Governance of and by platforms

Tarleton Gillespie

SAGE Handbook of Social Media,
edited by Jean Burgess, Thomas Poell, and Alice Marwick
Sage, forthcoming (2017)

Platforms rose up out of the exquisite chaos of the web. Their founders were inspired by the freedom it promised, but also hoped to provide spaces for the web's best and most social aspects. But as these platforms grew, the chaos found its way back onto them – for obvious reasons: if I want to say something, be it inspiring or reprehensible, I want to say it where people are likely to hear me. Today, we by and large speak on platforms when we're online. Social media platforms put people at "zero distance" (Searls, 2016) from one another, afford them new opportunities to speak and interact, and organize them into networked publics (Varnelis, 2008; boyd, 2011) – and though the benefits of this may be obvious, even seem utopian at times, the perils are also painfully apparent.

While scholars have long discussed the dynamics of free speech online, much of that thinking preceded the dramatic migration of online discourse to platforms (Balkin, 2004; Godwin, 2003; Lessig, 1999; Litman, 1999). By *platforms,* I mean sites and services that host public expression, store it on and serve it up from the cloud, organize access to it through search and recommendation, or install it onto mobile devices. This includes Facebook, YouTube, Twitter, Tumblr, Pinterest, Google+, Instagram, and Snapchat… but also Google Search and Bing, Apple App Store and Google Play, Medium and Blogger, Foursquare and Nextdoor, Tinder and Grindr, Etsy and Kickstarter, Whisper and Yik Yak. What unites them all is their central offer: to host and organize user content for public circulation, without having produced or commissioned it. They don't make the content, but they make important choices about that content: what they will distribute and to whom, how they will connect users and broker their interactions, and what they will refuse. With this growing and increasingly powerful set of digital intermediaries, we have to revisit difficult questions about how they structure the speech and social activity they host, and what rights and responsibilities should accompany that (DeNardis and Hackl, 2015; MacKinnon et. al., 2014; Gillespie, 2015; Grimmelman, 2015; Obar and Wildman, 2015; van Dijck, 2013; Wagner, 2013).

Traditional private information providers – publishers, broadcasters, resellers, telecommunications – already have established legal obligations for the speech they facilitate, in the U.S. and elsewhere (Baker, 2001; Benkler, 1998; Braman, 2004; Entman & Wildman, 1992; Freedman, 2008; Hendershot, 1999; Horwitz,

1991a; Horwitz, 1991b; Streeter, 1996). But traditional communication policies have proven hard to apply, honor, and enforce online (Bar & Sandvig, 2008; Braman, 2014; Castronova, 2014; Johnson & Post, 1996; Lessig, 1999; Tushnet, 2008). Even Internet-centric solutions formulated in an earlier moment, such as limited liability, safe harbor, and takedown measures for search engine and Internet Service Providers (ISPs), are arguably an ill fit for social media platforms (MacKinnon et al, 2014). Today, platforms face more vocal calls to both permit contentious speech and curate it, from policymakers, from users, from foreign governments, from activists, and from the press.

This essay will begin by discussing the governance of platforms: the policies that have emerged in the past decade specifying their liabilities (or lack thereof) for the user content and activity they host. In the U.S., these regulations are limited by a fundamental reluctance to constraint speech, whereas internationally, these same platforms face a wider array of restrictions. It will then consider governance by platforms. This is related to the first, but is not the same. Social media platforms have increasingly taken on the responsibility of curating the content and policing the activity of their users: not simply to meet legal requirements, or to avoid having additional policies imposed, but also to avoid losing offended or harassed users, to placate advertisers eager to associate their brands with a healthy online community, to protect their corporate image, and to honor their own personal and institutional ethics. Some of these interventions are welcomed by users, while others have been more contentious. The regulatory framework we impose on platforms, and the ways in which the major platforms enact those obligations and impose their own on their users, are settling in as the parameters for the how public speech online is and will be privately governed.

