Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation

Keller, D.; Leerssen, P.

DOI
10.1017/9781108890960

Publication date
2020

Document Version
Final published version

Published in
Social Media and Democracy

10

Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation

Daphne Keller and Paddy Leerssen

INTRODUCTION

We live in an era of increasing worry that internet platforms like Facebook or Twitter, which mediate our online speech, are also fomenting hatred, spreading misinformation, and distorting political outcomes. The 2016 US presidential election, in particular, unleashed a torrent of concern about platform-borne harms. Policymakers around the world have called for laws requiring platforms to do more to combat illegal and even merely "harmful" content.

From the perspective of platforms themselves, these proposals have a lot in common. Regardless of their *substantive* mandates – to address content that is misleading, hateful, or violent, for example – they all require similar *operational* processes to comply. Platforms already have these processes in place to enforce current laws and their discretionary Community Guidelines. Any new efforts to regulate content online will likely build on existing systems, personnel, and tools – and inherit both their strengths and their weaknesses. That makes it important to understand those systems.

Reliable information about platforms' content-removal systems was, for many years, hard to come by; but data and disclosures are steadily emerging as researchers focus on the topic and platforms ramp up their transparency efforts. This chapter reviews the current and likely future sources of information.

Some content takedowns are required by law, while others are performed voluntarily. Legal takedowns are shaped by intermediary liability laws, which tell platforms what responsibility they have for unlawful content posted by

Daphne Keller directs the Program on Platform Regulation at the Stanford Cyber Policy Center and was formerly Associate General Counsel to Google. Paddy Leerssen is a PhD candidate at the Institute for Information Law (IViR), University of Amsterdam, and a nonresident fellow at the Stanford Center for Internet and Society.

users. Platforms operating under legal frameworks like the US Digital Millennium Copyright Act (DMCA) or the EU's eCommerce Directive typically meet their legal obligations using "notice-and-takedown" systems. Larger platforms invest heavily in these operations and sometimes supplement them with proactive efforts to identify and eliminate illegal material. However, the evidence we review suggests that platforms perform poorly at enforcing the law consistently at scale.

Platforms' voluntary content removals are based on private rulesets: Community Guidelines. These private standards often prohibit a broad margin of lawful speech beyond that which actually violates the law. Community Guidelines may draw on platform operators' own moral beliefs or social norms. They may also simply aim to shape the user experience for business purposes. A real estate listing site, for example, might exclude photos that do not show buildings and property. Most websites prohibit spam, pornography, and harassment for comparable reasons. Platforms can also use Community Guidelines to simplify their legal enforcement efforts and avoid conflict with governments, by simply prohibiting more speech than the law does. Governments, for their part, may avoid challenges under constitutions or human rights law if removal of legal content is attributed to private, rather than state, action (Angelopoulos et al. 2016). Therefore, as policymakers and the public have increasingly demanded that platforms remove content that is harmful or offensive, but not necessarily illegal, these discretionary rules have become ever more important.

New platform efforts to weed out prohibited content will, inevitably, have a lot in common with these existing systems. That is partly good news, because policymakers are not drafting on a blank slate. Lawyers, researchers, and platform employees have two decades of experience in the ways that content-removal systems work in practice. It is also, however, partly bad news. Evidence suggests that platforms do not do a great job as enforcers of speech rules. Even when they apply their own, self-defined Community Guidelines, the results often appear erratic. It is hard for independent researchers to quantify platforms' accuracy in applying Community Guidelines standards, though, since the rules themselves are poorly understood and there is little reliable information about the specific text, images, videos, or other content removed.

Platforms' performance under laws like US copyright law or German hate speech law can be easier to research, in part simply because the rules come from public law rather than platforms' discretionary Community Guidelines. There is also somewhat more information available about how platforms apply legal rules and what specific content they take down. [1] This has allowed independent experts to assess the platforms' removal practices – and document significant problems. Platforms that receive notices alleging illegality, and hence exposing

[1] This chapter does not attempt to list the research on content takedown in more strongly speech-repressive countries such as China. See, e.g., Human Rights Watch (2006).

them to legal risk, commonly err on the side of caution, removing even lawful information. Some simply take down any content identified in a complaint. This "over-removal" is a constant byproduct of notice-and-takedown systems. Platforms have removed information ranging from journalism and videos documenting police brutality in Ecuador (Vivanco 2014) to media coverage of fraud investigations in the United States (Cushing 2017) to criticism of religious organizations (Galperin 2008) to scientific reporting (Oransky 2013; Timmer 2013).

Platforms historically have had little incentive to share detailed information about content removal with the public. Compiling records of evolving content takedown processes, which may use different tools and standards or be managed by different internal teams over time, is burdensome; and any disclosure, particularly one that admits error, can be used against platforms in court or in the press. Yet the longer-term benefits of greater transparency, for both society and platforms themselves, are becoming ever more evident. Without it, public debates about platform responsibility can become exercises in speculation. Laws passed without a practical foundation in platforms' real-world operations and capabilities can be burdensome for the companies and their users, yet fail to achieve lawmakers' legitimate goals.

Whether in recognition of this problem, or because of increasing pressure from civil society, academia, and other quarters, some platforms have provided substantially more public transparency in recent years. This chapter will review major sources of information released by platforms, as well as independent research concerning content takedown operations. We will begin in the section "Takedown and Intermediary Liability Laws" by very briefly reviewing intermediary liability law, which plays a central role in structuring platforms' content-removal operations. One particularly robust academic study will serve to illustrate common platform takedown practices and research themes.

The section titled "Sources of Information," which makes up the bulk of the chapter, provides a broader review of the current empirical literature and likely sources of future information. First, we discuss disclosures from platforms and other participants in content moderation, such as users and governments. Second, we discuss independent research from third parties such as academics and journalists, including data analysis, interviews, and surveys. Finally, before concluding the chapter, we will list specific questions and areas for future empirical research.

Debates about proposed new laws ranging from the EU's Terrorist Content Regulation to Singapore's "fake news" law should be informed by empirically grounded assessments of platforms' capacity to comply and the potential unintended consequences of their compliance efforts. Without better information about platforms' true strengths and weaknesses as speech regulators, we should not expect to see well-designed laws.

TAKEDOWN AND INTERMEDIARY LIABILITY LAWS

This section discusses intermediary liability laws, which form the legal backdrop for content-moderation discussions. As we explain, these laws determine when and how platforms are legally required to remove content. Following a brief legal analysis, we show how such laws operate in practice with the help of one particularly thorough study: "Notice and takedown in everyday practice" (Urban, Karaganis, and Schofield 2016).

Intermediary Liability Laws

Intermediary liability laws tell internet intermediaries such as ISPs, search engines, or social media companies what legal responsibility they have for their users' speech. As a matter of black-letter law, they are typically separate from underlying substantive legal doctrines that define things like defamation or hate speech. Yet by prescribing when and how platforms must take action, intermediary liability laws strongly influence what speech actually gets taken down.

At a high level, intermediary liability laws must balance three, often competing goals. The legal details in national law typically reflect lawmakers' judgment about how best to balance them. One goal is to *prevent harm*. Generally, the better job a law does of incentivizing platforms to take down illegal or otherwise harmful content, the more it will serve this goal. Another, often competing goal is to *protect lawful online speech and information*. A law that requires aggressive policing by platforms may run afoul of this goal, leading platforms to take down lawful and valuable speech in order to avoid legal risk. A third goal is to *promote innovation*. Early intermediary liability laws were conceived in part as means to protect nascent industries. Today, intermediary liability laws may profoundly affect competition between incumbent platforms and start-ups.

The balance of priorities between these three goals is a matter of national values and policy choices; but the question of what specific legal rules will, in practice, serve each goal is in part an empirical one, tied to the real-world practices of platforms responding to the law's requirements and incentives.[2]

Internationally, most intermediary laws share two basic elements. First, platforms are immune from legal claims arising from users' unlawful speech as long as they do not get too involved in developing that speech. National laws diverge as to how "neutral" platforms must be to qualify for immunity and the degree of content moderation they can engage in without being exposed to liability. One relative outlier in this respect is the US Communications Decency

[2] Many laws provide somewhat different rules for different claims (like copyright vs. terrorism) or different kinds of intermediaries (like hosts vs. ISPs). For a review of doctrinal variables, see Keller (2019a).

Act (CDA), which grants platforms unusually broad immunities, even when they become aware of unlawful content, for the express purpose of encouraging them to moderate and weed out "objectionable" content.[3]

Second, most intermediary liability laws give platforms obligations once they "know" about illegal content. In much of the world, platforms that learn about material like defamation or terrorist propaganda on their services must take down that content or face legal consequences. Laws vary substantially, however, in what counts as "knowledge." Under some national rules, platforms can only be legally required to take down users' speech if a court has adjudicated it unlawful.[4] Elsewhere, the law leaves platforms to decide for themselves what speech violates the law.

Within this framework, one important source of variation comes from the procedures that the law provides for platforms taking content down. The US Digital Millennium Copyright Act (DMCA) is one of the most procedurally detailed intermediary liability laws.[5] It spells out formal prerequisites for "notices" from rightsholders, steps for "counter-notice" by accused users, and other details including penalties for bad-faith notices against lawful speech. The Manila Principles, a set of model intermediary liability rules endorsed by civil society groups around the world and supported in the human rights literature, lists additional procedural protections – including public transparency requirements to illuminate errors, bias, or abuse in notice-and-takedown systems.[6]

A rapidly developing intermediary liability policy debate concerns platforms' potential obligations to proactively monitor or police users' speech. Until recently, most countries' laws built on the assumption that platforms could not, realistically, monitor user speech on an ongoing basis and accurately identify illegality. Important laws like the US DMCA and EU eCommerce Directive expressly disclaimed monitoring obligations,[7] making platforms responsible only for unlawful content they became aware of, usually through notice from third parties. Platforms' voluntarily-developed filtering tools have since changed policymakers' expectations, though the exact operation of those tools is poorly understood. One major new law, the EU's Copyright Directive, effectively

[3] 47 U.S.C. § 230.

