Content Regulation in the Digital Age

Submission by
New America's Open Technology Institute

Content regulation in the digital age poses a number of challenges for private companies, state actors, and civil society representatives alike. Consistent with OTI's prior and forthcoming work, this document outlines some of the major practices and considerations with respect to content regulation (Q5), algorithmic filtering (Q7), and transparency around these practices (Q8).

Q5: Content Regulation Processes

What processes are employed by companies in their implementation of content restrictions and takedowns, or suspension of accounts? In particular, what processes are employed to:

1. *Moderate content before it is published;*
2. *Assess content that has been published for restriction or take down after it has been flagged for moderation; and/or*
3. *Actively assess what content on their platforms should be subject to removal?*

Companies use an array of processes to remove and takedown content as well as to suspend accounts. The two most common categories of approaches to content moderation are human moderation and algorithmic moderation. Most companies employ a combination of both of these approaches.

Human-led content moderation is typically executed by teams of content reviewers who evaluate flagged content and accounts and assess whether they violate a company's terms of service and content policies. These general reviewers make judgments on most basic moderation cases. In companies with large and comprehensive content moderation teams, if these general moderators are unsure of how to respond to a certain flag, they can escalate it to higher level individuals in the company.[1] Today, a large amount of content moderation services are provided by outsourced employees in countries such as the Philippines and India.[2] However, some companies also train and employ their own teams of content specialists in offices around the world, who can accurately moderate content based on local contexts and laws. These employees are not cheap, however. Therefore it is difficult for smaller companies that are resource-strapped to build and train their own content moderation teams and scale these operations quickly, an expectation most companies face.[3]

The second approach to content moderation—namely, algorithmic content moderation—often involves the use of digital hashes and natural language processing tools in order to better inform content moderation processes. Later in this comment, we analyze the use of algorithms for content moderation amongst leading Internet companies, and highlight some of their benefits and

[1] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review* 131 (March 2017)
[2] Klonick, "The New Governors."
[3] Klonick, "The New Governors."

drawbacks. We also specifically note the criticality of increasing and refining the use of algorithmic content moderation to detect, identify, and voluntarily take down content.

Where scale and tools allow, some content can be screened and moderated before publication in order to detect content that has been pre-determined as illegal. In the United States, for example, this includes child pornography and copyrighted content. Most online platforms can detect child pornography materials throuhg the use of a picture recognition algorithm called PhotoDNA.[4] PhotoDNA was originally developed by Microsoft and has expanded to become a powerful tool used by companies such as Microsoft, Twitter, Google and Facebook, law enforcement and organizations such as the National Center for Missing & Exploited Children alike.[5] PhotoDNA works by converting existing illegal child porn images online (of which there are over 720,000[6]) into a grayscale format. It then overlays the images onto a grid and assigns each square a numerical value. The designation of a numerical value converts the square into a hash, or digital signature, which remains tied to the image and can be used to identify other recreations of the image online. PhotoDNA is particularly effective as it is not susceptible to alterations such as resizing and color alterations, and it can therefore detect hashes across a broad online spectrum in only microseconds.[7]

Hash-based technologies are particularly successful in pre-publication as they enable platforms to remove illegal content, which companies are obligated to remove. Following the success of the PhotoDNA technology with moderating child pornography materials, Henry Farid of Dartmouth College expanded the functionality of the tool so that it could also be applied to extremist and terror-related content.[8] In December 2016, Facebook, Twitter, Microsoft and YouTube created a shared industry database of hashes for extremist content, in an attempt to standardize and scale up content moderation in this area.[9] This however, is a field of content in which legality is not as easily defined. Pre-publication moderation practices, therefore, have been scrutinized for chilling freedom of expression, especially for marginalized and minority groups.

---

[4] Klonick, "The New Governors."

[5] Microsoft, "New Technology Fights Child Porn by Tracking Its “PhotoDNA”," *Microsoft*, last modified December 15, 2009, https://news.microsoft.com/2009/12/15/new-technology-fights-child-porn-by-tracking-its-photodna/#sm.0001mpmupctevct7pjn11vtwrw6xj.

