent this from accidentally supporting authoritarianism and mass disinformation or harassment campaigns? That should be a question when it comes to deciding what to do about language technologies and developing broader norms and rules around the use and deployment of this technology. And these issues need to be addressed at a systemic level.

C. Current Legal Frameworks: YouTube, Content ID and the DMCA

Platforms need legal protections and immunities to not be held responsible for user-generated content, and to be able to devise their own content moderation policies. Laws like the Communications Decency Act of 1996 (CDA) and the Digital Millennium Copyright Act of 1998 (DMCA) provide the foundations for how platforms approach content moderation. These laws provide intermediaries crucial safe harbors against liability.

---

46. Emma Roth, *Microsoft, GitHub, and OpenAI Ask Court to Throw Out AI Copyright Lawsuit*, THE VERGE (Jan 28, 2023, 7:02 PM), https://www.theverge.com/2023/1/28/23575919/microsoft-openai-github-dismiss-copilot-ai-copyright-lawsuit [https://perma.cc/GZ55-B7KN].

47. Benj Edwards, DeviantArt Upsets Artists with its New AI Art Generator, Ars Technica (Nov. 11, 2022, 5:47 PM), https://arstechnica.com/information-technology/2022/11/deviantart-upsets-artists-with-its-new-ai-art-generator-dreamup/ [https://perma.cc/8PE5-BSJU].

48. James Vincent, *Getty Images Sues AI Art Generator Stable Diffusion in the US for Copyright Infringement*, THE VERGE (Feb. 6, 2023, 11:56 AM), https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion [https://perma.cc/TPJ8-6P4U].

49. Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L. J. 1353, 1384 (2018); Jon Porter, *WhatsApp Sues Indian Government Over New Rules*, THE VERGE (May 26, 2021, 5:41 AM), https://www.theverge.com/2021/5/26/22454381/whatsapp-indian-government-traceability-lawsuit-break-encryption-privacy [https://perma.cc/L9PJ-8BRF].

50. About End-to-end Encrypted Backup, WhatsApp, https://faq.whatsapp.com/490592613091019.

Section 230 of CDA is considered to be a cornerstone of internet legislation in the United States and credited as the reason so many platforms have been able to grow and flourish here.[51] This provision immunizes the "providers and users of an interactive computer service" who publish content posted by third parties.[52] Despite this immunity, there is always a threat of liability if platforms do not comply with the conditions in this law.[53] Because of this inherent vested interest, platforms are not neutral parties in the process of content moderation, nor can the technologies they develop for the purpose of content governance be considered objective and equally applicable to all kinds of users. Section 230 has also allowed platforms discretion for how they moderate content, including the design of AI tools. It is important to note that AI tools need not be neutral, but an examination of how they are created is necessary when grappling with their widespread and consequential impacts.

Section 512(d) of the DMCA gives platforms immunity from monetary damages for referring to or linking to a location that contains material that infringes someone's copyright.[54] This immunity is crucial for platforms because platforms cannot monitor and take down billions of web pages. They also cannot be held liable for every instance of problematic content as that would lead them to expend all their resources in avoiding liability instead of providing a service. Since the internet is vast, and constantly changing and updating, there is no realistic way that platforms can monitor every action and post to ensure there is nothing illegal on their platforms.[55] Therefore, automated tools are essential in helping platforms comply with the law.

In order to retain this immunity, platforms have to operate the 'notice and takedown' system as described above. This obligation is operationalized with the help of AI tools such as the Content ID system on YouTube, which essentially a private system of copyright regulation to maintain immunity against legal claims for what third parties post on their platforms.[56]

Once they receive a notice about content that is allegedly infringing someone's copyright, platforms are mandated to respond to the notice by either taking content down or if they decide it is not infringing, then letting the sender know of their decision.[57] Platforms are in a position of power as they are required by law to exercise unprecedented authority over potential wrongdoing. In operating 'notice and takedown' systems, platforms make several important determinations. Some are technical like checking if the notice complies with the requirements of Section 512.[58] Others are more discretionary, such as deciding whether the infringing content is violating someone's copyright or may fall within an exception like 'fair use'. They

---

51. *See generally* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

52. Communications Decency Act, 47 U.S.C. § 230.

53. Mark A. Lemley, *Rationalizing Internet Safe Harbors*, 6 J. TELECOMMS. & HIGH TECH. L. 101, 102 (2007).

54. Digital Millennium Copyright Act, 17 U.S.C. § 512.

55. *See* Gillespie, *supra* note 2, at 1.

56. Jennifer M. Urban, Joe Karaganis & Brianna L. Schofield, *Notice and Takedown in Everyday Practice 10* (UC Berkeley, Research Paper No. 2755628, 2017).

57. 17 U.S.C. §§ 512(c)(2),(d)(2); URBAN ET AL., *supra* note 56, at 16.

58. URBAN ET AL., *supra* note 56, at 29.

also have to decide how to respond to a notice and whether to take down content as soon as they receive a notice or not.[59] The way these determinations are made are not specifically mandated in the law which means a lot of private—and in the case of Content ID, automated—determinations ultimately affect the content that stays online and content that is removed.

These decisions have implications for what kind of disputes are brought to platforms and what kind are brought to court, who brings them, what is taken down, and the overall efficacy and impact on access to copyrighted material online. Furthermore, since the law places the responsibility for maintaining a 'notice and takedown' system and its enforcement on platforms, they end up functioning as adjudicators and enforcers of certain decrees which may or may not be subject to judicial scrutiny.