[4] Marco Civil da Internet, Federal Law no. 12.965 (2014) (Brazil); Copyright Act, Law No. 20.435 (Chile); Corte Suprema de Justicia de la Nacion, 8/10/2014, *Rodriguez, Maria Belen c. Google Inc. / da.os y perjuicios* (Argentina).

[5] 17 U.S.C. § 512.

[6] Manila Principles on Intermediary Liability, www.manilaprinciples.org; David Kaye, the UN Special Rapporteur on Freedom of Expression, has reported that the Manila Principles "establish baseline protection for intermediaries in accordance with freedom of expression standards" (Kaye 2017).

[7] 17 USC 512(m); Council Directive 2000/31, 2000 O.J. (L 178) 1 (EC) ("eCommerce Directive"), Article 15.

requires filtering.[8] Other proposals in areas like terrorist content are pending.[9] Critics ranging from technologists to three UN rapporteurs have raised serious concerns about filters (Cannataci et al. 2018; O'Brian and Malcolm 2018). As a 2019 letter from civil society organizations including European Digital Rights (EDRi), Article 19, and the American Civil Liberties Union (ACLU) put it, filters remain "untested and poorly understood technologies to restrict online expression," with great potential to silence protected expression ranging from parody to human rights reporting, with resulting harm to "democratic values and individual human rights."[10]

Another emerging issue comes from both platforms' and governments' reliance on Community Guidelines instead of law as a basis for removing online content. Platforms' discretionary rules often prohibit legal expression, and until recently it was generally assumed that platforms had extremely wide latitude to do so.[11] National constitutions and human rights laws protect internet users from *state* interference with their legal exercise of speech rights, but platforms are generally free to ban any speech they want; and, because Community Guidelines are privately defined and enforced, platforms' decisions are generally not subject to review by courts.

In recent years, though, governments, particularly in Europe, have increasingly turned to platforms' Community Guidelines as enforcement mechanisms. For example, both the European Commission's Hate Speech Code of Conduct and Disinformation Code of Practice call on platforms to voluntarily prohibit specified content, often in reliance on Community Guidelines.[12] Law enforcement bodies including Europol, for their part, often use Community Guidelines or Terms of Service rather than law as a basis for

[8] Directive 2019/790, 2019 O.J. (L130) 92 (EC)("Copyright Directive"), Article 17.

[9] European Parliament. Legislative resolution of April 17, 2019 on the proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (provisional edition), P8_TA-PROV(2019)0421. www.europarl.europa.eu/doceo/document/TA-8-2019-0421_EN.pdf

[10] Center for Democracy and Technology. Civil Society Letter to the European Parliament on the proposed Regulation on Preventing the Dissemination of Terrorist Content Online, February 2019. https://cdt.org/files/2019/02/Civil-Society-Letter-to-European-Parliament-on-Terrorism-Database.pdf

[11] US courts have consistently upheld platforms' right to take down users' lawful speech. Internationally, a different picture may be emerging. In 2018, first instance courts in both Germany and Brazil upheld user claims in this situation. See Keller (2019b), pp. 12–13. Experts have raised concerns that states may violate human rights obligations when they rely on private Community Guidelines to prohibit online expression and information. See Angelopolous et al. (2016), pp. 50–51; Kuczerawy (2017).

[12] European Commission (2016) (agreement with Microsoft, YouTube, Twitter, and Facebook); European Commission (2018b) (agreement with Facebook, Twitter, and Google). While the Hate Speech Code (European Commission 2016) calls for removal of content posted by users, the Disinformation Code (European Commission 2018b) concerns advertising content and counsels *against* prohibiting "false" content from ordinary users.

asking platforms to take down content (Europol 2016; Chang 2018). Civil liberties organizations have decried these arrangements, saying they replace democratic lawmaking and courts with privatized and unaccountable systems (European Digital Rights 2016; see also Chang 2018).

A Case Study: Notice and Takedown in Everyday Practice

By far the most thorough analysis of intermediary liability compliance operations is "Notice and takedown in everyday practice" (Urban et al. 2016). To produce it, researchers reviewed takedown notices affecting some 4,000 individual webpage URLs in Google's web search and image search products and interviewed platform operators, rightsholders, and other participants in the notice-and-takedown ecosystem. Although focused on copyright (the area for which the richest public dataset is available), its thorough analysis documents trends and issues with close analogs for content removal under other laws. Many of its lessons, particularly those relating to automation and large-scale operations, are relevant to removal under Community Guidelines as well.

One of the study's most important findings is the divergence between the operations and capabilities of mega-platforms like Google and other, more modest internet intermediaries (Urban et al. 2016, pp. 28–29, 73–74). Smaller or more traditional companies generally employed teams of three or fewer people for this function and carried out substantive individual review of each notice (Urban et al. 2016, pp. 29, 36). Many described "opting to take down content even when they are uncertain about the strength of the underlying claim" in order to avoid exposure to liability (Urban et al. 2016, p. 41). They also reported notifiers' "deliberate gaming" of the takedown process, "including to harass competitors, to resolve personal disputes, to silence a critic, or to threaten the [platform]" (Urban et al. 2016, p. 40).

The picture for larger players was very different. The difference began with scale: In contrast to the dozens or hundreds of notices received by smaller operations, Google received more than 108 million removal notices for web search during the study's six-month period (Urban et al. 2016, pp. 29, 77). Both large-scale notifiers and platforms relied heavily on automation. Rightsholders and their outsourced vendors automated the notification process, for example by using search queries to identify lists of URLs (Urban et al. 2016, p. 92). The resulting "robonotices" sometimes included errors, like a request from the musician Usher to take down a film version of Edgar Allan Poe's *Fall of the House of Usher* (Urban et al. 2016, pp. 90–91).

Platforms reported automatically accepting many such automated requests, in particular from "trusted" sources. As a result, some provided no human review at all for the majority of automated notices they received (Urban et al. 2016, p. 29). Some also proactively policed content, using "measures such as ex-ante filtering systems, hash-matching based 'staydown' systems, [and] direct

back-end takedown privileges for trusted rightsholders" (Urban et al. 2016, p. 29).

In their quantitative analysis, Urban and colleagues documented considerable error in DMCA operations. Among notices submitted to Google web search, for example, they found questionable legal claims in 28 percent (Urban et al. 2016, p. 88). Some 4.2 percent – which extrapolates to 4.5 million requests across the six-month dataset – seemed to be simple errors, requesting removal of material that did not relate to the notifier's legal claim (Urban et al. 2016, p. 88). Among notices submitted to Google's image search product, the figure was 38 percent (Urban et al. 2016, pp. 98–99). Fully 70 percent of image search notices were called into serious doubt if calculations included one individual's barrage of improper takedown demands (Urban et al. 2016, pp. 98–99).[13]

Platforms interviewed for the Urban study also reported a low rate of DMCA counter-notices from users challenging erroneous takedowns. Many platforms received no counter-notices at all (Urban et al. 2016, p. 44). This finding is consistent with figures released by the Motion Picture Association of America in 2013, showing a 0.000032 percent rate of counter-notices to DMCA removal requests filed by member companies; only 8 counter-notices were identified for 25,235,151 notified URLs (Boyden 2013). Platform transparency reports, similarly, typically report counter-notice rates of well below 1 percent for copyright claims (Bridy and Keller 2016). Given the widespread documentation of over-removal, these figures suggest that wrongful removals are going unchallenged.

SOURCES OF INFORMATION

This section reviews numerous sources of empirical information, both from platforms and other participants involved in content takedown and from independent researchers. Some, including major platform transparency reports, are published on regular schedules and should be fruitful sources of future information.

Disclosures by Platforms and Other Participants in Content Takedown

Platforms reveal information about content moderation in manifold ways. These include (1) periodic transparency reports; (2) primary source

[13] Any study's assessment of error rates, while particularly essential for policymaking, is also subject to challenge – since it requires researchers, like the platforms themselves, to make judgment calls in legal gray zones. Some critics disputed Urban, Karaganis, and Schofield's (2016) study methodology on these and other grounds. See, e.g., Ford (2017). Urban, Karaganis, and Schofield (2017) and other experts – including one of this chapter's authors and law professor Annemarie Bridy – defended the study's standards for classifying notices. See Urban et al. (2017); Bridy and Keller (2017).

information shared with academic research archives such as Harvard's Lumen Database; (3) notices to affected individuals about takedown decisions; and (4) incidental public statements and other disclosures about specific content issues. Increasingly, (5) governments also require platforms to perform public filings about content moderation, and some also publish data about their own involvement in takedown procedures. (6) Third-party audits also reveal information about takedown practices, as do (7) leaked information from platforms.

Transparency Reports

Many platforms publish periodic transparency reports, which typically disclose aggregate data about requests for content removal. An index of transparency reports maintained by the civil society organization Access Now lists reports from more than seventy companies,[14] including Google,[15] Facebook,[16] Twitter,[17] Amazon,[18] Tumblr,[19] Medium,[20] Reddit,[21] Github,[22] and WordPress.[23] These can provide important quantitative overviews of the big picture – or at least part of it. They typically aggregate data about removal requests, along with the platform's rate of compliance. They may also disclose the frequency with which users accused of wrongdoing choose to appeal or challenge platforms' decisions. Transparency reports have historically focused on legal removal requests. In 2018, however, Facebook,[24] Twitter,[25] and YouTube[26] all published their first Community Guidelines enforcement reports.