[6] Klonick, "The New Governors."

[7] Microsoft, "Photo DNA: Step by Step," *Microsoft*, https://www.microsoft.com/global/en-us/news/publishingimages/ImageGallery/Images/Infographics/PhotoDNA/flowchart_photodna_Web.jpg. A similar digital hash system, known as contentID was developed by YouTube to identify content that violates copyright laws. ContentID allows content creators on the platform to assign a “digital fingerprint” to their work so it can be compared and screened against content that is uploaded onto the platform in the future. Similar to the PhotoDNA technology, it flags any similarities in hashes and thus enables for fast and efficient pre-publication moderation.

[8] Kaveh Waddell, "A Tool to Delete Beheading Videos Before They Even Appear Online," *The Atlantic*, June 22, 2016, https://www.theatlantic.com/technology/archive/2016/06/a-tool-to-delete-beheading-videos-before-they-even-appear-online/488105/.

[9] Facebook, "Partnering to Help Curb Spread of Online Terrorist Content," *Facebook*, last modified December 5, 2016, https://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/.

Collaborations between companies to create more effective content moderation strategies however, are encouraged and useful.

Concerns regarding the limiting of freedom of speech and expression online are further heightened by the lack of transparency by companies regarding their content moderation practices and policies. This gray area is particularly concerning as it enables platforms to become the arbiters of justice and determine what kinds of content should be permissible and what kinds of content should be silenced in the modern-day public forum. While social media sites are privately run, and therefore subject to private rules and policies, the lack of transparency around these rules raises concerns about the state of democratic discourse online. These concerns have for example emerged numerous times with the Apple App Store. According to Apple's App Store Review Guidelines, porn, racism and "mean-spirited" content is not permitted.[10] As a result Apple has rejected content from the App Store that disseminates pornography as well as malware, which are clearly objectionable. However, the company has also rejected content that disseminates negative information on the company by for example, discussing the negative aspects of smartphone production. The company has also removed apps that are considered "politically problematic"[11] such as Metadata, an app that alerted a user every time a US-led drone strike occurred. The censoring of these kinds of apps demonstrate a hindrance of freedom of speech and the creation of barriers to accessing vital information. However, without transparency into what policies guided these decisions, it is difficult to make a case to reverse these decisions or to prevent further instances of these from taking place.

With respect to content that has already been published, companies regularly receive reports and flags to remove or moderate content. These flags come from users as well as from governments and government agencies. In the case of regular government requests, most companies follow a standard procedure. Upon receiving a government request to remove or moderate content, Microsoft will for example, assess the rationale behind the request, evaluate the authority and jurisdiction of the requesting party or agency, and gauge how the request interfaces with Microsoft's terms of service and the policies it has in place for all of its consumer online services (Bing, OneDrive, BingAds and MSN). Based on the results of this review process, the company will decide whether or not to moderate or remove the content in question[12].

Companies' content moderation practices are also guided by international laws and policies. In the European Union for example, individuals have the "right to be forgotten" online.[13] As a result, search engines are required to provide users the option to de-list URL's and remove search results featuring their names[14] if the results are "inadequate, inaccurate, no longer relevant

---

[10] Louise Matsakis, "Apple's Long History of Rejecting 'Objectionable Content' From the App Store," *Motherboard*, July 17, 2017, https://motherboard.vice.com/en_us/article/a3dwq8/apples-long-history-of-rejecting-objectionable-content-from-the-app-store?utm_source=mbtwitter.

[11] Matsakis, "Apple's Long,".

[12] Microsoft, "Content Removal Requests Report," Microsoft, https://www.microsoft.com/en-us/about/corporate-responsibility/crrr.

[13] Microsoft, "Content Removal," *Microsoft*.