Scholars have suggested that the safe harbor provisions in the DMCA are "confusing and illogical."[60] The "notice and takedown" system encourages platforms to comply with a complaint if it is made, no matter how frivolous it may be. Empirical studies of DMCA notices have found that "30% of them were legally dubious . . . ."[61] The law allows for counter-notices from people whose content is removed, but most people do not pursue that option.[62] One study of DMCA notices reviewed 876 notices received by various platforms and individuals.[63] It found that the primary users of these notices were corporations and business entities.[64] The overall effect of the system is that it provides more incentive for platforms to takedown doubtful content as soon as they receive a notice.[65]

When platforms rely on AI tools to institute and enforce a "notice and takedown" system, the laws especially seem to have significantly limited scope. Content ID is a prime example.[66] AI tools are not simply making decisions mandated by law, but are enacting actions not required by law, thus making them more overbroad than what the law may require.[67] The Content ID system falls outside the DMCA requirements, as until some time ago, copyright owners participating in this system did not even have to submit a formal DMCA notice.[68]

---

59. *Id.* at 29–30.

60. Lemley, *supra* note 53, at 102.

61. *Id.* at 114 (citing Jennifer M. Urban & Laura Quilter, *Efficient Process or "Chilling Effects"? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA HIGH TECH. L. J. 621, 667 (2006)).

62. Urban & Quilter, *supra* note 61, at 623.

63. *Id.* at 641. (citing Chilling Effects, now at https://lumendatabase.org/).

64. Daniel Seng, *The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices*, 18 VA. J. L. & TECH. 369, 373–374 (2014).

65. *Id*. at 376.

66. *How Content ID Works - YouTube Help*, GOOGLE, support.google.com/youtube/answer/2797370?hl=en [https://perma.cc/AY28-ZF8Y].

67. Jennifer Urban et al. describe these measures in the context of the DMCA as "DMCA Plus" practices that go a step beyond responding to notice and takedown requirements and instituting automated content filters like the Content ID system. URBAN ET AL., *supra* note 56, at 29.

68. Parker Higgins, *YouTube Upgrades Its Automated Copyright Enforcement System*, ELEC. FRONTIER FOUND. (Oct. 5, 2012), www.eff.org/deeplinks/2012/10/youtube-upgrades-its-automated-copyright-enforcement-system [https://perma.cc/7PGC-EUNC].

Content ID works by scanning and detecting copyrighted materials in all videos uploaded to YouTube.[69] It may seem like platforms can carry out this task in a neutral manner since the filter applies to all uploads, but by its nature this is not a neutral endeavor. Large-scale content creators, for example, have more access to the YouTube Content ID system than small independent creators.[70] Scholars can critique this system and track its effectiveness in light of an already established legal framework, namely copyright law.[71] But as I discuss below, there is no corresponding legal recourse to judge the decisions made by AI tools, and these decisions are sometimes submitted for review by private bodies such as the Meta Oversight Board.[72] Furthermore, platforms have an impact outside their jurisdictional boundaries as the internet lacks territorial limitations.[73] Therefore, there could be disparate effects of AI tools in different national contexts depending on how content governance laws and AI accountability frameworks are instituted.

Within the operationalization of the Content ID system, there is another dimension which involves privileged groups of people that have a say in how the system is used. Platforms like Google receive thousands of notices every month. Because of this volume of requests, Google instituted a Trusted Copyright Removal Program (TCRP) which allowed "agents," individuals who are authorized to send a DMCA notice, to participate as trusted submitters of copyright claims because they do so with "high accuracy."[74] Google does not "delay the processing" of these submitters compared to "non sophisticated submitters" who submit "incomplete or abusive" notices.[75] However, the mere participation of an agent in this program does not mean that all takedowns asserted by them are legitimate.[76] Furthermore, the fact that such a mechanism was instituted shows how ineffective individualized solutions like notice and counter-notices are against a large-scale system of content management.

The TRCP and YouTube's Content ID system are examples of how platforms build on and respond to requirements in laws like the DMCA to converge business interests with the management of user content. With the help of AI tools, platforms create rules and norms that come to be taken for granted as the rational and proper way to deal with online copyright disputes. Since the Content ID system is

---

69. GOOGLE, *supra* note 66.

70. URBAN ET AL., *supra* note 56, at 139

71. *See, e.g.*, *DMCA Notices*, LUMEN, https://www.lumendatabase.org/topics/29 [perma.cc/4RW9-PZX6].

72. Adi Robertson, *Facebook Oversight Board Overturns Hate Speech and Pandemic Misinformation Takedowns*, THE VERGE (Jan. 28, 2021, 11:31 AM), https://www.theverge.com/2021/1/28/22254155/facebook-oversight-board-first-rulings-coronavirus-misinformation-hate-speech [https://perma.cc/GJ5U-GTL4].

73. *Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations*, AMNESTY INT'L (Sep. 29, 2022), https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ [https://perma.cc/KH6H-MPPQ].

74. Seng, *supra* note 64, at 414.

75. *Id*.

76. *Id*. at 430.

automated, there is no legal assessment of whether something might be fair use. Yet the attention towards AI tools is incidental and not central to any policy responses. This system has clear power asymmetries, and this asymmetry is bound to be exacerbated in the context of the AI tools used for content prediction, detection, and creation.