Transparency reports have major limitations. The aggregated data in transparency reports only shows the platforms' own assessments, and not the merits of the underlying cases. That means researchers cannot evaluate the accuracy of takedown decisions or spot any trends of inconsistent enforcement. Also, most transparency reports only cover particular categories of takedowns – often only those initiated by governments or copyright-holders. This leaves open questions about platforms' responses to legal allegations brought by individuals under, say, French defamation law or Brazilian privacy law.

Transparency reports also vary widely in the ways they classify data, making apples-to-apples comparisons between companies difficult. In particular, reports that track how many *notices* a company received cannot fruitfully be compared to reports tracking how many *items of content* they were asked to remove, since one notice may list any number of items.

Transparency reports also vary greatly in detail. Take, for instance, the aforementioned Community Guideline reports. YouTube's report documents the number of channels and videos removed for eleven different types of standards violations (e.g., spam, nudity, promotion of violence and extremism)

[14] Access Now (2016). [15] Google (n.d.). [16] Facebook (2018a). [17] Twitter (2017).
[18] Amazon (2015). [19] Tumblr (n.d.). [20] Medium (2015). [21] Reddit, Inc. (2015).
[22] GitHub (2015). [23] Automattic (n.d.). [24] Facebook (2018a). [25] Twitter (2018).
[26] Google (2018a).

(Google 2018a). It also specifies how these videos were detected, whether through automated flagging, individual trusted flaggers, users, NGOs, or government agencies. Facebook's report is even more detailed; it also registers how often users appealed against removal decisions and how often content was later restored (either proactively by Facebook or following a user appeal) (Facebook 2018b). In addition to this numerical reporting, Facebook provides details about operations. Its expected staff of more than 20,000 people are working on content moderation, including native speakers of more than 50 languages and teams working around the clock. Separately, Facebook has published a detailed public version of its Community Standards (Facebook 2018b) and a guide to understanding the figures from the report (Facebook 2018). Twitter's report, on the other hand, is significantly less detailed. It only documents the number of unique accounts reported and actioned for six different categories of violations, without specifying appeal or reinstatement rates or reporting mechanisms other than those from known government entities (Twitter 2018).

One important external assessment of company transparency reports' strengths and weaknesses can be found in the Electronic Frontier Foundation's periodic *Who Has Your Back* report (Gebhart 2018). Another can be found in the Ranking Digital Rights Corporate Accountability Index, which rates technology companies on numerous measures relating to transparency and protection of users' free expression and privacy rights.[27] The Open Technology Institute's Transparency Reporting Toolkit also provides a valuable comparison of existing reports and recommendations for improved practices (Open Technology Institute 2018). It draws on the widely endorsed Santa Clara Principles, which call for "numbers, notice, and appeal" as essential elements in platforms' content-removal operations.[28]

Primary Source Information Shared by Platforms

For researchers to draw their own conclusions about platforms' content-removal practices, they need to know what content was actually removed. To date, the best source for such information has been the Lumen database ("Lumen"), an archive hosted at Harvard's Berkman Klein Center.

Lumen archives legal takedown notices from any platform – or sender – that chooses to share them. Senders' personal information, and occasionally – as in the case of child abuse content – content location URLs are redacted, but researchers can otherwise review the entire communication. At last check, the database held some 9.3 million notices, targeting approximately 3.35 billion URLs.[29] The majority comes from Google; other contributors include Twitter, WordPress, the Internet Archive, Kickstarter, Reddit, and Vimeo. Because the

[27] The Ranking Digital Rights 2018 Corporate Accountability Index. https://rankingdigitalrights.org/index2018/

[28] The Santa Clara Principles. https://santaclaraprinciples.org/

[29] Email correspondence with Adam Holland, Lumen Project Manager.

notices identify material that is often still available at the listed URL, researchers can look at the specific content alleged to be illegal and assess whether a platform made the right decision. The Lumen database has enabled extensive academic research.[30]

Demands for other forms of "primary source" transparency are increasing. An important recent proposal from the French government, for example, calls for transparency sufficient for auditors to review specific takedown decisions.[31] Similarly, proposed amendments to the German Network Enforcement Law (*Netzwerkdurschsetzungsgesetz*, or NetzDG) – discussed in further detail in the section titled "Public Filings and Other Government Disclosures" – also call for the auditing of content moderation practices and the creation of a public "clearing house" to adjudicate user complaints about wrongful removals (German Parliament 2018). Such independent oversight mechanisms, if they proved operationally and economically feasible, might also allow more detailed third-party research into the substance of content-moderation decisions.

Notice to Affected Individuals

Platforms also provide potentially useful information to individuals affected by takedown requests. In particular, they may (1) respond to a person who requested removal, letting them know if the request was honored; (2) notify the user whose content was taken down; or (3) "tombstone" missing material, putting up a notice for users who are trying to visit a missing page or find information. YouTube's "this video is not available" notices are perhaps the most visually familiar example for many internet users.

Information gleaned from notices in individual cases often drives news cycles about particularly controversial decisions. It cannot show researchers the big picture, but it can play an important role in surfacing errors, by putting information in the hands of the people most likely to care and take action.

Issue-Specific Platform Disclosures

For high-profile content-moderation issues, platforms are increasingly issuing in-depth public statements to explain their policies. Some offer detailed

[30] This includes Urban, Karaganis, and Schofield (2017) and numerous other scholarly works including those cited in Section III.B.1, below. See also Brief of Amici Curiae Chilling Effects Clearinghouse Leaders in Support of Appellee at *8–16, Perfect 10, Inc. v. Google, Inc., 653 F.3d 976 (9th Cir. 2011), 2010 WL 5813411, www.eff.org/document/amicus-brief-chilling-effects (citing other works). Equivalent transparency for removal of hosted content, while highly valuable, would be more difficult. It would require either the host or Lumen to actively preserve allegedly illegal content, in the face of a removal request. In many cases, as with privately shared Facebook posts, disclosing the content could also conflict with privacy obligations to users.

[31] See French Secretary of State for Digital Affairs (2019), p. 20, which calls for an "independent and extra-judicial mechanism for reviewing the platform's decision."

information on platforms' assessment of individual cases, which is typically lacking from aggregate transparency reports.

An important example is the terrorist attack on Christchurch of March 15, 2019. Video footage of the attack was livestreamed on Facebook and spread virally to several other websites. Facebook ultimately issued two public announcements describing their efforts to remove this graphic footage. They provide detailed timelines of events, starting with the first livestream, as well as data on how often it was subsequently viewed, shared, re-uploaded, and ultimately removed.[32] In its response to the Christchurch incident, Facebook also took the unprecedented step of inviting a legal academic, Kate Klonick, to sit with response teams. Klonick later published her observations (Klonick 2019).

Facebook also issued several statements regarding its efforts to remove "coordinated inauthentic behavior," to protect elections in, for example, Ukraine,[33] Israel,[34] India, and Pakistan[35] and targeting fake accounts originating from Russia[36] and Iran.[37] These posts explain how the platform identifies inauthentic behavior, what patterns it has found, and what removal decisions it made including examples as well as aggregate data.

Cloudflare, a web infrastructure company, published a particularly influential blog post about content moderation in the aftermath of the Charlottesville riots of August 11, 2018. It explained why the company had decided to terminate its services to The Daily Stormer, a white supremacist website. The author, CEO Matthew Prince, was remarkably self-critical and highlighted the "the risks of a company like Cloudflare getting into content policing" (Cloudflare 2017).

Perhaps the most in-depth example of issue-specific reporting is Google's report on *Three Years of the Right to Be Forgotten* (Google 2018b). This document is unique in the degree of detail it provides about the company's internal process in assessing individual removal requests. It provides anonymized examples of individual cases, such as one request to remove search results for "an interview [the notifier] conducted after surviving a terrorist attack" and another for "a news article about [the notifier's] acquittal for domestic violence on the grounds that no medical report was presented to the judge confirming the victim's injuries" (Google 2018b, p. 10; Google took down both). The report lists several specific factors and classifications Google uses to resolve requests. One factor, for example, is the identity of the "requesting entity." Google classifies the requesting entity for each item of disputed online content using the six categories in Table 10.1 (Google

[32] "The first user report on the original video came in 29 minutes after the video started, and 12 minutes after the live broadcast ended. In the first 24 hours, we removed more than 1.2 million videos of the attack at upload . . . Approximately 300,000 additional copies were removed after they were posted" (Facebook 2019a).

[33] Facebook (2019c). [34] Facebook (2019d). [35] Facebook (2019e). [36] Facebook (2019f). [37] Facebook (2019g).

TABLE 10.1 *Breakdown of all requested URLs after January 2016 by the categories of requesting entities*

Requesting entity	Requested URLs	Breakdown	Delisting rate
Private individual	858,852	84.5%	44.7%
Minor	55,140	5.4%	78.0%
Nongovernmental public figure	41,213	4.1%	35.5%
Government official or politician	33,937	3.3%	11.7%
Corporate entity	22,739	2.2%	0.0%
Deceased person	4,402	0.4%	27.2%

Note. Private individuals make up the bulk of requests.

2018b, p. 5). Based on these granular criteria, Google generates aggregate numbers and statistical analysis.

The report is also valuable because it illustrates concretely how a platform might break down complex claims into standardized elements or checkboxes for rapid, large-scale processing. Independent researchers could review these elements to assess, for example, how adequate they seem as an alternative to judicial review of parties' competing privacy and free expression rights. They could also use the reported factors and elements as a concrete, debatable starting point in discussing what information platform employees should reasonably track and report about each takedown decision.