[14] Google, "Search Removals Under European Privacy Law," *Google*, https://transparencyreport.google.com/eu-privacy/overview.

or excessive".[15] The guidelines on the right to be forgotten have been further expanded on in Russia, where it has been mandated that if a user provides proof to support their request, a company must comply.[16]

A recently passed law in Germany has also had profound ramifications on how content is moderated on online platforms. According to the law, which aims to stymie the spread of hate speech and extremist content, companies must remove "illegal, racist or slanderous" comments and posts within 24 hours of their posting or they face fines of approximately $57 million.[17] Following the introduction of this law, free expression advocates spoke out about concerns that companies would censor permitted expression given the pressure to comply with the 24-hour time frame. The laws have been further broadened to allow companies seven days to review the content after it has been removed in order to determine whether it should be reposted.[18] However, given the rapid nature of technology, seven days is enough time to stifle important voices and movements.

Another common example of differing laws and policies influencing content moderation practices is through geo-blocking. Geo-blocking is the process of blocking a user's access to certain types of content based on their geographic location.[19] These moderation approaches are typically put in place at the request of local governments or due to local laws and policies. For example, in 2012 a video titled *The Innocence of Muslims* was uploaded to YouTube. The video showed Muslims burning the homes of Egyptian Christians and depicted Muhammad as a "bastard, homosexual, womanizer, and violent bully."[20] The release of the video resulted in mass outcry and protests in the Islamic world and as a result the governments of Libya and Egypt demanded that access to the video be blocked in their countries. YouTube complied with the requests, but left the video up in other countries as it did not violate their guidelines for acceptable content.[21]

Companies such as Google and Microsoft also regularly receive requests by users and organizations to moderate and remove copyrighted content from their platforms. Typically, these requests are placed based on the stipulations put forth by the Digital Millennium Copyright Act.[22] Requests that are submitted require users to include links to the content in question. Requests are typically processed manually by human moderators. If these moderators determine the

---

[15] Microsoft, "Content Removal," *Microsoft*.
[16] Microsoft, "Content Removal," *Microsoft*.
[17] Melissa Eddy and Mark Scott, "Delete Hate Speech or Pay Up, Germany Tells Social Media Companies," *New York Times*, June 30, 2017, https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html?mtrref=undefined&gwh=3B9C3FF4D784BC314A10080CCC48B6CC&gwt=pay.
[18] Eddy and Scott, "Delete Hate."
[19] Klonick, "The New Governors."
[20] Klonick, "The New Governors."
[21] Klonick, "The New Governors."
[22] Google, "Requests to Remove Content Due to Copyright," *Google*, https://transparencyreport.google.com/copyright/overview.

request is legally sufficient, it results in the removal of those links from search results on search engines such as Bing and Google.[23]

User-flagged content is a significant component of the content that is moderated by companies on a daily basis. This comes with both its benefits and its challenges. User and civil society input and support in flagging and reporting content for moderation has proven to be valuable for companies. As a result, some platforms such as YouTube have decided to capitalize on and encourage the involvement of users and organizations to make their platforms safer. For example, YouTube introduced the Heroes program, which offers users perks such as sneak peek access to new products and features in exchange for them reporting a certain amount of objectionable content that violates their Terms of Service. In addition, in 2014, YouTube created a list of 20 "super flaggers," individuals and organizations whose flagging contributions were considered accurate, valuable and as enhancing the quality of content and safety on the platform. One of the most recognized of these "super flaggers" is the U.K.'s Metropolitan Police Counter Terrorism Internet Referral Unit, members of which regularly flag for moderation and removal content that is extremist in nature.

Twitter also regularly utilizes the input of users and organizations in order to shape and inform their moderation practices. In February 2016, the company announced the establishment of a Trust and Safety Council which was composed of members of civil society around the world with a vested interest in preventing online issues such as cyberbullying, child pornography, online harassment and cyberstalking. These organizations regularly provide input to the organization on cases and forms of content to look out for as well as how the platform and moderation practices can be made safer.[24] This is a practice that other companies should adopt, as it fosters strong relationships with organizations working in tandem to combat objectionable content online and strengthens overall efforts.