Platforms now have similar technologies that filter content for various purposes. For example, Google has a tool called "Perspective API" that is used to detect hate speech.[77] The underlying technology is developed using language models that are trained and labeled by humans to identify categories of harmful speech.[78] The problems of context and accuracy with such systems are more complicated than simply checking for potential copyright infringement. The notice-and-takedown system and the AI tool developed in response is an example of the way in which automated tools are viewed as an incidental feature of speech governance but not given specific regulatory attention. The following sections analyze how we may situate automated tools within discussions of legal and policy interventions.

II. CONTENT MODERATION AND CONTENT GENERATION AS PART OF THE SAME SYSTEM

The discussion above shows that insufficient attention is paid to AI tools in current legal frameworks regulating platforms. It also illustrates what we overlook when AI tools are seen as ancillary or incidental to content governance. Furthermore, it highlights the complexity of the challenges now that AI tools are used in both content moderation and content generation. This section synthesizes these challenges and identifies some of the shared aspects of content moderation and content generation that need regulatory attention.

A. Focusing on the Underlying Technology

First, I take a closer look at language technologies, which have well-documented technical shortcomings. When they are used in the already complex online speech ecosystem, language technologies have the potential to exacerbate the harms resulting from ineffective content moderation. The unaccountable use of language technologies is a challenge shared by both content moderation and content generation. Language technologies pose a number of risks ranging from factual inaccuracies to the ability to generate high quality text for mass disinformation campaigns. Language technologies that encode a range of biases including sexist, racist, and stereotypical associations are used to develop automated content governance tools.[79] They are used in the generation of disinformation and

---

77. Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 BIG DATA & SOC'Y 7 (2020).

78. *Id.*

79. Emily Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, FACCT '21, 610, 617 (2021).

misinformation, and in content moderation, including the detection of hate speech, abuse, and other prohibited forms of speech on social media platforms.[80] The inaccuracies and issues laden within the language models will result in automated detection tools that do not work well. Paradoxically, even when they work as designed, language models have become more efficient at generating content, including harmful content.[81]

We may categorize features of language technologies that are relevant to both content moderation and content generation in three ways: the source of data used to build the technology, the scale of the model used to build the AI tool, and the capability of the tool that has been designed and deployed.

B. Data Sources

The source of the data used to build language technologies matters greatly because when dataset creators use large amounts of web text to see it as representative of all of humanity, they risk perpetuating "dominant viewpoints, increasing power imbalances, and further reifying inequality."[82] The choice of where training data is sourced from has a clear impact on the outcomes of an automated decision-making system in both content moderation and generation contexts. For example, stereotypes against a religious community ingrained in a selection of media sources will show up in a model trained on these sources.[83] Researchers have documented that choices such as the base language of a training model will shape the resulting models and advantage the speakers of that language.[84] Furthermore, biases expressed in the form of word connotations and context can be embedded in the models.[85] The discriminatory effects may not be intentional, but choice of base language and training data are deliberate, and hence there is more potential to direct policy efforts towards them. However, not enough regulatory attention has been given to how platforms make the choices and tradeoffs for language and training data choices.

---

80. Ioannis Mollas, Zoe Chrysopolou, Stamatis Karlos & Grigorios Tsoumakas, *ETHOS: an Online Hate Speech Detection Dataset*, 8 COMPLEX & INTELLIGENT SYS. 4663 (2022), https://doi.org/10.1007/s40747-021-00608-2; Mladen Karan & Jan Snajder, *Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context*, *in* PROCEEDINGS OF THE THIRD WORKSHOP ON ABUSIVE LANGUAGE ONLINE 129, 132–33 (Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran & Zeerak Waseem, eds., 2019).

81. Tiffany Hsu & Stuart A. Thompson, *Disinformation Researchers Raise Alarms about A.I. Chatbots*, N.Y. TIMES (Jun. 20, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html [https://perma.cc/L9J3-DYW4].

82. Bender et al., *supra* note 79, at 614.

83. IRENE SOLAIMAN, MILES BRUNDAGE, JACK CLARK, AMANDA ASKELL, ARIEL HERBERT-VOSS, JEFF WU, ALEC RADFORD & JASMINE WANG, OPENAI, RELEASE STRATEGIES AND THE SOCIAL IMPACTS OF LANGUAGE MODELS 11 (Aug. 2019) (biased outputs can be useful to detect further biases in training data, but fine tuning remains a challenge).

84. GABRIEL NICHOLAS & ALIYA BHATIA, LOST IN TRANSLATION: LARGE LANGUAGE MODELS IN NON-ENGLISH CONTENT ANALYSIS 23 (2023).

85. *Id.*

Even if there is diversity in language choice, comprehensiveness does not have the connotations one would think of when it comes to datasets. Researchers have thus far been unable to create datasets that are comprehensive enough to account for the "fluidity and variances in human language and expression."[86] As a result, automated tools cannot be used in "different cultures and contexts, as they are unable to effectively account for the various political, cultural, economic, social, and power dynamics that shape how individuals express themselves and engage with one another."[87] The issue then is with the practices, choices, and values that go into determining what the dataset comprises and not just the data.