Public Filings and Other Government Disclosures

Valuable information about platform operations sometimes surfaces to the public through court or other public filings.[38] Documents made public in the *Viacom v. YouTube* case, for example, made headlines for their revelations about both parties to the suit (Anderson 2010; YouTube 2010). Information disclosed in response to consultations by governments or transnational bodies has appeared in publications including a 2012 European Commission staff report (European Commission 2012) and reports from the office of the UN Free Expression Rapporteur (Kaye 2017).

More recently, large platforms including Facebook, YouTube, and Twitter have published important reports as part of their compliance with Germany's NetzDG law.[39] That law is best known for its unusually strict content-removal

[38] See, e.g., European Commission (2018a); US Copyright Office (2015); US Patent and Trademark Office (2015); Torrent Freak (2018) (citing testimony of Google legal director disclosing use of hash matching on Google Drive).

[39] Netzwerkdurchsetzungsgesetz vom 1. September 2017 (BGBl. I S. 3352). ("Network Enforcement Law" or "NetzDG"). www.gesetze-im-internet.de/netzdg/BJNR335210017.html; Google (2018c); Twitter (2018b); Facebook (2018b).

rules, but it also imposes unprecedented public reporting requirements. Platforms' biannual reports include information such as staffing numbers, wellness resources available to staff, operational processes, the number of consultations with external legal counsel, and turnaround time for responding to notices, broken down by the specific legal violation alleged. While researchers have no independent means of assessing accuracy of the platforms' legal determinations, the reports are rich in other statistics and operational detail.

For example, in the second half of 2018, YouTube received NetzDG notices identifying more than 250,000 items. The most reported category was Hate Speech or Political Extremism (83,000 plus complaints), followed by Defamation or Insults (51,000 plus) and Sexual Content (36,000 plus). In response, YouTube removed 54,644 items, with takedown rates varying per content category (Google 2018c). The report also shows whether notices were submitted by users vs. German government agencies.

As Figure 10.1 from the report shows, YouTube also relied heavily on Community Guidelines. Nonetheless, it looked to German law to resolve the legal status of more than 10,000 items.

Facebook's NetzDG reports paint a very different picture. In the same period, they reportedly received only 500 NetzDG complaints, involving 1,048 items of content – only a fraction of what YouTube and other major platforms received (Facebook 2018b). This is likely because their NetzDG

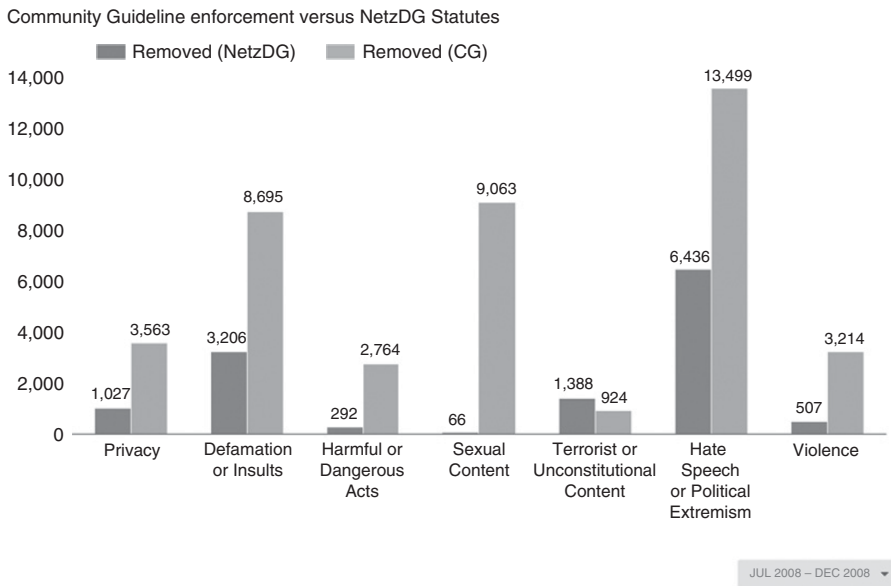

FIGURE 10.1 Community guideline enforcement versus NetzDG statutes
Source. Google (2018c)

complaint form was less visible on their website compared to YouTube or Twitter. Yet Facebook's Community Guideline removals in the same period numbered in the millions and were not included in the NetzDG report. On July 2, 2019, Germany's Federal Office of Justice fined Facebook €2 million for incomplete reporting, claiming that, because it was unclear which complaints based on Community Guidelines had in fact identified unlawful material, "the number of received complaints about unlawful content is incomplete," and the reports therefore created a "distorted picture" (Bundesamt für Justiz 2019). This raises the question whether NetzDG also requires platforms to assess and report on the lawfulness of content decisions that are not referred to them under the NetzDG framework.

The European Commission also persuaded Twitter, Google, and Facebook to publish monthly compliance reports in the run-up to the EU elections of May 2019, as part of the Code of Practice on Disinformation (European Commission 2018b). These reports describe a range of activity related to disinformation, including media literacy and fact-checking efforts but also certain forms of content moderation. Importantly, these are some of the few reports to discuss the enforcement of advertising standards, including procedural information on the approval, review, and transparency mechanisms for political advertising as well as quantitative data on advertising activity and removal decisions. In May 2019, for instance, Google detected 16,690 EU-based Google accounts in violation of its misrepresentation policies, and Twitter removed 1,418 ads in violation of their Unacceptable Business Practices Policy (which prohibits, e.g., misleading content) (Google 2019; Twitter 2019). Finally, the US House Intelligence Committee has also published important datasets, obtained from Facebook, about Russian political advertising during the 2016 US presidential elections (US House of Representatives 2019).

Platforms also disclose takedown information to the European Commission for inclusion in the Commission's reporting on the Hate Speech Code of Conduct (European Commission 2016, 2017). Under the Code of Conduct, expert organizations notify participating platforms about content the organizations have identified as illegal hate speech. Platforms complied with 28 percent of such notices in the Commission's first review and 59 percent in its second – a development billed by the Commission as "important progress." Yet the figures represent only notifiers' and platforms' rate of agreement about what content should come down; and, while the Code of Conduct refers to "illegal" hate speech, platforms (and perhaps also notifiers) presumably actually assess notices under their Community Guidelines. Without independent access to notices and affected content, we cannot know what standards either side is applying, how consistently and accurately those standards are enforced, or how they relate to any country's laws.[40]

[40] The Code covers "illegal hate speech," and the Commission says its reports quantify platforms' compliance with "notifications concerning illegal hate speech." In practice, platforms

Some related but more modest disclosures come from governments themselves, sometimes in conjunction with platforms. A 2016 report from Europol, for example, discusses terrorist content referred by its Internet Referral Unit (IRU) to platforms for takedown (Europol 2016). As of then, the IRU had referred 9,787 items to 70 different platforms. Its success rate was greater than 90 percent. Because IRUs seek removal under platforms' Community Guidelines, however, this figure reflects only success in predicting platforms' applications of those rules – or convincing platforms to adopt law enforcement agents' interpretation. Independent researchers have no means of ascertaining what portion of referred or removed content violates any laws.

Some European states also operate their own IRUs at the national level. Their operations have been criticized for a lack of transparency, but occasional disclosures have occurred. The UK's Counter-Terrorist Information Referral Unit (CTIRU) published data on its website on December 2016 claiming that it was instigating the removal of more than 2,000 pieces of content per week and was on course to have removed more than 250,000 pieces of content in total by the end of the year (UK Metropolitan Police 2016). The CTIRU has also published information in Parliamentary Hearings[41] and in response to Freedom of Information requests submitted by University College Dublin law professor T. J. McIntyre (2018).

More government reporting about their involvement in content moderation may one day be required by law. The EU Parliament's draft of the Terrorist Content Regulation, for example, includes detailed requirements for transparency about law enforcement referrals to platforms.[42]

Audits

Published reports from independent auditors represent a small but likely growing category of disclosure. The Global Network Initiative has published reports of privacy and content-removal practices going back to 2013 for companies including Facebook, Google, LinkedIn, Microsoft, and Yahoo, for example. The reports, which draw on internal but nonprivileged information shared by the companies, include general assessments and case studies.[43] Other

presumably also honor notifications – and Commission counts them as "successful" – for content that does *not* violate the law but does violate the platforms' Community Guidelines. The increase from 28 percent compliance to 59 percent could mean that notifiers got better at predicting platforms' internal rules.

[41] Hansard. 2017. HL Deb 787 Col. 1261. http://bit.ly/2kctmPL

[42] European Parliament. Legislative resolution of 17 April 2019 on the proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (provisional edition), P8_TA-PROV(2019)0421. www.europarl.europa.eu/doceo/document/TA-8-2019-0421_EN.pdf

[43] Global Network Initiative, Company Assessments. https://globalnetworkinitiative.org/company-assessments/

one-off independent audits have become a common response to tech industry scandals and may produce relevant information going forward.[44]

A Human Rights Audit of the Internet Watch Foundation ("*IWF Audit*") report provides insights into an important non-platform participant in content takedowns (MacDonald 2014).[45] The UK's Internet Watch Foundation (IWF) is a private, nonprofit organization that works with police, companies, and the public to identify child sexual abuse material online. It then conveys lists of URLs to intermediaries to be blocked. The *IWF Audit*, prepared at IWF's request by an outside human rights expert, details the group's internal operations and suggests improved processes for, among other things, appeals and difficult legal judgment calls. This report is unique among empirical research – and important to developments well beyond child protection – in its focus on the interplay of state and private action. For example, it discusses the role that IWF as a private organization plays in speeding takedown requests initiated by police – requests that would otherwise require additional judicial process (MacDonald 2014, p. 5).