Despite the fact that the moderation of flagged and reported content is a major component of the content moderation practices of online platforms, there is little transparency regarding what exact procedures and guidelines these companies follow. In May 2017, the *Guardian* released findings they had acquired from reviewing over 100 internal training manuals, spreadsheets and flowcharts related to Facebook's internal content moderation guidelines. The guidelines touched on topics including violence, terrorism, pornography, hate speech, racism and self-harm. For example, the guidelines stated that the platform would allow users to livestream attempts to self-harm as it didn't "want to censor or punish people in distress" and that some photos of non-sexual physical abuse and bullying of children were permissible unless they featured a "sadistic or celebratory element".[25] The release of these findings sparked outcry regarding some of the practices followed by the company and resulted in calls for the guidelines to be refined in order to make the platform safer.

---

[23] Microsoft, "Content Removal," *Microsoft*.
[24] Patricia Cartes, "Announcing the Twitter Trust & Safety Council," *Twitter*, last modified February 9, 2016, https://blog.twitter.com/official/en_us/a/2016/announcing-the-twitter-trust-safety-council.html.
[25] Nick Hopkins, "Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence," *Guardian*, May 21, 2017,
https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.

Although most companies remain secretive regarding the policies that guide their content moderation practices, some platforms have released portions of their internal manuals in an attempt to promote transparency. Tumblr for example, released a copy of their internal guidelines shortly after the Facebook leak. Among the forms of content moderated on the platform are accounts that "actively promote self harm", including eating disorders.[26] While this has in some ways augmented the safety of users on the platform, it has also raised a number of freedom of speech concerns for blogs that discuss the topic of self-harm, rather than promote it. The cases of Facebook and Tumblr therefore demonstrate the continuous struggles companies face in moderating content as they must continuously balance safety with freedom of speech and expression.[27]

Although companies regularly face criticisms that their content moderation practices are not sensitive to the perspectives of individuals and groups from different ethnicities, genders, religions, cultures and nations, it is important to recognize that a number of positive developments have been made over the past decade to improve the applicability of content moderation practices globally. For example, under the leadership of Nicole Wong, former Vice President and Deputy General Counsel for Google, YouTube made a number of strides towards making content moderation practices more considerate of varying user perspectives and backgrounds and towards preventing the over censorship of users.[28] In 2006 for example, a video of Saddam Hussein's hanging was posted on YouTube, as well as a video of his corpse lying in the morgue. Wong's team decided to remove the video of his corpse from the platform, citing it as an example of gratuitous violence. A number of other examples exist as well.[29]

Platforms such as Flickr are also notable for their attempts to clearly define their responsibility when it comes to assessing content on their platforms. Upon uploading images to the platform, users must rate their photos as either "safe," "moderate," or "restricted."[30] The company then actively assesses and patrols content based on the ratings assigned the images, rather than assessing all content simultaneously. When users flag or report images, their complaints are phrased "I don't think this photo is flagged at the appropriate level," demonstrating how unlike other platforms, Flickr's framework for content moderation is not determined by a set of rules that can be violated, but rather by a classification system that can result in misclassifications.[31] Although this approach to content moderation and assessment reduces the platform's responsibility and enables them to triage content, it also raises significant safety concerns as users have no substantive, high-priority way of notifying the platform of the prevalence of

---

[26] Eva Galperin, "What the Facebook and Tumblr Controversies Can Teach Us About Content Moderation," *Electronic Frontier Foundation*, last modified March 2, 2012, https://www.eff.org/deeplinks/2012/03/what-facebook-and-tumbler-controversies-can-teach-us-about-content-moderation.

[27] Galperin, "What the Facebook," *Electronic Frontier Foundation*.

[28] Klonick, "The New Governors."

[29] For additional information, see Klonick, "The New Governors." (discussing the influence of historically relevant events such as the Iranian Green Movement and the Arab Spring on content moderation practices)

[30] Crawford and Gillespie, "What is a Flag."