governance OF platforms

Platforms vary, in ways that matter both for the influence they can assert over users and for how they should be governed. It is deceptively easy, in public debates and in scholarship, to simply point accusingly at Facebook and move on, without acknowledging the variety of purpose, scope, membership, economics, and design across the sites and services that call themselves platforms. In fact, 'platform' is a slippery term, in part because there may be little that unites different sites as a category, and in part because it gets deployed strategically, by stakeholders and critics alike. (Gillespie, 2010) As shorthand, it too easily equates a site with the company that offers it, implies that social media companies act with one mind, and downplay the people involved. Platforms are socio-technical assemblages and complex institutions; they're not even all commercial, and even the commercial ones are commercial in different ways. At the same time,

'platform' is a widely used term by the companies themselves. And many discourses of responsibility and liability (legal and otherwise) conceive of institutions as singular entities, and for good reason.

in the middle

In the language of U.S. information policy, "platform" as a term has not enjoyed much traction. Most of the policies that currently apply to social media platforms were crafted before their emergence, to address a broader category of online services and access providers. The preferred term of art, "online intermediaries," which replaced an earlier and now archaic term, "interactive computing services," is broader. The Organisation for Economic Co-operation and Development (OECD) definition helps highlight what's common to all these terms: "Internet intermediaries bring together or facilitate transactions between third parties on the Internet. They give access to, host, transmit and index content, products and services originated by third parties on the Internet or provide Internet-based services to third parties." (OECD, 2010) The definition highlights two important aspects: (1) online intermediaries come between and facilitate the connection of others; and (2) the content they transmit is produced by others.

Contemporary social media platforms fit this category, but they also complicate it. They are not 'content producers' (though in practice they do produce lots of ancillary content along the way); rather, they host, store, organize, and circulate the content of others. While the hosting provided by platforms is more involved than that of ISPs, and the organizing of content provided by platforms is more involved than that of search engines, these differences are of degree more than of kind, given that all network services store and circulate content as part of their service, at least temporarily or incidentally.

By calling them intermediaries, let's recognize that social media platforms are fundamentally in the middle -- that is, they mediate between users who produce content and users who might want it. This makes them similar to not only search engines and ISPs, but also traditional media. They too face a regulatory framework premised on the fact that they mediate between producers and audiences, between speakers and listeners.

Social media platforms are not only in the middle between user and user, and user and public, but between citizens and law enforcement, policymakers, and regulators charged with governing their behavior. Online, illicit activity can be difficult to pinpoint and difficult to police: users can enjoy the anonymity provided by some sites, and the obscurity provided by encryption and transient Internet connections; illicit content moves easily across regional jurisdictions, and has oblique or cumulative effects. Since platforms gather people and collect traces of their activity, they present a compelling opportunity to policymakers to

govern users through them. The governance of platforms is marked by, and struggles with, this middle-ness, and the thorny questions of convenience and responsibility that come with it.[1]

Public and policy concerns around illicit content, at first largely focused on sexually explicit and graphically violent images, have expanded in recent years: to include with additional categories like hate speech, self-harm, and extremism; and to deal with the enormous problem of user behavior targeting other users, including misogynistic, racist, and homophobic attacks, trolling, harassment, and threats of violence. And questions about the responsibility of platforms are expanding as the range of platforms expand: to social platforms that circulate goods (auction sites like eBay, exchange sites like Craigslist, and e-commerce platforms like Etsy), that circulate money or investment (Kiva, Venmo), that circulate labor (Amazon Mechanical Turk, Uber, Taskrabbit), or that trade access to physical services (AirBnB). Each of these intersects with other regulatory frameworks, but each also includes fundamental questions about whether and how platforms should be responsible for their (independent, amateur, non-salaried) users' speech and actions.

the myth of the impartial platform

Social media platforms have long positioned themselves as open, impartial, and noninterventionist, perhaps in part to avoid liability and regulation, and in part because their founders fundamentally believe it to be so. Twitter, for example, begins its Rules with, "We respect the ownership of the content that users share and each user is responsible for the content he or she provides. Because of these principles, we do not actively monitor and will not censor user content, except in limited circumstances described below." It's a curious statement, given the list of prohibitions that follow.