Leaked Information

In addition to their publicly available Community Guidelines, platforms also issue more detailed rules and instructions for their content-moderation staff. These documents are confidential, but they have been leaked to the press on several occasions. They shed some light on the way that platforms' general principles are enforced in practice; in order to instruct their moderators at scale, platforms are often forced to reduce complex speech issues to simplified rules of thumb.

For instance, Facebook instruction manuals leaked to *The Guardian* told content moderators that the phrase "Someone should shoot Trump" was a credible threat of violence, whereas "Let's beat up fat kids" was not (Hopkins 2017). In documents leaked to the *Süddeutsche Zeitung*, Facebook instructed moderators to treat people as public figures, with fewer privacy protections, as long as they "were mentioned in news reports five times or more in the past two years" (Krause and Grasegger 2016). Gawker's 2012 leaks alleged that Facebook contractors were instructed to "escalate" to Facebook employees any "maps of Kurdistan," "burning Turkish flag(s)," and "All attacks on Ataturk (visual and text)" – suggesting that Facebook made concessions to public pressure from Turkey (Chen 2012). More recently, the *New York Times* has also published similar documents (Fisher 2018).

[44] For instance, Facebook recently responded to accusations of bias with new auditing measures, in which "one adviser will conduct an audit of Facebook's impact on minority communities and communities of color, while another will advise the company on the potential bias against conservative perspectives" (Ong 2018).

[45] An additional valuable resource documenting content removal mechanisms by IWF and other UK entities is McIntyre (2018).

Independent Research

A growing and important body of information about platforms' takedown practices comes from outside researchers. Some (1) analyze data released by platforms, while others (2) perform surveys and interviews with platform staff and other participants in the takedown ecosystem.[46] (3) Others have run their own tests and experiments with platform services and associated software.

Analysis of Data Disclosed by Platforms

As mentioned in the section on "Primary Source Information Shared by Platforms," researchers using the Lumen database can do an important thing most others cannot: review the content that platforms actually removed. A handful of Lumen-based reports, like *Notice and Takedown*, take this approach (Urban et al. 2016). Most concern copyright, since the bulk of the data Lumen holds relates to the US DMCA.[47]

One recent exception came from law professor Eugene Volokh, who discovered that numerous claimants had falsified court orders and used them to convince Google to take content out of its search results (Volokh and Levy 2016). Volokh's detective work, which involved hiring an actual detective, turned on clues like the dates and official stamp numbers that appeared on the putative court documents – details that were only available because Lumen gives researchers access to exact copies.

Reporting and Interviews with Participants in Takedown Processes

Other researchers have carried out the painstaking work of tracking global developments and seeking out and interviewing individual participants. Rebecca McKinnon laid important groundwork for this in her 2012 book

[46] Platforms have recently taken a few steps toward greater disclosures to independent researchers. Facebook, for example, committed to the Social Science One data-sharing project discussed in the main text; and Twitter agreed to experiment with new means of surfacing rules to users in a joint project with researchers (Benesch and Matias 2018). These are important developments, but it is not clear if either will lead to new public information specifically on the topic of content removal.

[47] Heins and Beckles (2005) found 47 percent of notices stated weak claims or involved speech with important legal defenses; Urban and Quilter (2006), pp. 621, 641, noted 55 percent of notices involved disputes between commercial competitors and 31 percent presented significant legal questions; Bar-Ziv and Elkin-Koren (2017) found a few notifiers accounted for a high 66 percent rate of abusive requests on Israel's .il domain, but the remainder of requests had a rate perhaps as low as 5 percent invalid or questionable requests; Seng (2015), p. 1, found that 8.3 percent of notices had "formal" errors and 1.3 percent had "substantive" errors. A member survey by the Copyright Alliance, a copyright-holder advocacy organization, provided a rare quantification of notifiers' experiences. Of 219 respondents to a survey, 62 percent reported filing DMCA notices that received no response at all from platforms. Examples in the report suggested many may have been smaller companies (see Copyright Alliance 2016).

Consent of the Networked (MacKinnon 2012). The *Notice and Takedown* report, discussed in the section titled "A Case Study: Notice and Takedown in Everyday Practice," builds on interviews and extensively documents the self-reported behaviors of copyright owners, platforms, and other players in the notice-and-takedown ecosystem (Urban et al. 2016, pp. 10–13, 116–117).

Noteworthy contributions relating to Community Guidelines enforcement include Kate Klonick's article "The new governors: The people, rules, and processes governing online speech," for which the author reviewed public reporting to date on the topic and interviewed early platform content moderators to understand the growth of their rules. Facebook, for example, started with vague rules, remembered by one employee as "Feel bad? Take it down" (Klonick 2018). Tarleton Gillespie explored content moderation at length in a 2018 book (Gillespie 2018). Other researchers have reported on platforms' publicly documented removal rules (Venturini et al. 2016; York et al. 2018).

Reporters like Julia Angwin have experimented with platform content toleration, documenting things like anti-Semitic ad targeting terms on Facebook (Angwin, Varner, and Tobin 2017). Academics and civil society advocates affiliated with Onlinecensorship.org have used crowdsourcing to gather and quantify users' reports of experiences with platform takedowns (Anderson, Carlson et al. 2016; Anderson, Stender et al. 2016), including apparent disparate impact on vulnerable and minority groups (York and Gullo 2018). Documentarians have produced at least two films about the on-the-ground experience of individual frontline content moderators working in places like India or the Philippines for vendors under contract with US-based platforms.[48] Daniel Kreiss and Shannon McGregor have performed in-depth interviews with Facebook and Google's advertising staff, in order to study how company standards around political advertising are developed and enforced (Kreiss and McGregor 2019); and academics including Sarah Roberts have analyzed more closely the role – and vulnerability – of this global workforce (Roberts 2019).

Academics also engage with platform employees at conferences and other discussion events, such as the Content Moderation at Scale (or "COMO") series initiated by Professor Eric Goldman of Santa Clara University.[49]

[48] *The Cleaners* (2018), directed by Hans Block and Moritz Riesewieck (Berlin: Gebrueder Beetz Filmproduktion); *Field of Vision – The Moderators* (2017), directed by Ciaran Cassidy and Adrian Chen (New York: First Look Media), YouTube video, April 14. www.youtube.com/watch?v=k9moaxUDpro

[49] Santa Clara University School of Law, 2018. Content Moderation & Removal at Scale conference, Santa Clara, CA, February 2. https://law.scu.edu/event/content-moderation-removal-at-scale. In one widely reported session, Emma Llansó of the Center for Democracy and Technology and Mike Masnick of the blog Techdirt invited the audience to "make the call" on a series of vexing takedown decisions – and found, even within the relative cultural homogeneity of a Washington, DC audience, no consensus (Goldman 2018).

Independent Trials and Experiments

Because platforms' content-removal decisions are taken behind closed doors, some researchers have been creative in nosing out useful information. European researchers in the early 2000s, for example, experimented with posting famous out-of-copyright literature – John Stuart Mill's *On Liberty* in one case (Ahlert, Marsden, and Yung 2004) and an essay by the nineteenth-century Dutch satirist Multatuli in another (Leyden 2004) – and then submitting copyright infringement notices to see if intermediaries would take them down. Most intermediaries complied. More recently, the researcher Rishabh Dara sent a wide array of content-removal requests to intermediaries in India and tracked their responses in detail (Dara 2011). There, too, platforms generally erred on the side of caution. Dara's research is relatively unique in its focus on non-copyright claims. For example, by invoking a law against advocacy of gambling, he caused a news site to take down user comments concerning a proposed change in Indian gambling law (2011, p. 15).[50] University of Haifa researchers Mayaan Perel and Niva Elkin-Koren did similar research in Israel to assess the use of algorithms in copyright takedown processes – a method they call "black box tinkering" (Perel and Elkin-Koren 2017).

Independent research may be particularly relevant for the study of algorithmic content filtering systems, which have become increasingly central to legal debates and large platforms' moderation practices. Reliance on these technologies concerns civil rights activists, since they perform poorly in decisions that require nuanced assessments of context.[51] Distinguishing terrorist propaganda from journalist commentary on terrorism, for instance, or distinguishing content piracy from parody or other fair uses, is difficult to automate.

In a 2018 report, the Center for Democracy and Technology reviewed commercially available text-based filters and found an accuracy rate in the 70–80 percent range (Center for Democracy and Technology 2017). Filters performed particularly poorly in assessing jokes or sarcasm or in languages not spoken by their developers (Center for Democracy and Technology 2017, pp. 14, 19). A 2017 report by Princeton Computer Science professor Nick Feamster and Evan Engstrom of the start-up–advocacy group Engine provides greater technical detail, analyzing one of the few open-source (and hence publicly reviewable) filtering tools, Echoprint (Engstrom and Feamster 2017). The authors found a 1–2 percent error rate in simple duplicate matching, including both false positive and false negatives.

[50] This research might be difficult to replicate, given ethics concerns about targeting the lawful speech of real-world internet users for unjustified removal.

[51] Article 1. 2018. Joint Letter on European Commission regulation on online terrorist content. www.article19.org/resources/joint-letter-on-european-commission-regulation-on-online-terrorist-content/; Reda (2017).

Under a broader view of content moderation, platforms also shape discourse through the design of their ranking and recommender algorithms, such as Facebook's News Feed and YouTube's Recommended videos (Keller 2019b). A growing body of literature in computer science and communications science seeks to ascertain the operation and effects of these complex systems.[52] The design of these algorithms is currently unregulated, but several governments have recently proposed to do so.[53] Most of these initiatives also explicitly demand greater transparency in algorithmic recommendations.[54]

CONSEQUENCES OF PLATFORM CONTENT REMOVAL

Most empirical research on platform content takedowns focuses on removal decisions themselves. Research on more complex questions about how removals affect individual users or society at large is generally harder to come by.