[31] Crawford and Gillespie, "What is a Flag."

certain images that should not be on the platform at all.[32] Given that different companies assume a varying degree of responsibility in moderating content, transparency into their policies is important for creating a safe internet.

As companies patrol and assess content that could be classified as extremist messaging, hate speech or discriminatory, one of the primary concerns is that these efforts could also silence counter messaging and counter speech, which aim to introduce counter narratives to objectionable content. According to Facebook for example, they remove 288,000 hate speech posts every month.[33] However, according to civil society members and activists, because the platform's content moderation guidelines are not clearly laid out, it is difficult to assess whether they are biased against certain groups, whether the company is doing the most they possible can to create a safe space for individuals and groups online and whether their content moderation efforts are in fact silencing counter narratives.[34]

Q7: Automation and Content Moderation

What role does automation or algorithmic filtering play in regulating content? How should technology as well as human and other resources be employed to standardize content regulation on platforms?

Automation and algorithmic filtering—for the purposes of this section referred to collectively as 'algorithmic content moderation'—are deep-rooted at internet platform companies and telecommunications firms, among others. They are integral and powerful tools that help keep online communications and content clean, secure, and legal. Companies that operate over the Internet should be empowered and encouraged to continue expanding the application of these tools to keep the internet safe.

Today's leading technology companies operate at a scale so massive and global that it necessitates the use of algorithmic technology to keep their platforms safe. For example, 60 billion messages are sent over Messenger and WhatsApp together, per a 2016 report.[35] In that same year, over 500 million tweets were sent each day on Twitter,[36] and more than 100 million people were using Instagram's then-new stories feature.[37] With such tremendous reach in their user networks, these internet services host and disseminate a universe of content that is vastly too

---

[32] Crawford and Gillespie, "What is a Flag."

[33] Tracy Jan and Elizabeth Dwoskin, "A White Man Called Her Kids The N-Word. Facebook Stopped Her From Sharing It.," *Washington Post*, https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html?utm_term=.81303e2184cd.

[34] Jan and Dwoskin, "A White."

[35] Lauren Goode, "Messenger and WhatsApp process 60 billion messages a day, three times more than SMS," *The Verge*, April 12, 2016, https://www.theverge.com/2016/4/12/11415198/facebook-messenger-whatsapp-number-messages-vs-sms-f8-2016

[36] Kit Smith, "44 Twitter Statistics for 2016," *brandwatch*, May 17, 2016, https://www.brandwatch.com/blog/44-twitter-stats-2016/.

[37] Darrell Etherington, "Instagram Stories has 100 million daily active users after just 2 months," *TechCrunch*, October 6, 2016, https://techcrunch.com/2016/10/06/instagram-stories-has-100-million-daily-active-users-after-just-2-months/.

difficult for the human mind to contemplate, let alone moderate. Utilization of algorithmic content moderation can help these companies defeat this fundamental issue of scale and analyze troves of information in real time and with great efficiency.[38] Algorithmic content moderation has been employed by Internet services for several years and should continue as the industry standard for keeping the web safe.

Algorithmic technologies can be used in the context of content moderation in two primary ways. First, algorithms may be developed by an Internet company to identify content that, with some level of confidence, violates the company's standards for content or private communications. For instance, many leading Internet companies are actively refining machine learning algorithms to detect content that violates their platform policies against child exploitation and violent material. [39] Second, algorithms may be developed to automatically take down content that clearly violates the company's policies – like Internet ads that display imagery of cigarettes and alcohol, among many other examples. There are several industry-standard algorithmic content moderation tools that are in wide use today that help companies achieve some combination of this automatic filtering and content takedown, including Microsoft's PhotoDNA Cloud Service, which helps clients automatically detect the sharing of child exploitation images;[40] YouTube's Content ID service, which helps copyright owners protect their digital content;[41] and various spam detection tools.[42] These algorithmic technologies typically rely on identifying patterns of behavior (e.g., high rates of spam message) or matching new uploads with previously identified unwanted content (e.g., keywords, URLs, or image hashes).[43]