This fundamental mystification of the role of platforms began when platforms did: from their earliest presentation they have often characterized themselves as open to all comers; in their promotion they often suggest that they merely facilitate public expression, that they are impartial and hands-off hosts, with an "information will be free" ethos, and that being so is central to their mission (Gillespie, 2010; Vaidhyanathan, 2012). Though users seem to be recognizing that platforms intervene in myriad ways, are growing increasingly concerned about it, platforms continue to perform their impartiality.

This is odd, considering that, from a different view, everything on a platform is designed and orchestrated. While social activity would exist without Facebook or Twitter, the kind of social activities that occur there depend powerfully on the space and structure they provide (Baym & boyd, 2012; Bruns and

[1] It is visible not only in the regulation of illicit content, but in legal efforts to protect user privacy, and in digital copyright law. Much of the way we think of intermediaries as protected legal entities was forged in the "copyright wars" of the 2000s (Yu, 2003). This essay will focus on the regulation of illicit content, though many of the tensions involved are relevant for ongoing concerns about copyright and privacy.

Burgess, 2015; Couldry & van Dijck, 2015; Gerlitz & Helmond, 2013; Langlois, 2013; Sandvig, 2015; Shepherd & Landry, 2013; van Dijck, 2013; Weltevrede et al, 2014). These structures are certainly not neutral: they are designed to invite and shape participation, toward particular ends. This includes what kind of participation they invite and encourage; what gets displayed first or most prominently; how the platforms design navigation from content to user to exchange; the pressures exerted by pricing and revenue models; and how they organize information through algorithmic sorting, privileging some content over others, in opaque ways. And it includes what is not permitted, and how and why they police objectionable content and behavior (Gillespie, 2014; Grimmelmann, 2015).

the rise of safe harbor

In the 1990s, policymakers in the U.S. and elsewhere became aware of growing concerns about the proliferation of illicit content on the web, especially pornography and piracy.[2] In such cases, it proved difficult to directly pursue online "publishers" for their illegal or illicit behavior, particularly when those publishers were individuals, usually amateurs, sometimes anonymous, and hard to locate and identify. Because of this, some lawsuits brought in the U.S. for libel, publication of private documents, and the distribution of hate speech, began targeting not the individual user but the Internet service provider disseminating the content (Ardia, 2010; Kreimer, 2006; Mann & Belzley, 2005).

In the United States, Congress crafted a legislative response to some of these issues, the Communication Decency Act (CDA), as part of a massive telecommunications bill. Passed in 1996, the CDA made it illegal to provide "obscene or indecent" material to minors. But the ban was determined to be unconstitutional[3] by the Supreme Court only a year later. However, parts of the law survived, including the defenses it provided for "interactive computer service providers" – safe harbors against any liability for harmful material their users might provide. Because these safe harbors were not at issue in the *Reno v. ACLU* Supreme Court decision, they have remained a part of U.S. telecommunication law, known as Section 230.

The Section 230 safe harbor has two parts (Mueller 2015). The first is that intermediaries cannot be held liable for the speech of their users, since they merely provide access to the Internet or other network services; they are not 'publishers' of their users' content, in the legal sense. Presumably, this implies that

[2] This was fueled by a vocal and urgent panic, in American culture, about the availability of pornography online, a concern not unwarranted but wildly overstated (Maddison, 2010; Marwick, 2008). Copyright infringement lawsuits began to appear (Ginsburg, 1995), but they were more concerned with images and private documents than music and movies, which would emerge as a problem later with the rise of peer-to-peer file sharing services like Napster (Litman, 2001; Yu, 2003).

[3] *Reno v. American Civil Liberties Union,* 521 U.S. 844 (1997).

intermediaries do not need to police what their users say and do. The second, less familiar part adds a twist. If an intermediary does decide to police what their users say or do, they don't lose their safe harbor protection by doing so. In other words, policing content on their own accord does not suddenly make them 'publishers,' nor does it require them to meet some standard of effective policing. This second half was crafted so that, even though the safe harbor makes it possible for intermediaries not to intervene, it would also not discourage them from doing so, by making them more liable for it than if they had simply turned a blind eye.