There is a concentration of automated tools that are heavily based on the English language, and yet deployed in contexts of global significance such as social media content moderation—this is of particular concern as it was recently revealed by the Facebook whistleblower that a large amount of disinformation and hate speech are left undetected by Facebook due to its inability to moderate non-English content globally.[88] The choice of language is significant because of its social and political implications.[89] Beyond the inability of these tools to parse through non-English text, the utility of the tools should depend on the representativeness of the people impacted. Furthermore, there is little to no notice when an automated tool makes a content decision, especially in global contexts. The tools designed and used by platforms are supposed to respond to the scale and complexity of the content moderation challenges that platforms face. Yet we see repeated failures and systemic unknowns with how the tools are developed.

C. Dataset Size

Unfortunately, the size of text sources and data collected does not guarantee diversity.[90] Instead, the source of the data matters because it may be rife with problematic content despite being large in volume. For instance, Amanda Levendowski distinguishes between different kinds of data used to build datasets and argues that public domain datasets are "low-friction" because they allow AI researchers without access to large troves of data to perform machine learning inference easily, or to train their models on existing, public domain datasets.[91] Public-

---

86. SPANDANA SINGH, NEW AMERICA, EVERYTHING IN MODERATION: AN ANALYSIS OF HOW INTERNET PLATFORMS ARE USING ARTIFICIAL INTELLIGENCE TO MODERATE USER GENERATED-CONTENT 19 (2019).

87. *Id.*

88. Deepa Seetharaman, Jeff Horwitz & Justin Scheck, *Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts.*, WALL STREET J. (Oct. 17, 2021, 9:17 AM), https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184.

89. Daniel Rivkin, *New Language-learning Algorithms Risk Reinforcing Inequalities, Social Fragmentation, per U-M Study*, U. MICHIGAN (Apr. 27, 2022), https://news.umich.edu/new-language-learning-algorithms-risk-reinforcing-inequalities-social-fragmentation-per-u-m-study/ [https://perma.cc/3AAY-BHZ4].

90. Bender et al., *supra* note 79, at 623.

91. Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 589 (2018).

domain data may contain demonstrable biases, especially given that many works enter into the public domain after the expiration of their 96-year copyrights. Established biases have also been shown to exist in downstream visual and linguistic representations.[92] Additionally, size poses a problem to effective documentation of what is in the dataset. Bender and her coauthors refer to this as the "incurring documentation debt," a situation in which datasets are both undocumented and too large to effectively document.[93]

The size issue is also relevant when we encounter the danger of models that are too large and hence the social views embedded in them are presumed to be static and there is no way of currently knowing how often these models are updated to reflect societal changes.[94] It is not always easy to identify and detect harmful speech in context, as the large datasets miss more subtle forms of harmful content, such as gender bias, microaggression, dehumanization, and other more contextual forms of speech.[95] Furthermore, researchers have shown that the margin of error in a dataset often places the burden on "underserved, disenfranchised, and minority groups."[96] Moreover, the idea of "good" data is also contentious. If the language model has been trained on credible media or literary sources, but these sources consistently espouse a certain view about a religion or ethnicity, this may still be "good" data because it is representative of larger discourses. For instance, mainstream news media and entertainment in the West have depicted bias towards Arabs and Muslims.[97] And relying on a large volume of this "good" data may still encode biased representations about a marginalized community.[98]

D. Technological Capacity

The development process of datasets includes balancing considerations like replicability and generalizability.[99] If a model is customized and usable for one

---

92. *See generally id.*; Joy Buolamwini & Timnit Gebru, *Gender Shades*: *Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACHINE LEARNING RES. 77 (2018).

93. Bender et al., *supra* note 79, at 613.

94. *Id.* at 614.

95. *Id.* at 617.

96. GILLESPIE, *supra* note 2, at 3.

97. *See generally* LAURENS DE ROOIJ, MUSLIMS, MINORITIES, AND THE MEDIA: DISCOURSES ON ISLAM IN THE WEST (2023); *see also* Keon West & Joda Lloyd, *The Role of Labeling and Bias in the Portrayals of Acts of "Terrorism": Media Representations of Muslims vs. Non-Muslims*, 37 J. MUSLIM MINORITY AFFS. 211 (2017); Esra Özcan, *Lingerie, Bikinis and the Headscarf: Visual Depictions of Muslim Female Migrants in German News Media*, 13 FEMINIST MEDIA STUD. 427 (2013).

98. *See, e.g.*, Jack G. Shaheen, *Reel Bad Arabs: How Hollywood Vilifies a People*, 588 ANNALS AM. ACAD. POL. & SOC. SCI. 171 (2003); Alexander Hotz, *Newsweek 'Muslim Rage' Cover Invokes a Rage of its Own*, GUARDIAN (Sept. 17, 2012), https://www.theguardian.com/media/us-news-blog/2012/sep/17/muslim-rage-newsweek-magazine-twitter [https://perma.cc/WF3Q-GPUX].

99. *See generally*, Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton & Alex Hanna, *Data and Its Discontents: A Survey of Dataset Development*

context, what kinds of resources would be required to do the same for every other context? In this situation, comprehensiveness would take a different meaning. And developing a comprehensive database would not necessarily mean including any and all categories, but specific ones that would work well when operationalized. Therefore, it is a challenge to develop a tool that is capable of reliable application "across different groups, regions, and sub-types of speech."[100] Some of the capabilities of these AI tools relate directly to design choices. For example, Google's Perspective AI failed to identify gendered harassment online because it failed to understand that "language itself can code meanings that are intended only for specific audiences, and thus evade algorithmic recognition."[101] Tools are also unable to differentiate between the different uses of language and behavior. For instance, "excessively liking someone's pictures or using certain slang words may be construed as harassment on one platform or in one region of the world," but this will not be detected as so by an automated tool.[102] This also calls into question the defensibility of scale as a justification to use automated content moderation tools, especially when the underlying technology is unable to manage the scale of the content that is being moderated.[103]