Unfortunately, while these and other technologies are in wide use across the industry to enable algorithmic content moderation, they are not flawless. Some these drawbacks include that they typically do not include all policy-violating or otherwise egregious content in their databases or frameworks; their potential to be circumvented by sophisticated content sharers; the frequent presence of bias or a lack of representation in the data and algorithms used to build these tools; the complexity of perpetually evolving communications; and the non-transferability of some of these tools across different Internet platforms. Additionally, these tools often require human involvement to mitigate algorithmic errors.[44]

---

[38] Kate Cox, "Facebook's Robots Are Working Hard On Content Moderation So Humans Don't Have To," *Consumerist, June 1, 2016, https://consumerist.com/2016/06/01/facebooks-robots-are-working-hard-on-content-moderation-so-humans-dont-have-to/.*

[39] See, for instance, Sarah Perez, "YouTube promises to increase content moderation and other enforcement staff to 10K in 2018," *TechCrunch*, December 5, 2017, https://techcrunch.com/2017/12/05/youtube-promises-to-increase-content-moderation-staff-to-over-10k-in-2018/; or Josh Constine, "Facebook spares humans by fighting offensive photos with AI," *TechCrunch, May 16, 2016,* https://techcrunch.com/2016/05/31/terminating-abuse/.

[40] Linda Kinkade, "How child predator was caught by tiny clue in photo he posted online," *CNN*, March 19, 2017, http://www.cnn.com/2016/04/21/us/project-vic-child-abuse/index.html.

[41] Allegra Frank, "YouTube is changing the Content ID system in an effort to help creators," *Polygon*, April 28, 2016, https://www.polygon.com/2016/4/28/11531228/youtube-content-id-changes-copyright-dispute-jim-sterling.

[42] Emma Llanso, "Automation in Content Moderation: Capabilities and Limitations," Centre for European Policy Studies, last modified September 2017, https://www.ceps.eu/sites/default/files/Emma%20Llanso%20CDT.pdf.

[43] Llanso, "Automation in Content," Centre for European Policy Studies.

[44] Llanso, "Automation in Content," Centre for European Policy Studies.

One issue in particular that would would draw attention to is the tendency for algorithms to perpetuate discriminatory or biased outcomes in some cases. Algorithmic content moderation is designed to mimic human rationality and judgment to the extent possible, but researchers have consistently found that machine learning algorithms, including in the context of content moderation, have in the past adopted various demographic biases and were therefore unable to moderate content without biased discrepancies. Researchers have accordingly suggested that, to promote fairness in these contexts, tests and trainings conducted on an ongoing basis might be required to test, monitor and evaluate content moderating technologies.[45]

Many of these concerns can be more effectively addressed if the leading companies responsible for performing algorithmic content moderation are more transparent about the ways they develop these algorithms and the rules they use to enforce them. For example, Internet companies can be more transparent about how they define offending content, be it hate speech, violence, or nudity. Some are more transparent about their practices with others, but over time, our hope is that the industry can work with the public to increase efforts around transparency in relation to content moderation procedures.

We would also highlight the ongoing necessity for human involvement in the definition and development of content moderation processes. Even if a platform company develops a procedure for identifying and removing graphical content that, for instance, includes nudity, it is highly likely that manual review will be required to handle content that has escaped the grasp of the moderation algorithm. For example, Microsoft Azure's Content Moderation tool, which offers off-the-shelf industry-standard algorithm-powered moderation for images, text, and video, deploys a combination of machine learning and data matching to detect violating content, but also includes a human review tool to better advise the algorithmic processes underlying the tool over time.[46] Humans can often augment algorithmic models, particularly when prediction confidence needs to be contextualized for the real world. Indeed, humans can focus on the edge cases and help the models learn and get better over time.[47]

As a final note, we would bring attention to ongoing debates about the moderation of online extremist and violent content. There is presently intense pressure on Internet companies to develop stronger tools to detect particular types of content, including hateful conduct and terrorist/extremist speech.[48] OTI recognizes the difficult challenges for internet companies in identifying such content while ensuring that permitted content is not inappropriately removed.. We encourage internet companies to continue to refine existing moderation tools, collaborate

---

[45] Reuben Binns et al., "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation" (Cornell University, 2017), https://arxiv.org/abs/1707.01477.