Section 230 leans on a legislative distinction in U.S. telecommunication law, between publishers who *provide* information (and therefore can be held liable for it) and distributors that merely *circulate* the information of others (and should not) – commonly known as the "content/conduit" distinction. Since ISPs offer 'access' to the Internet, and do not produce the content they help circulate, the law prioritizes the free movement of information, and asserts strong limits on any liability for the content therein.[4] As with telephone systems, holding a provider liable for what users say or do might encourage that provider to monitor users proactively. This would be not only practically impossible and financially unbearable,[5] but also politically undesirable. Legislators and technologists feared that this might also discourage online innovation, out of fear of lawsuits (CDT, 2010).

Outside of the US, few nations offer the kind of safe harbor provided in Section 230. MacKinnon et.al. (2014, 42) dub the U.S. approach "broad immunity," the most lenient of the three types of intermediary liability regimes they identify. Most of the European Union nations, as well as Russia and many of the South American nations, offer intermediaries "conditional liability," which is more akin to the U.S. rules for copyright: platforms are not liable for what their users post or distribute, as long as they have no "actual knowledge" of, and did not produce or initiate the illegal or illicit material, and if they respond to request

[4] It is worth noting that Section 230 was for "offensive material" and explicitly excluded "cases involving federal criminal law, intellectual property law, and electronic-communications privacy law." So the safe harbor it establishes for ISPs and platforms does not apply to these other concerns. This explains why the platform obligations for child pornography are very different than for other categories of harmful speech, because child pornography is a federal, criminal offense. It also explains why the arrangements are different for copyright infringement. The Digital Millennium Copyright Act, also passed in 1996, offered ISPs and search engines protection against a charge of contributing to copyright infringement as well, but this safe harbor comes with some obligations, the most notable being that intermediaries must comply with "notice and takedown" requests from copyright owners who have identified their work as being circulated through their service (Fifer and Carter, 2004). Also, the law is concerned with how much knowledge the intermediary had of the infringing material, and their relative ability to patrol for it. In court cases that followed, peer-to-peer networks and other online services found they did not enjoy the DMCA safe harbor when they had "materially" contributed to the circulation of pirated content, when they enjoyed some financial benefit from it, or even when they had "induced" it by promoting their service as designed for piracy.

[5] This is before innovations such as digital fingerprinting and other forms of automated content identification, techniques that now make it possible for platforms and ISPs to "know" of illicit content on their service, even in real time.

from the state or the courts to remove illicit third-party content. Nations such as China and many of the nations in the Middle East impose "strict liability," (MacKinnon 2014, 40) requiring Internet intermediaries to prevent the circulation of illicit or unlawful content. This generally means proactively removing or censoring, often in direct cooperation with the government. Without a regulatory bulwark against state intervention, these private actors are much more beholden to government demands, and in some cases even rules that prohibit political speech. Finally, some nations, for example in sub-Saharan Africa, have not instituted laws articulating the responsibilities of Internet intermediaries in any form, leaving intermediaries there uncertain about what they might or might not be liable for.

But while it is the most generous, even the U.S. safe harbor for intermediaries embodies conflicting views of online service providers (OSPs). As Mueller writes, Section 230

> "was intended both to immunize OSPs who did nothing to restrict or censor their users' communications, and to immunize OSPs who took some effort to discourage or restrict online pornography and other forms of undesirable content. Intermediaries who did nothing were immunized in order to promote freedom of expression and diversity online; intermediaries who were more active in managing user-generated content were immunized in order to enhance their ability to delete or otherwise monitor 'bad' content." (805)

These competing impulses, between allowing intermediaries to stay out of the way and encouraging them to police their users, continue to shape the way we think about the role and responsibility of Internet intermediaries, and has extended to how we regulate social media platforms.

the pressures on safe harbor

From a legal standpoint, broad and conditional safe harbors are profoundly advantageous for Internet intermediaries. Notice-and-takedown requirements generate real challenges for platforms, and are prone to abuse, but are far preferable for platforms than being held liable for what their users post (Urban et. al., 2016). As Tushnet (2008, p. 1002) notes, "Current law often allows Internet intermediaries to have their free speech and everyone else's too." However, while safe harbor provisions have held up for nearly a decade, platforms face three distinct challenges that reveal the limitations of the safe harbor provision, and in some cases are fueling calls for its reconsideration.