One possible exception is the growing body of research on online influence and the distortion of democratic political processes. Areas of empirical inquiry include "fake news," Russian electoral interference, bot-based message amplification, and political bias in platforms' content-moderation policies. Current and likely future work in this area is comparatively robust and is discussed throughout this volume. A promising source for future research is Facebook's Social Science One project with the Social Science Research Council, which will provide some access to anonymized user data for independent research on "the effects of social media on democracy and elections."[55]

Another relevant issue is the charge, increasingly raised in the United States, Germany, and elsewhere, that major California-based platforms are biased against political conservatives. Individual takedown decisions often drive news coverage or social media concern about this possibility. To meaningfully assess the claim, however, researchers would need far more information about overall takedown patterns. Even with that data, researchers may continue to disagree on what qualifies as legitimate political speech and which speakers fall into the category of "conservatives." For example, commentators have

[52] E.g. see Hargreaves et al. (2018) on documenting patterns in Facebook content recommendations for Italian news media, based on observational data from dummy accounts; see Cornia et al. (2018) on surveying the effects of changes to the Facebook news feed on various news organizations.

[53] Helberger, Leerssen, and van Drunen (2019). See also French Secretary of State for Digital Affairs (2019), p. 3, which proposes an "[o]bligation of transparency of the function of ordering content" and a "duty of care towards [platforms'] users"; The European Commission (2018b), in its Code of Practice on Disinformation, requires platforms to "[d]ilute the visibility of disinformation by improving the findability of trustworthy content."

[54] European Commission (2018b). See also: Regulation 2019/1150 (EU).

[55] Social Science One. https://socialscience.one

disagreed on the appropriate classification of the American Nazi Party (Hanania 2019; Graves 2019).

Beyond election-related topics, empirical research on the broader impact of platform takedown decisions is rare. One particularly pressing question concerns the connection between online speech and offline violence. Observers around the world have pointed to social media as a causal factor in violence in areas from Myanmar to Libya (Walsh and Zway 2018; McLaughlin 2018). A 2018 study from Germany, claiming to quantify Facebook's impact on physical assaults against immigrants, drew both headlines and condemnation of its methodology (Taub and Fisher 2018; Masnick 2018).

Research on terrorism, radicalization, and recruitment is comparatively advanced, but experts are divided on the true role of online materials. A 2017 review of literature to date, for example, cited divergent opinions but some movement toward "consensus that the internet alone is not generally a cause of radicalisation, but can act as a facilitator and catalyser of an individual's trajectory towards violent political acts" (Meleagrou-Hitchens and Kaderbhai 2017, pp. 19, 39; Keller 2018).

Other researchers have cited data suggesting that open platforms, which permit public visibility and counter-speech, may be less conducive to real-world violence than more isolated internet echo chambers (Benesch 2014; Munger 2017).[56] A related empirical question concerns online speech and public participation by members of vulnerable or minority groups. Many thinkers express concern, for example, that toleration for lawful but offensive or threatening speech on platforms like Twitter effectively diminishes the public presence of ethnic minorities, women, and other frequently attacked groups (West 2017). Civil rights organizations have also charged platforms with disproportionately silencing members of minority groups.[57] Questions about disparate impact or bias in takedown operations are all but impossible to truly answer, however, in the absence of representative datasets revealing individual content-removal decisions.

Other consequences of platform takedown operations may affect any user. Individuals who are locked out of their accounts with major platforms like Facebook or Google, for example, may find themselves unable to access other online services that depend on the same login information. Those who depend on hosting services to maintain their writing or art may find their own sole

[56] Research by Susan Benesch, for instance, indicated that speech believed to be correlated to violence during Kenyan election was overrepresented in closed Facebook discussion, compared to public exchanges on Twitter (Benesch 2014).

[57] In 2017, for example, seventy civil rights and social justice organizations wrote to Facebook to complain of bias in its content-removal decisions (Levin 2017). In 2018, YouTube faced public outcry from LGBTQ users who said their videos were unfairly penalized (The Guardian 2017); see also Duguay, Burgess, and Suzor (2018). One author of this chapter has argued elsewhere that platforms' overzealous efforts to counter Islamist extremism can be expected to disproportionately harm users speaking Arabic or talking about Islam (Keller 2018, pp. 20–26).

copies deleted (Macdonald 2016); and several studies suggest that internet users who believe their speech is being monitored curtail their writing and research (Marthews and Tucker 2017, Pen America 2013, Penney 2016).

EMPIRICAL QUESTIONS ABOUT PLATFORM CONTENT TAKEDOWNS

The empirical research summarized in this chapter answers some important questions about platform content takedowns and illuminates others. Key considerations that should inform policy decisions are listed here. Current and future research addressing these questions will improve both our understanding and public decision-making on questions involving platforms and online speech.

- Accuracy rates in identifying prohibited material
  - In notices from third parties generally
  - In notices from expert or "trusted" third parties
  - In flags generated by automated tools
  - In platform decision-making
- Areas of higher or lower accuracy
  - For different claims (such as defamation or copyright)
  - For different kinds of content (such as images vs. text; English language vs. Hindi; news articles vs. poems)
  - For different kinds of notifiers (such as "trusted experts")
- Success rates of mechanisms designed to prevent over-removal
  - Legal obligations or penalties for notifiers
  - Legal obligations or penalties for platforms
  - Counter-notice by users accused of posting unlawful content
  - Audits by platforms
  - Audits by third parties
  - Public transparency
- Costs
  - Economic or other costs to platforms
  - Economic or other costs to third parties when platforms under-remove (prohibited content persists on platforms)
  - Economic or other costs to third parties when platforms over-remove (when platforms take down lawful or permitted content)
- Filters
  - Accuracy in identifying duplicates
  - Accuracy in classifying never-before-seen content
  - Ability to discern or assess when the same item of content appears in a new context (such as news reporting)
  - Relative accuracy for different kinds of prohibited content (such as nudity vs. support of terrorism)

  - Relative accuracy for different kinds of files or media (such as text vs. MP3)
  - Effectiveness of human review by platform employees to correct filtering errors
  - Cost, including implementation and maintenance costs for platforms that license third-party filtering technology
  - Impact on subsequent technical development (such as locking in particular technical designs)
- Community Guidelines
  - Rules enforced
  - Processes, including appeal
  - Accuracy and cost of enforcement
  - Governments' role in setting Community Guidelines
  - Governments' role in specific content-removal decisions
- Consequences of removal, over-removal, and under-removal
  - Public information and discourse, including trust in media
  - Electoral outcomes
  - Violence
  - Commercial interests of notifiers
  - Commercial interests of businesses impacted by removals
  - Disparate impact based on race, gender, etc.

CONCLUSION

Public understanding of platforms' content-removal operations, even among specialized researchers, has long been limited. This information vacuum leaves policymakers poorly equipped to respond to concerns about platforms, online speech, and democracy. A growing body of independent research and company disclosures, however, is beginning to remedy the situation. Through improved public transparency by platforms, and thoughtful inquiry and evaluation by independent experts, we may move toward new insights and sounder public policy decisions.

REFERENCES

Access Now. (2016). *Transparency Reporting Index*. Access Now report. www.accessnow.org/transparency-reporting-index/

Ahlert, C., Marsden, C., & Yung, C. (2004). How "liberty" disappeared from cyberspace: The mystery shopper tests internet content self-regulation. Programme in Comparative Media Law & Policy at the Oxford Centre for Socio-Legal Studies research paper.

Amazon. (2015). *Amazon Information Request Report*. Amazon report. http://do.awsstatic.com/certifications/Transparency_Report.pdf

Anderson, J., Carlson, K., Stender, M., West, S. M., & York, J. C. (2016). *Censorship in Context: Insights from Crowdsourced Data on Social Media Censorship*. Onlinecensorship.org report. https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-second-report-censorship-in-context-pdf

Anderson, J., Stender, M., West, S. M., & York, J. C. (2016). *Unfriending Censorship: Insights from Four Months of Crowdsourced Data on Social Media Censorship*. Onlinecensorship.org report. https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-first-report-download

Anderson, N. (2010). Smoking guns, dark secrets aplenty in YouTube-Viacom filings. *Ars Technica*, March 18. https://arstechnica.com/tech-policy/2010/03/smoking-guns-dark-secrets-spilled-in-youtube-viacom-filings/

Angelopolous, C., Brody, A., Hins, A. W. et al. (2016). *Study of Fundamental Rights Limitations for Online Enforcement Through Self-Regulation*. Report. https://perma.cc/8QAW-79QT

Angwin, J., Varner, M., & Tobin, A., (2017). Facebook enabled advertisers to reach "Jew haters." *ProPublica*, September 14. www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

Automattic. (n.d.). *Transparency Report*. Automattic report. https://transparency.automattic.com

Bar-Ziv, S., & Elkin-Koren, N. (2017). Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown. *Connecticut Law Review*, *50*(2), 1–46.

Benesch, S. (2014). Countering dangerous speech to prevent mass violence during Kenya's 2013 elections. Dangerous Speech Project website, February 9. https://dangerousspeech.org/countering-dangerous-speech-kenya-2013

Benesch, S., & Matias, J. N. (2018). Launching today: New collaborative study to diminish abuse on Twitter. *Medium*, April 6. https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc

Boyden, B. (2013). *The Failure of the DMCA Notice and Takedown System*. George Mason University Center for the Protection of Intellectual Property. https://sls.gmu.edu/cpip/wp-content/uploads/sites/31/2013/08/Bruce-Boyden-The-Failure-of-the-DMCA-Notice-and-Takedown-System1.pdf

Bridy, A., & Keller, D. (2016). *U.S. Copyright Office Section 512 Study: Comments in Response to Notice of Inquiry*, March 31.