[46] Microsoft, "Azure Developer Tools," *Microsoft*, https://azure.microsoft.com/en-us/tools/.

[47] Microsoft, "Human-In-The-Loop," *Microsoft*, last modified June 27, 2017, https://docs.microsoft.com/en-us/azure/cognitive-services/content-moderator/review-tool-user-guide/human-in-the-loop.

[48] See, for example, Heather Stewart, "May calls on internet firms to remove extremist content within two hours," *The Guardian*, September 19, 2017, https://www.theguardian.com/uk-news/2017/sep/19/theresa-may-will-tell-internet-firms-to-tackle-extremist-content.

with civil society advocates, and embrace transparency as they further advance technologies and implement new policies and practices.

Q8: Transparency
What information should companies disclose about how content regulation standards under their terms of service are interpreted and enforced? Is the transparency reporting they currently conduct sufficient?

In the wake of new terrorist threats, changing hate speech laws, and the growing user bases of major social media platforms, tech companies are under more pressure than ever with respect to how they treat the content on their platforms. Companies have been under attack to be more proactive, remove more content, and remove it faster,[49] while also coming under fire for taking down too much content or lacking appropriate remedial measures.[50]

This confusion has led to further confusion: Some have complained that social media platforms are pushing a particular agenda via their content moderation efforts,[51] the left is calling for those same platforms to take down more extremist speech,[52] and free expression advocates are deeply concerned about whether the companies taking down too much content, or if there's too much collateral damage happening.[53]

Meanwhile, there is a lot of confusion about what exactly the companies *are* doing with respect to content moderation. The few publicly available insights into these processes reveal both bizarre and idiosyncratic rule sets that could benefit from greater transparency. The question of how to address that need for transparency, however, is difficult. There is a clear need for hard data about specific company practices and policies on content moderation, but what does that look like? What numbers should be reported? What data would be most valuable? And what is the most accessible and meaningful way to report this information?

Consistent with OTI's work pushing for transparency around government requests for user information,[54] we are advocating for a similar level of transparency around content moderation. The benefits to the companies of such transparency are significant: For those companies under pressure to "do something" about terrorist speech online, this is a an opportunity to outline the lengths to which they have gone to do just that; for companies under fire for "not doing enough," a transparency report would help them to explain that there is no magic artificial intelligence

[49] Natasha Lomas, "Tech Giants Told to Remove Extremist Content Much Faster," *TechCrunch*, September 20, 2017, https://techcrunch.com/2017/09/20/tech-giants-told-to-remove-extremist-content-much-faster/.
[50] Jan and Dwoskin, "A White."
[51] Steven Overly and Ashley Gold, "Tech firms' fight against hate could haunt them," *Politico*, August 19, 2017, https://www.politico.com/story/2017/08/19/tech-firms-fight-hate-backlash-241807.
[52] Jan and Dwoskin, "A White."
[53] Emma Llanso, "Takedown Collaboration by Private Companies Creates Troubling Precedent," *Center for Democracy and Technology*, December 6, 2016, https://cdt.org/blog/takedown-collaboration-by-private-companies-creates-troubling-precedent/.
[54] Liz Woolery, Kevin Bankston, and Ryan Budish, "The Transparency Reporting Toolkit - Guide and Template," *New America*, last modified December 29, 2016, https://www.newamerica.org/oti/policy-papers/transparency-reporting-toolkit-reporting-guide-and-template/.

wand they can wave and make harrassment online disappear;[55] and finally, public disclosure about content moderation and terms of service practices will go a long way toward building trust with users – a trust that has crumbled in recent years.[56] Equally significant, before we can have an intelligent conversation about hate speech, terrorist propaganda, or other worrisome content online, this transparency is necessary. Facebook has said they are entering a new of era transparency for the platform.[57] Twitter has published some data about content removed for violating its TOS,[58] Google followed suit for some of the content removed from YouTube,[59] Microsoft[60] has published data on "revenge porn" removals, and Automattic recently shared a blog post about its own efforts to moderate content on Wordpress. While each of these examples is a step in the right direction, what we need is a full and consistent push across the sector. So what does that look like?