Framing a problem has an impact on technological capacity. Consider responses to gendered online harassment. Scholars argue that digital security is often framed in terms of securing the device, and not so much in understanding the social and intimate nature of violence, both online and offline.[104] This in turn affects all of the resulting data and design choices. Furthermore, its "manifestations and mutations" are difficult to keep up with, and therefore, it would be important to continuously update the language tools tasked with detecting and filtering online harassment. Even with better tools, these interventions fall short because, as researchers argue, "the socio-technical aspects of how violence happens are not fully addressed by re-design alone."[105]

It is presently difficult to draw conclusions about content-generating tools because the extent of generative risks of language models is currently unknown. Consider the large language model GPT-3. This is a large model capable of "text summarization, chatbots, search, and code generation."[106] There is uncertainty about

---

and Use in Machine Learning Research, 2 PATTERNS, Nov. 2021; Victoria Stodden & Sheila Miguez, *Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research*, 2 J. OPEN RES. SOFTWARE e21 (2014); Morgan Klaus Scheuerman, Alex Hanna & Emily Denton, *Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development*, 5 PROC. ACM HUMAN-COMPUT. INTERACTION 1 (2021).

100. SINGH, *supra* note 86, at 17.

101. Maya Indira Ganesh & Emanuel Moss, *Resistance and Refusal to Algorithmic Harms: Varieties of 'Knowledge Projects'*, 183 MEDIA INT'L AUSTL. 90, 98 (2022).

102. NEW AMERICA, *supra* note 100.

103. Gillespie, *supra* note 2.

104. Ganesh & Moss, *supra* note 101, at 11.

105. *Id.* at 9.

106. Alex Tamkin & Deep Ganguli, *How Large Language Models Will Transform Science, Society, and AI*, STAN. UNIV. (Feb. 5, 2021), https://hai.stanford.edu/news/how-

whether a particular use is harmful now or may be harmful in the future. The capabilities of a model also make it difficult to forecast the impact on society.[107] The opacity and lack of explainability, combined with the uncertainty about the impact, make content generation an AI accountability issue that needs deeper examination.

E. Intersection of Technical and Normative Features

Here, I highlight the importance of recognizing power asymmetries when we ask for more transparency about dataset development. When asked to share data, companies cite trade secret protection over these tools, and others have pointed out the risk with releasing such information that may be used for malicious purposes.[108] In the context of AI tools, researchers have recommended that Facebook should release "error rates from automated decisions" including releasing "the false positive, true positive, and false negative rates, as well as precision and recall" to better understand how automated tools work.[109] This is a pervasive challenge for lawmakers and researchers as it is quite difficult to do so because, as one research group found, Facebook did not share a list of classifiers for speech because "it would have been impractical for the group to meaningfully review them, since there are a large number of classifiers which are constantly changing."[110] This means that even if platforms release some information, it is difficult to parse through and understand without context and awareness about how the data fits into the overall system. Not only do we need a system that mandates information-sharing arrangements with regulatory agencies and researchers, but we also need standards to foster specificity and knowing what data to ask for.

However, the challenges with transparency are exacerbated with power asymmetries in how and where it is used. One such example is the lack of transparency around how content databases are shared amongst large platforms.[111] The demands for accountability should not be limited to sharing data about the technical system but also the institutions involved in creating the AI tools in the first place. Consider the creation and sharing of the Global Internet Forum to Counter Terrorism (GIFCT) database. This was created by industry actors in response to pressure from governments to do more to filter terrorist content. Tech companies, including Microsoft, Google, and Meta, created a set of norms around terrorist content by sharing a hash database of the content removed by their respective platforms so that the other platforms could use that information to build their own tools.

---

large-language-models-will-transform-science-society-and-ai [https://perma.cc/NET9-CMRJ].

107. *Id.*

108. NEW AMERICA, *supra* note 100.

109. Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler & Danieli Evans Peterman, *Report of the Facebook Data Transparency Advisory Group*, JUST. COLLABORATORY AT YALE L. SCH. 8–9 (2019).

110. *Id.* at 15.

111. Chloe Hadavas, *The Future of Free Speech Online May Depend on This Database*, SLATE (Aug. 13, 2020), https://slate.com/technology/2020/08/gifct-content-moderation-free-speech-online.html [https://perma.cc/5Y64-BHK7].

Critics have decried the opacity surrounding the use procedures of GIFCT. For example, Syrian human rights abuse records have been mistakenly flagged as terrorist speech because the tools do not differentiate between content shared by human rights activists or those involved in the atrocities depicted in the content.[112] This example led to many unanswered questions about what constitutes terrorist speech, how often is the content checked individually, whether certain languages are overrepresented in the hash database, and how the database contributes to the creation of automated tools.[113] There is also no public scrutiny over how GIFCT is updated and used.[114] Furthermore, researchers do not have access to GIFCT, making any sort of research on the underlying mechanics of the database difficult.[115]

GIFCT is an example of the opacity with which automated tools are developed devoid of accountability and concentrate power in the hands of a few private companies.[116] When an AI tool flags civil rights activists' speech as extremist content for reposting it for journalistic purposes, it becomes apparent that the issues in an automated system are the result of an intersection between the technology, regulatory environment, and the entities that make choices about how to train and use the tool.[117]It is an opportunity to reflect on what data is being used as well as *who* is engaged in collecting and labeling that data. The issues with the moderation of just one specific category of content, i.e., terrorist content, should also raise alarm about how other kinds of category-based tools are labeled and designed.