(2017). U.S. Copyright Office Section 512 study: Comments in response to second notice of inquiry. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920871

Bundesamt fuer Justiz. (2019). Bundesamt für Justiz erlässt Bußgeldbescheid gegen Facebook. Bundesjustizamt.de. www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702.html

Cannataci, J., Kaye, D., & Ní Aoláin, F. (2018). Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; the right to privacy; and promotion and protection of human rights and fundamental freedoms while countering terrorism. United Nations Special Rapporteur communication OL OTH 71/2018, December 7. https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234

Center for Democracy and Technology. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy and Technology report. https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/

Chang, B. (2018). From Internet Referral Units to international agreements: Censorship of the Internet by the UK and EU. *Columbia Human Rights Law Review*, *49*(2), 114–212.

Chen, A. (2012). Inside Facebook's outsourced anti-porn and gore brigade, where "camel toes" are more offensive than "crushed heads." Gawker.com, February 16. http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads

Cloudflare. (2017). Why we terminated Daily Stormer. Cloudflare.com (blog), August 16. https://blog.cloudflare.com/why-we-terminated-daily-stormer/

Copyright Alliance. (2016). *Comments of the Copyright Alliance Before the U.S. Copyright Office*, Docket No. 2015-7. https://copyrightalliance.org/wp-content/uploads/2016/11/Copyright-Alliance-Section-512-Comments1.pdf

Cornia, A., Sehl, A., Levy, D., & Nielsen, R. K. (2018). *Private Sector News, Social Media Distribution, and Algorithm Change*. Reuters Institute Digital News Project report. https://reutersinstitute.politics.ox.ac.uk/our-research/private-sector-news-social-media-distribution-and-algorithm-change

Cushing, T. (2017). *Someone under Federal Indictment Impersonates a Journalist to File Bogus DMCA Notice*. Techdirt, May 23. www.techdirt.com/articles/20170518/09500537404/someone-under-federal-indictment-impersonates-journalist-to-file-bogus-dmca-notice.shtml

Dara, R. (2011). *Intermediary Liability in India: Chilling Effects on Free Expression on the Internet*. https://cis-india.org/internet-governance/intermediary-liability-in-india.pdf

Duguay, S., Burgess, J., & Suzor, N. (2018). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, *26*(2), 237–252. https://doi.org/10.1177/1354856518781530

Engstrom, E., & Feamster, N. (2017). *The Limits of Filtering: A Look at the Functionality & Shortcomings of Content Detection Tools*. Engine report. www.engine.is/the-limits-of-filtering

European Commission. (2016). *Code of Conduct on Countering Illegal Hate Speech Online: First Results on Implementation*. European Commission report. http://ec.europa.eu/information_society/newsroom/image/document/2016-50/factsheet-code-conduct-8_40573.pdf

(2017). *Code of Conduct on Countering Online Hate Speech: Results of Evaluation Show Important Progress*. European Commission report. http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=71674

(2018a). Public consultation on measures to further improve the effectiveness of the fight against illegal content online. Europa.eu. https://ec.europa.eu/digital-single-market/en/news/public-consultation-measures-further-improve-effectiveness-fight-against-illegal-content-online

(2018b). EU Code of Practice on Disinformation. Europa.eu. https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

European Digital Rights. (2016). Europol: Non-transparent cooperation with IT companies. EDRi.org. https://edri.org/europol-non-transparent-cooperation-with-it-companies/

Europol. (2016). *EU Internet Referral Unit: Year One Report*. Europol report. www.europol.europa.eu/sites/default/files/documents/eu_iru_1_year_report_highlights.pdf

Facebook. (2018a). *Facebook Government Requests*. Facebook report. https://govtrequests.facebook.com

(2018b). *NetzDG Transparency Report (January–June 2018)*. Facebook report. https://fbnewsroomus.files.wordpress.com/2018/07/facebook_netzdg_july_2018_english-1.pdf

(2019a). A further update on the New Zealand terrorist attack. Facebook Newsroom, March 20. https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

(2019b). Update on New Zealand. Facebook Newsroom, March 18. https://newsroom.fb.com/news/2019/03/update-on-new-zealand/

(2019c). More coordinated inauthentic behavior from Russia. Facebook Newsroom, May 6. https://newsroom.fb.com/news/2019/05/more-cib-from-russia/

(2019d). Removing coordinated inauthentic behavior from Israel. Facebook Newsroom, May 16. https://newsroom.fb.com/news/2019/05/removing-coordinated-inauthentic-behavior-from-israel/

(2019e). Removing coordinated inauthentic behavior and spam from India and Pakistan. Facebook Newsroom, April 1. https://newsroom.fb.com/news/2019/04/cib-and-spam-from-india-pakistan/

(2019f). More coordinated inauthentic behavior from Russia. Facebook Newsroom, May 6. https://newsroom.fb.com/news/2019/05/more-cib-from-russia/

(2019g). Removing more coordinated inauthentic behavior from Iran. Facebook Newsroom, May 28. https://newsroom.fb.com/news/2019/05/removing-more-cib-from-iran/

Fisher, M. (2018). Inside Facebook's secret rulebook for global political speech. *New York Times*, December 27. www.nytimes.com/2018/12/27/world/facebook-moderators.html

Ford, G. S. (2017). Notice and takedown in everyday practice: A review. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2963230

French Secretary of State for Digital Affairs. (2019). *Interim Mission Report: Creating a French Framework to Make Social Media Platforms More Accountable*. Mission report. www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf

Galperin, E. (2008). Massive takedown of anti-Scientology videos on YouTube. Electronic Frontier Foundation, September 5. www.eff.org/deeplinks/2008/09/massive-takedown-anti-scientology-videos-youtube

Gebhart, G. (2018). Who has your back? Censorship edition 2018. Electronic Frontier Foundation, September 10. www.eff.org/who-has-your-back-2018

German Parliament. (2018). *Proposal of Representative Kunast and Others to Further Develop the Network Enforcement Law*. Document 19/5950. http://dip21.bundestag.de/dip21/btd/19/059/1905950.pdf

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

GitHub. (2015). *GitHub's 2014 Transparency Report*. Github report. https://github.com/blog/1987-github-s-2014-transparency-report

Goldman, E. (2018). COMO: Content moderation at scale conference recap. Eric Goldman (blog). https://blog.ericgoldman.org/archives/2018/07/como-content-moderation-at-scale-conference-recap.htm

Google. (n.d.). *Google Transparency Report*. Google report. https://transparencyreport.google.com

(2018a). *YouTube Community Guidelines Enforcement Report*. Google report. https://transparencyreport.google.com/youtube-policy/removals?hl=en

(2018b). *Three Years of the Right to be Forgotten*. Google report. https://g.co/research/rtbf_report

(2018c). *Removals under the Network Enforcement Law*. Google report. https://transparencyreport.google.com/netzdg/youtube

(2019). *EU Code of Practice: May Report*. Google report. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60042

Graves, Z. (2019). Does Twitter have an anti-conservative bias, or just a Nazi bias? Techdirt, July 18. www.techdirt.com/articles/20190221/16154641652/does-twitter-have-anti-conservative-bias-just-anti-nazi-bias.shtml

The Guardian. (2017). YouTube changes restrictions on gay-themed content following outcry. *The Guardian*, March 21. www.theguardian.com/music/2017/mar/21/youtube-changes-restrictions-gay-lgbtq-themed-content-tegan-sarah

Hanania, R. (2019). It isn't your imagination: Twitter treats conservatives more harshly than liberals. *Quillette*, July 18. https://quillette.com/2019/02/12/it-isnt-your-imagination-twitter-treats-conservatives-more-harshly-than-liberals/

Hargreaves, E., Agosti, C., Menasche, D., Neglia, G., Reiffers-Mason, A., & Altman, E. (2018). Biases in the Facebook News Feed: A case study on the Italian elections. arXiv.org. https://arxiv.org/abs/1807.08346

Heins, M., & Beckles, T. (2005). *Will Fair Use Survive? Free Expression in the Age of Copyright Control*. Brennan Center for Justice report. www.brennancenter.org/publication/will-fair-use-survive-free-expression-age-copyright-control

Helberger, N., Leerssen, P., & van Drunen, M. (2019). Germany proposes Europe's first diversity rules for social media platforms. LSE Media Policy Project (blog), May 29. https://blogs.lse.ac.uk/mediapolicyproject/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms/

Hopkins, N. (2017). Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*, May 21. www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

Human Rights Watch. (2006). *"Race to the Bottom": Corporate Complicity in Chinese Internet Censorship*. Human Rights Watch report. www.hrw.org/reports/2006/china0806/china0806web.pdf

Kaye, D. (2017). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, A/HRC/35/22, United Nations, August 18. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G17/077/46/PDF/G1707746.pdf?OpenElement

Keller, D. (2018). Internet platforms: Observations on speech, danger and money. Hoover Institution Aegis Paper Series No. 1807. www.hoover.org/research/internet-platforms-observations-speech-danger-and-money

(2019a). Build your own Intermediary Liability law: A kit for policy wonks of all ages. Balkinization (blog), June 11. https://balkin.blogspot.com/2019/06/build-your-own-intermediary-liability.html?m=1

(2019b). Who do you sue? State and platform hybrid power over speech. Hoover Institution Aegis Paper Series No. 1902. www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131* (6), 1599–1669.