First and perhaps most obviously, most of these laws were not designed with social media platforms in mind. When Section 230 was being crafted, few such platforms existed. U.S. lawmakers were addressing a web largely populated by ISPs and amateur web "publishers" – amateurs posting home pages, companies designing websites, and online communities having discussions. Besides ISPs who simply provided access to

the network, the only intermediaries at the time were ISPs that also doubled as content "portals," like AOL and Prodigy; the earliest search engines like Altavista and Yahoo; and operators of BBS systems, chatrooms, and newsgroups. The law predates not just Facebook, but MySpace, Friendster, Napster and its peer-to-peer brethren, even Google's search engine. Blogging was in its infancy, well before the invention of tools like Blogspot and Wordpress; eBay, Craigslist, and Match.com were less than a year old; and the ability to comment on a web page had not yet been modularized into a plugin.

Although they were not included or anticipated in the law, social media platforms have generally claimed that they enjoy its safe harbor. But many of the assumptions that animated intermediary liability (particularly the questions of whether the intermediary has knowledge of illicit content, could conceivably intervene in its circulation, and benefits financially from it) are tested by contemporary social media platforms. U.S. regulatory traditions, such as protecting "conduits" from liability so they are not encouraged to monitor or censor the content traveling through them, are an ill fit for YouTube's ContentID (which can automatically identify copyrighted music in user-submitted videos) or Facebook's NewsFeed algorithm (which constructs a curated feed from user posts designed to keep users interested and attentive to advertising) or the anonymous attacks possible with Yik Yak.

Second, while intermediary liability regimes are typically nation-specific, platforms largely are not. ISPs are almost exclusively located in the nation in which regulation is imposed and enforced, both in terms of the (physical and legal) location of the company, its material infrastructure, and its users. This is not the case for the likes of Twitter, Instagram, or Wikipedia. Currently, most of the major social media platforms are, as corporate and legal entities, based in the United States, where they enjoy the broadest safe harbor, but they serve millions of users living in nations that impose much stricter liability, or have specific requirements about responding to state or court requests to remove content.

Major social media platforms have had to develop their own policies on how to respond to requests from foreign governments to remove content. Google famously pulled out of China rather than filter its search results according to Chinese dictates (although there were certainly a variety of motivations for the move).[6] LinkedIn remained, by honoring the Chinese government's policies and seeking financial investment from Chinese firms.[7] Twitter will remove tweets in response to government requests, but does so only for users in that nation rather than removing them from the entire service, and will indicate what has been

[6] Branigan, T. (2010). Google Angers China by shifting service to Hong Kong. *The Guardian*, march 23. Retrieved from https://www.theguardian.com/technology/2010/mar/23/google-china-censorship-hong-kong

[7] Mozur, P. and Goel, V. (2014). To Reach China, LinkedIn Plays by Local Rules. *New York Times*, October 5. Retrieved from http://www.nytimes.com/2014/10/06/technology/to-reach-china-linkedin-plays-by-local-rules.html

removed and at whose behest.[8] Facebook, in at least one case, removed content at the request of the Pakistani government by rendering it invisible to searches emanating from that country. Many of the major platforms publish data on the number of removal requests they receive, by country and by category of request. Some have even included "warrant canaries" in their policy statements, a sentence stating that no government subpoenas had been served – which they would remove when it was no longer true, to alerting those in the know that a subpoena had been served without violating a gag order.

Because Western platforms have been cautious about how they respond to removal requests from foreign governments, some nations have threatened to block content they deem illegal or offensive. China and the Islamic nations of the Middle East and North Africa have been most aggressive in this tactic. This typically involves providing local ISPs with a 'blacklist' of pages deemed criminal or otherwise unacceptable. This tactic, of course, is made more complicated by massive platforms such as social networking sites and discussion platforms, where the offending post or video is just one element of a massive, complex, and constantly changing database. As Palfrey (2010) observes, this tends to result in 'overfiltering,' where a nation will threaten to block not a single YouTube video or Facebook user, but YouTube or Facebook in its entirety. What often follows is a high-stakes game of chicken: platforms do not relish being entirely blocked from an entire nation of users; at the same time, doing so is risky for the government as well, as it may have costs in terms of public sentiment. For countries with a stronger commitment to freedom of expression or independent telecommunications, this tendency to block legitimate content along with the offensive is an unpalatable one.