III. FOSTERING ACCOUNTABILITY: A SYSTEMS-LEVEL APPROACH

Thus far, this article has highlighted the lack of sufficient attention toward the automated tools used in content governance. It has also shown the intertwined technical and normative features of these tools that need deeper examination. This section takes a broad perspective on how we may begin to address this oversight. In this section, I offer a framework to analyze AI tools and develop an approach that focuses on the process and stakeholders involved in creating and using these tools. I do this by using the recently proposed Algorithmic Accountability Act as an example of how regulators may apply it to content governance tools.

A case-by-case approach is ineffective when it comes to AI tools because of the aggregate nature of the automated tool's impact. In response to the limitations of individual rights-based frameworks, we need more precise proposals that incorporate ex ante rights administration and institutional design.[118] In the context of AI, this

---

112. *Id.*

113. Courtney C. Radsch, *GIFCT: Possibly the Most Important Acronym You've Never Heard Of*, JUST SECURITY (Sept 30. 2020), https://www.justsecurity.org/72603/gifct-possibly-the-most-important-acronym-youve-never-heard-of/ [https://perma.cc/9K79-EZDQ].

114. *Id.*

115. *Id.*

116. Radsch, *supra* note 113.

117. *See* Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 641–42 (2020).

118. *See* Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526 (2022).

means paying closer attention to earlier normative decisions about the design of algorithms and user interfaces. Even when people have called for transparency about data, we need to understand how the tools are created, trained, and deployed.[119] I discuss current approaches in legal scholarship addressing systems-level interventions for AI accountability, and suggest how this may be used to study the AI tools used in content moderation and generation.

A. Normative and Regulatory Framework for AI Accountability

Scholars have noted that automated decision-making systems implicate the due process clauses of the Fifth and Fourteenth Amendments.[120] This is because the use of these systems involves the right to be given notice—which requires individuals to be informed of how they will be impacted, the evidence used to make the decisions, and the government agency's decision-making process.[121] While notice would fulfill certain normative goals, it still has limited utility when platforms are moderating millions of individual pieces of content every day. Instead, algorithmic governance requires systemic regulation, and as evidenced by the complex system of technical and institutional design discussed above, collaboration between private and public actors.[122] As Evelyn Douek notes, "error choice is baked in at the moment of ex ante system design."[123] Andrew Selbst and Solon Barocas note that explaining the outcomes of a single case does not provide enough information about the logic or normative values of a system and that it is difficult to provide explanations of causal results for each case.[124]

At a systemic level, there are two elements needed in solutions that address constitutional concerns: ex ante rules to create transparency and impose disclosure requirements, and making aggregate-level litigation remedies more available.[125] Aziz Huq makes the case for a due process analysis of system-level choices because the problems are systemic in the first place.[126] He notes that for an action against the use of an automated decision-making system to be successful, we need more clarity about how large aggregates of data are used. We can also achieve better transparency

---

119. *See* Spandana Singh, *Everything in Moderation*, NEW AMERICA (July 22, 2019), https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/ [https://perma.cc/ENG7-VYQF].

120. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1281 (2007).

121. *See* JERRY L. MASHAW, DUE PROCESS IN THE ADMINISTRATIVE STATE 176 (1985); *see also* Vargas v. Trainor, 508 F.2d 485, 489–90 (7th Cir. 1974) (holding that the notice violated due process in this case because public benefits recipient was not told why benefits were reduced).

122. Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1533–34 (2019).

123. Douek, s*upra* note 118, at 548.

124. Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 87, 1085, 1105 (2018).

125. Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875, 1938. (2019).

126. *Id.* at 1910–11.

by making design choices more available.[127] He also states that there is currently no exploration of the risks arising from the *creation* of large data aggregates.[128] Thus, an understanding of ex ante decisions and processes involved in the creation of content moderation and content generation tools is essential for a regulatory framework. We would need to examine how well the training data, model, and outcomes correspond to one another. We would also need to interrogate the algorithmic design choices starting from the point when subjective decision-making is involved, which means looking at data collection and labeling stages.[129] Furthermore, marginalized groups are sometimes impacted by machine learning systems as a class, especially in cases of bias and stereotyping. To that end, they should be able to act as a group to contest them rather than as individuals.

The concept of "value sensitive design" is instructive here. This constitutes a range of methodologies to identify stakeholders and values in designing systems. Examples of these methods include "the development of value scenarios" and "working with panels of experiential experts." What that looks like in platform governance will invariably involve a range of academic, civil society, and policy stakeholders. [130]

Section 3 of the proposed Algorithmic Accountability Act (AAA) would require the FTC to promulgate regulations to impose certain obligations on companies, such as documentation requirements.[131] These regulations will address the ways consumers are impacted in areas such as education, employment, healthcare, and housing. It also states that the FTC may consider what is an appropriate assessment at each specific point in the technology development life cycle. This is useful because it allows us to envision the kinds of impact assessments we would need. As discussed above, various features of the technology's development implicate different concerns, and it is important that the FTC break down the stages of development so that a one-time assessment or one-size-fits-all approach is not seen as the norm.