What would meaningful transparency around content moderation and terms of service enforcement look like? Existing Transparency reports that detail the number of requests for user information received offer a logical starting point. These reports are aimed at a lay audience while providing insight into the volume and scale of requests that impact users' privacy and freedom of expression. Ultimately, platforms like Facebook and Twitter ought to share information that best educates readers, policymakers, advocates, and others about how content is handled on their sites.

While there have been calls for publication of such information, there has been little specificity with respect to *what exactly* should be published. No doubt this is due, in great part, to the opacity of individual companies' content moderation policies and processes: It is difficult to identify specific data that would be useful without knowing what data is available in the first place. Anecdotes and snippets of information from companies like Automattic and Twitter offer a starting point for considering what information would be most meaningful and valuable.[61]

Data that offers some show of the scope and volume of content removals, account removals, and other forms of account or content interference/flagging would, in itself, be a logical starting point.[62] Information about content that has been flagged for removal by a government actor (as is

---

[55] Llanso, "Automation in Content," Centre for European Policy Studies.
[56] "The Fate of Online Trust in the Next Decade," *Pew Research Center*, last modified August 10, 2017, http://www.pewinternet.org/2017/08/10/the-fate-of-online-trust-in-the-next-decade/.
[57] https://newsroom.fb.com/news/2017/10/update-on-our-advertising-transparency-and-authenticity-efforts/
[58] "Combating Violent Extremism," *Twitter*, last modified February 5, 2016, https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html; Twitter, Inc., Transparency Report, 2016, https://transparency.twitter.com/en/removal-requests.html.
[59] "Why Flagging Matters," YouTube Official Blog, last modified September 15, 2016, https://youtube.googleblog.com/2016/09/why-flagging-matters.html.
[60] Microsoft, "Content Removal," *Microsoft*.
[61] See, e.g, Transparency Report, 2016, https://transparency.twitter.com/en/removal-requests.html; Tackling Extremist Content on WordPress.com, *Automattic*, December 6, 2017, https://transparency.automattic.com/2017/12/06/tackling-extremist-content-on-wordpress-com/.
[62] Danielle Keats Citron, "Extremist Speech and Compelled Conformity" (March 27, 2017). Notre Dame Law Review (Forthcoming); U of Maryland Legal Studies Research Paper No. 2017-12. Available at SSRN: https://ssrn.com/abstract=2941880

the case with YouTube and the U.K.'s Counterterrorism Internet Referral Unit,[63] for example), should be included. More granular information, such as the reasons why a piece of content may be removed, or which part of the terms of service was violated, would provide even more meaningful transparency.

Ultimately, the quantitative data (i.e., numbers and percents) is valuable, but when combined with qualitative data, a company can do more than shed light on an opaque process, it can tell a story. As transparency reporting around government requests for user information has evolved, many of the companies publishing those reports have added qualitative data, including anecdotes and examples, to their reports. However, as with those traditional transparency reports, a certain amount of user education may be necessary. For individuals to understand the content they are reading or the numbers they are poring over, they need information on how companies handle various policy issues and processes and how companies define various terms.[64]

Regardless of the approach, the most valuable information – that which will lead to greater accountability of both companies and governments – is meaningful insight into how companies apply policies and practices that impact users' freedom of expression and privacy.

[63] Crawford and Gillespie, "What is a Flag."
[64] Freedom Online Coalition Working Group 3 on Privacy and Transparency Online, *Submission to UN Special Rapporteur David Kaye: Study on Freedom of Expression and the Private Sector in the Digital Age*, www.ohchr.org/Documents/Issues/Expression/PrivateSector/FreedomOnlineCoalition.pdf.