(2019). Inside the team at Facebook that dealt with the Christchurch shooting. *The New Yorker*, July 18. www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting

Krause, T., & Grasegger, H. (2016). Facebook's secret rules of deletion. *Süddeutsche Zeitung* Internet Archive. https://web.archive.org/web/20170103040043/http://international.sueddeutsche.de/post/154543271930/facebooks-secret-rules-of-deletion

Kreiss, D., & McGregor, S. (2019). The "arbiters of what our voters see": Facebook and Google's struggle with policy, process, and enforcement around political advertising. *Political Communication*, *36*(2). www.tandfonline.com/eprint/9IUFbkGmZ4YHtNXM7sQ2/full?target=10.1080/10584609.2019.1619639

Kuczerawy, A. (2017). The power of positive thinking: Intermediary liability and the effective enjoyment of the right to freedom of expression. *Journal of Intellectual Property, Information Technology and E-Commerce Law*, *8*(3), 226–237.

Levin, S. (2017). Civil rights groups urge Facebook to fix "racially biased" moderation system. *The Guardian,* January 18. www.theguardian.com/technology/2017/jan/18/facebook-moderation-racial-bias-black-lives-matter

Leyden, J. (2004). How to kill a website with one email: Exploiting the European E-commerce Directive. *The Register*, September 10. www.theregister.co.uk/2004/10/14/isp_takedown_study

Macdonald, F. (2016). Google's deleted an artist's blog, along with 14 years of his work. *Science Alert*, July 18. www.sciencealert.com/google-has-deleted-an-artist-s-blog-with-14-years-of-his-work

MacDonald, K. (2014). *A Human Rights Audit of the Internet Watch Foundation*. Lord MacDonald Report. www.iwf.org.uk/sites/default/files/inline-files/Human_Rights_Audit_web.pdf

MacKinnon, R. (2012). *Consent of the Networked: The Worldwide Struggle For Internet Freedom*. New York: Hachette Book Group.

Marthews, A., & Tucker, C. E. (2017). Government surveillance and internet search behavior. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2412564

Masnick, M. (2018). Dubious studies and easy headlines: No, a new report does not clearly show Facebook leads to hate crimes. Techdirt, August 3. www.techdirt.com/articles/20180823/00122840491/dubious-studies-easy-headlines-no-new-report-does-not-clearly-show-facebook-leads-to-hate-crimes.shtml?threaded=true

McIntyre, T. J. (2018). Internet censorship in the United Kingdom: National schemes and European norms. In L. Edwards (Ed.), *Law, Policy and the Internet* (pp. 291–330). Oxford: Hart Publishing.

McLaughlin, T. (2018). How Facebook's rise fueled chaos and confusion in Myanmar. *Wired*, September 10. www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar

Medium. (2015). *Medium's 2015 Transparency Report*. Medium report. https://blog.medium.com/medium-s-2015-transparency-report-5c6205c48afe

Meleagrou-Hitchens, A., & Kaderbhai, N. (2017). *Research Perspectives on Online Radicalisation: A Literature Review, 2006–2016*. VOX-Pol report. www.voxpol.eu/new-vox-pol-report-research-perspectives-online-radicalisation

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. https://link.springer.com/article/10.1007/s11109-016-9373-5

O'Brian, D., & Malcolm, J. (2018). 70+ internet luminaries ring the alarm on EU copyright filtering proposal. Electronic Frontier Foundation, June 12. www.eff.org/deeplinks/2018/06/internet-luminaries-ring-alarm-eu-copyright-filtering-proposal

Ong, T. (2018). Facebook is recruiting external advisers to tackle claims of bias on its platform. *The Verge*, May 2. www.theverge.com/2018/5/2/17310894/facebook-recruiting-external-advisers-bias-minority-conservative

Open Technology Institute. (2018). *The Transparency Reporting Toolkit: Content Takedown Reporting*. Open Technology Institute report. www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/

Oransky, I. (2013). WordPress removes Anil Potti posts from Retraction Watch in error after false DMCA copyright claim. Retraction Watch, February 5. https://retractionwatch.com/2013/02/05/wordpress-removes-anil-potti-posts-from-retraction-watch-in-error-after-false-dmca-copyright-claim

Pen America (2013). *Chilling Effects: NSA Surveillance Drives U.S. Writers to Self-Censor*. Pen America report. https://pen.org/chilling-effects

Penney, J. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal*, *31*(1), 172. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2769645

Perel, M., & Elkin-Koren, N. (2017). Black box tinkering: Beyond disclosure in algorithmic enforcement. *Florida Law Review*, *69*(181). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2741513

Regulation 2019/1150, 2019 O.J. (L 186) 57 (EU)

Reda, J. (2017).When filters fail: These cases show we can't trust algorithms to clean up the internet. Julia Reda website. https://juliareda.eu/2017/09/when-filters-fail/

Reddit, Inc. (2015). *Reddit, Inc. Transparency Report, 2015*. www.reddit.com/wiki/transparency/2015

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. Media Studies Publications, Paper No. 12. https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1012&context=commpub

(2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Seng, D., (2015). "Who watches the watchmen?" An empirical analysis of errors in DMCA takedown notices. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2563202

Taub, A., & Fisher, M. (2018). Facebook fueled anti-refugee attacks in Germany, new research suggests. *New York Times*, August 21. www.nytimes.com/2018/08/21/world/europe/facebook-refugee-attacks-germany.html

Timmer, J. (2013). Site plagiarizes blog posts, then files DMCA takedown on originals. *Ars Technica*, February 5. https://arstechnica.com/science/2013/02/site-plagiarizes-blog-posts-then-files-dmca-takedown-on-originals

Torrent Freak. (2018). Google Drive uses hash matching to detect pirated content. TorrentFreak.com. https://torrentfreak.com/google-drive-uses-hash-matching-detect-pirated-content

Tumblr. (n.d.). *Transparency Report*. Tumblr report. www.tumblr.com/transparency

Twitter. (2017). *Twitter Transparency Report Copyright Notices January – June 2017*. Twitter Report. https://transparency.twitter.com/en/copyright-notices.html#copyright-notices-jan-jun-2017

(2018a). *Twitter Rules Enforcement Report*. Twitter report. https://transparency.twitter.com/en/twitter-rules-enforcement.html

(2018b). *Twitter Netzwerkdurchsetzungsgesetzbericht: Januar – Juni 2018*. Twitter report. https://cdn.cms-twdigitalassets.com/content/dam/transparency-twitter/data/download-netzdg-report/netzdg-jan-jun-2018.pdf

(2019). *EU Code of Practice: May Report*. Twitter report. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60043

UK Metropolitan Police. (2016). 250,000th piece of online extremist/terrorist material to be removed. Archive.org. https://web.archive.org/web/20180306202125/http://news.met.police.uk/news/250000th-piece-of-online-extremist-slash-terrorist-material-to-be-removed-208698

Urban, J. M., Karaganis, J., & Schofield, B. (2016). Notice and takedown in everyday practice. UC Berkeley Public Law Research, Paper No. 2755628. http://ssrn.com/abstract=2755628

(2017). Response to "Notice and takedown in everyday practice: A review." The Takedown Project. http://takedownproject.org/770-2

Urban, J. M., & Quilter, L. (2006). Efficient process or "chilling effects": Takedown notices under Section 512 of the Digital Millennium Copyright Act. *Santa Clara High Tech Law Journal*, 22(4): 621–693.

US Copyright Office. (2015). *Section 512 Study*. US Copyright Office report. www.copyright.gov/policy/section512/

US House of Representatives. (2019). Social Media Advertisements. Intelligence.House.Gov. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm

US Patent and Trademark Office. (2015). Multistakeholder Forum on the DMCA Notice and Takedown System. USPTO.Gov. www.uspto.gov/learning-and-resources/ip-policy/copyright/multistakeholder-forum-dmca-notice-and-takedown-system

Venturini, J., Louzada, L., Maciel, M., Zingales, N., Stylianou, K., & Belli, L. (2016). *Terms of Service and Human Rights: An Analysis of Online Platform Contracts*. Report. http://bibliotecadigital.fgv.br/dspace/handle/10438/18231

Vivanco, J. M. (2014). Censorship in Ecuador has made it to the Internet. Human Rights Watch news release, December 15. www.hrw.org/news/2014/12/15/censorship-ecuador-has-made-it-internet

Volokh, E., & Levy, P. A. (2016). Dozens of suspicious court cases, with missing defendants, aim at getting web pages taken down or deindexed. *Washington Post*, October 10. www.washingtonpost.com/news/volokh-conspiracy/wp/2016/10/10/dozens-of-suspicious-court-cases-with-missing-defendants-aim-at-getting-web-pages-taken-down-or-deindexed/?utm_term=.b61e8a5967b2

Walsh, D., & Zway, S. A. (2018). A Facebook war: Libyans battle on the streets and on screens. *New York Times*, September 4. www.nytimes.com/2018/09/04/world/middleeast/libya-facebook.html

West, L. (2017). I've left Twitter. It is unusable for anyone but trolls, robots and dictators. *The Guardian*, January 3. www.theguardian.com/commentisfree/2017/jan/03/ive-left-twitter-unusable-anyone-but-trolls-robots-dictators-lindy-west

York, J. C., Faris, R., Deibert, R., & Heacock, R. (2018). Policing content in the quasi-public sphere. OpenNet Initiative bulletin, September. https://opennet.net/policing-content-quasi-public-sphere

York, J. C., & Gullo, K. (2018). Offline/Online project highlights how the oppression marginalized communities face in the real world follows them online. Electronic Frontier Foundation, March 6. www.eff.org/deeplinks/2018/03/offlineonline-project-highlights-how-oppression-marginalized-communities-face-real

YouTube. (2010). Broadcast yourself. YouTube (official blog), March 18. https://youtube.googleblog.com/2010/03/broadcast-yourself.html