To implement any form of reporting requirement, we need better documentation, and many researchers have been working on creating documentation standards and proposing the use of datasheets. The standards on writing datasheets for natural

---

127. *See* CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE 37–38, 241–43 (2022).

128. *Id.* at 1905. In *State v. Loomis*, 881 N.W.2d 749, 761–69 (Wis. 2016), the court rejected an individual due process challenge to the use of COMPAS, an algorithm used to determine sentencing evaluations. Although the challenge was rejected, the Court did note the proprietary nature of COMPAS and the inability to understand more about how group data was being used to calculate risks about individuals. See also Andrea Roth, who introduced a taxonomy that can be used to categorize the various kinds of machine-based evidence that may be introduced in court. This taxonomy includes standards for what constitutes sufficient information about a dataset, and documentation related to it. Andrea Roth, *Machine Testimony*, 126 YALE L. J. 1972, 2026 (2017).

129. Mehtab Khan & Alex Hanna, *The Subjects and Stages of AI Dataset Development*, 19 OHIO ST. TECH. L. J. 171, 197–98 (2023).

130. Bender, et al., *supra* note 79, at 619.

131. Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. § 3 (2002).

language processing state that dataset creators should articulate the rationale and the task for which a dataset is being created.[132]

Saying we need more information about design choices is a vague objective without a clear set of questions about the information we are seeking. Documentation mandates should allow us to identify what specific design choices we need more clarity on. Based on the analysis in Part II, this means we also need more information about the tradeoffs and decisions involved at those stages, such as whether it was important to respond quickly to a content governance issue (such as during the pandemic), an assessment of the level of risk in a particular context (such as elections in a particular jurisdiction), and the level of technological capacity for moderating a certain kind of content (such as choices about which language to build the technology on).[133]

The AAA recognizes the importance of transparency and explainability by stipulating a provision for "transparency, explainability, contestability, and opportunity for recourse" for consumers. Section 5(1)(C) of the AAA requires covered entities to include in the assessment an explanation of why a "critical decision [is] being made and the purpose" for it.[134] Here, it is important to clarify the criteria that the FTC will use to evaluate whether such an explanation is adequate. For content governance tools, we first need to map the harms and risks, and then devise appropriate recourse.

This is why documentation standards need to be more specific. For example, questions in the datasheets may relate to more precise characteristics of a dataset, such as speaker demographic, annotator demographic, immediate source of the data, and speech and text characteristics.[135] These are important questions because in the context of content governance, who decides the classifications and labels of the text matters.

When a text database is created, "the personal judgments of the individuals annotating each document can impact what is constituted as hate speech, as well as what specific types of speech, demographic groups, and so on are prioritized in the training data."[136] This is an issue because even if the database is checked for these biases, the content of what is being taken down in a particular jurisdiction depends on the team responsible for that country. Here, we can anticipate a convergence between the tool's designers and users, and hence highlighting the need for more attention on the people involved in the process. Consider the following scenario: if the tool's designer is able to create something that detects hate speech in a particular context, it is still left to the discretion of the team in that particular place to use that tool to find and remove hate speech, especially in politically contentious situations. For example, one of the challenges to removing hate speech against minorities in

---

132. EMILY BENDER, BATYA FRIEDMAN & ANGELINA MCMILLEN-MAJOR, A GUIDE FOR WRITING DATA STATEMENTS FOR NATURAL LANGUAGE PROCESSING 11 (2021).

133. Douek, *supra* note 94, at 552.

134. Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. § 5(1)(c) (2002).

135. Bender, et. al, *supra* note 107.

136. SPANDANA SINGH, THE LIMITATIONS OF AUTOMATED TOOLS IN CONTENT MODERATION, in EVERYTHING IN MODERATION 17, 19 (2019); *see also* Wenjie Yun & Arkaitz Zubiaga, *Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions*, PEER J COMP. SCI. (2021).

India was that Facebook's team on the ground had ties to the ruling party that included individuals and affiliations who sponsored harmful views towards minorities.[137]

Current standards, community guidelines, and areas of contestation in content moderation can also serve as a roadmap to devising standards. The harmful categories that are presently policed on platforms can be used as a starting point to improve biases in datasets. Platforms do agree on the need to address certain kinds of content, such as CSAM or threats of violence, and the concept of "protected classes" in discrimination law is useful for thinking about some of the biases embedded in datasets.[138] Here we need to recognize the technical and normative issues discussed above that exist at a systemic level, so that we may be able to devise better ways to make the process more inclusive for communities around the world.

Relatedly, the impact of data selection and curation deserves as much importance as what happens once an application is deployed. Algorithmic Impact Assessments (AIAs) should also be applied to the process of curation and creation of a dataset because the process of data collection itself can be unethical.[139] The recent use of private chats on suicide helpline by a company to train NLP illustrates this point.[140] Furthermore, assessments of the data collection and curation process would shed light on the unfair working conditions endured by content moderators while generating data-forming content moderation tools.[141] Data for content moderation tools is collected and sourced through unethical means, like overworked moderators working under stressful conditions.[142] Data for content generation is labeled by poorly paid workers in global south countries.[143]

---

137. Billy Perrigo, *Facebook's Ties to India's Ruling Party Complicate Its Fight Against Hate Speech*, TIME (Aug. 27, 2020), https://time.com/5883993/india-facebookhate-speech-bjp [https://perma.cc/6QXT-9S28].

138. ALEX TAMKIN, MILES BRUNDAGE, JACK CLARK & DEEP GANGULI, UNDERSTANDING THE CAPABILITIES, LIMITATIONS, AND SOCIETAL IMPACT OF LARGE LANGUAGE MODELS 6 (2021).

139. An Algorithmic Impact Assessment is a tool for providing details about an algorithmic system. It involves conducting a variety of analysis and tests on the effects of an AI system. It is measured against several metrics and risk frameworks, such as potential biases, harms, impact on the environment, etc. *See generally* Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeline Clare Elish, Jacob Metcalf, *Assembling Accountability*, *in* DATA & SOC'Y (2021).

140. Alexandra S. Levine, *Suicide Hotline Shares Data with For-profit Spinoff, Raising Ethical Questions*, POLITICO (Jan. 28, 2022), https://www.politico.com/news/2022/01/28/suicide-hotline-silicon-valley-privacy-debates-00002617 [https://perma.cc/WU97-FSPA].

141. GILLESPIE, *supra* note 33.

142. *See* SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019); Billy Perrigo, *Inside Facebook's African Sweatshop*, TIME (Feb. 14, 2022), https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/ [https://perma.cc/2GBC-N4W7].

143. Billy Perrigo, *OpenAI Used Kenyan Workers*, TIME (Jan. 18, 2023) https://time.com/6247678/openai-chatgpt-kenya-workers/ [perma.cc/3N7L-SP82]; Keoni Mahelona, Gianna Leoni, Suzanne Duncan & Miles Thompson, *OpenAI's Whisper is Another Case Study in Colonization*, PAPAREO (Jan. 24, 2023), https://blog.papareo.nz/whisper-is-

The data pipeline and design audits can only work with good faith participation by private companies.[144] Key information is often protected under trade secrecy. Companies are also best placed to provide more information on the impacts of their technology due to the complex nature of how they are developed. As Selbst notes, it becomes difficult to develop any liability regime in the absence of knowledge about the development process.[145]

In addition to gradual participation and disclosure standards for AIAs, other systemic proposals directed at private companies recommend testing, design and documentation requirements, whistleblower protections, and a public-interest course of action.[146] Sonia Katyal calls for both ex ante and ex post evaluation of training data, as well as whistleblower protection for employees working at the companies developing these algorithms. Whistleblowers have been instrumental in shedding light on critical practices inside corporations and have received protection under US law to some extent. Katyal makes a case for whistleblower protection for employees working on AI, as these individuals could provide important information on discrimination, bias, and other AI-related harms.[147] For example, we learned critical information about AI tools such as it not being effective in non-English language contexts through a whistleblower at Facebook.[148] But learning this information in the absence of a clear way to hold platforms accountable meant that there was a lot of criticism but not as much cohesive action.

Lastly, the proposed AAA also leaves the publishing of impact assessment results to the discretion of covered entities.[149] Since information about an AI system could be proprietary, often protected by trade secret law, this requirement would have to be balanced with the public need for information and transparency, and there should be a provision that would create a mechanism for requests for information, especially for research purposes.

CONCLUSION

This article has illustrated why current legal frameworks do not fully capture the dynamic challenges posed by AI tools used in content moderation and content generation. It offers an analysis that centers the technical features and frames these as a regulatory issue. However, it is important to remember the limitations of pure technical fixes to the issues highlighted here. Examining data alone is not enough if we do not situate the practices under overarching legally vacuous spaces and

---

another-case-study-in-colonisation/ [perma.cc/2WRZ-RNMK].

144. Andrew Selbst, *An Institutional View of Algorithmic Impact Assessment*, 35 HARVARD J. L. TECH. 117, 117 (2021).

145. *Id.* at 125.

146. Sonia Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 62 (2019).

147. *Id.* at 126–29.

148. Karen Hao, *The Facebook Whistleblower says its Algorithms are Dangerous. Here's Why.*, MIT TECH. REV. (Oct. 5, 2021), https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/ [perma.cc/FTF2-RZV3].

149. Algorithmic Accountability Act, H.R. 5628, 118th Cong. § 6 (2023).

practices that embody certain values. Private governance initiatives like the Meta Oversight Board are exerting some degree of influence on how automated solutions are used, developed, and contested. The Meta Oversight Board is highlighting the importance of context when it comes to devising rules and policies. Its policy advisory opinion reflects the emphasis on speech context outside the US and EU.[150] Recently, the board has also recommended that Facebook inform users when a content decision was made by an automated tool.[151] The Supreme Court has also highlighted relevant questions about AI tools used in online speech governance.[152] These are all welcome steps but insufficient to address the issues highlighted above. What we need is to imagine a recourse that involves challenging not just a content-based decision but also the technological infrastructure and value system used to build it. Current frameworks for AI accountability allow us to envision a response that captures both.

150. Karissa Bell, *Facebook Agrees to Some Policy Changes in Response to Oversight Board Recommendations*, ENGADGET (Feb. 25, 2021), https://www.engadget.com/facebook-responds-oversight-board-recommendations-210841361.html [perma.cc/B3RJ-GJ3X].

151. *Id.*

152. Eric Goldman, *The Internet Survives SCOTUS Review (This Time)–Twitter v. Taamneh and Gonzalez v. Google*, TECH. & MKTG. L. BLOG (May 18, 2023), https://blog.ericgoldman.org/archives/2023/05/the-internet-survives-scotus-review-this-time-twitter-v-taamneh-and-gonzalez-v-google.htm [perma.cc/7ZRP-FCH4].