Third, presumptions about how much liability platforms should face, and for what reason, have been challenged by categories of content particularly abhorrent to users and governments. These hesitations are happening in all corners of the world: even U.S. policy, with the broadest safe harbor, has shifted in the face of specific concerns.

Most pressing has been, unsurprisingly, the issue of terrorism. Certainly, terrorist organizations have grown increasingly savvy in the use of social media platforms (Archetti, 2015).[9] At the same time, combatting terrorism is a compelling justification for the imposition of policies that may have other aims as well (MacKinnon, 2012). Even in the United States, where the ethos of the First Amendment typically provides information providers a powerful shield against government intrusion, terrorism serves as an effective

[8] Chao, L. and Efrati, A. (2012). Twitter can censor buy Country. *Wall Street Journal*, January 28. Retrieved from http://www.wsj.com/articles/SB10001424052970204573704577185873204078142

[9] Geller, E. (2016). Why ISIS is winning the online propaganda war. *The Daily Dot*, March 29. Retrieved from http://www.dailydot.com/politics/isis-terrorism-social-media-internet-countering-violent-extremism/. Waddell, K. 2016. The Government Is Secretly Huddling With Companies to Fight Extremism Online. *The Atlantic*, March 9. Retrieved from http://www.theatlantic.com/technology/archive/2016/03/the-government-is-secretly-huddling-with-companies-to-fight-extremism-online/472848/

rhetorical challenge to that. In Europe, this has meant an acceleration of the time in which platforms, once informed of terrorist content, must remove it. Under the UK Terrorism Act of 2006, platforms now have two days to comply with a takedown request, otherwise, they are deemed to have "endorsed" the terrorist content.[10] Several governments in the Middle East have instituted new laws (or attempted to) regarding terrorism that affect platforms. In Egypt, for example, a law drafted in 2014 gave authorities much wider latitude to intervene in and surveil online communication for suspected terrorist activity. Similar laws have been passed in Jordan, Qatar, and Saudi Arabia.[11]

In Europe, hate speech and racial discrimination have also fueled debates about the obligations of social media platforms. Germany and France both have laws prohibiting the promotion of Nazism, anti-Semitism, and white supremacy. The French law produced one of the earliest online content cases, in which Yahoo was compelled to prevent French users from accessing online auctions of Nazi memorabilia.[12] More recently, when anti-Semitic comments began appearing on Twitter under the hashtag #unbonjuif, or "a good jew," French courts pressed Twitter to turn over the user data behind the offending tweets (Mackinnon et al, 2014). Similar concerns have emerged in other parts of the world. In Argentina, an addition to their anti-discrimination law is currently under consideration that would require intermediaries to monitor and remove comments that were racist or discriminatory, and would even encourage them to remove the comment features of their sites entirely.[13]

Nations that do not share the American version of freedom of expression have been more willing to criminalize speech that criticizes the government or upsets public order. Some nations are extending limits on press freedoms to bloggers and even amateur speech on social media platforms. Laws that curtail the press online have appeared in Egypt, Iran, Pakistan, Tunisia, and the United Arab Emirates.[14] In other nations, including Kuwait and Lebanon, laws that prohibit the disruption of public order have been applied to political activists.[15] Some countries prohibit speech directly criticizing their leaders, and in some cases these rules have been extended to social media platforms. In 2012, authorities in Brazil arrested the head of Google Brazil for refusing to remove YouTube videos that targeted Brazilian political candidates,[16] and Facebook now

---

[10] JISC (Joint Information Systems Committee). (2007). Hosting Liability. https://www.jisc.ac.uk/guides/hosting-liability

[11] Radsch, C. (2015). Treating the Internet as the enemy in the Middle East, *Committee to Protect Journalists*. Apil 27. https://cpj.org/2015/04/attacks-on-the-press-treating-internet-as-enemy-in-middle-east.php

[12] Lasar, M. (2011). Nazi hunting: How France first "civilized" the Internet. *Ars technical*, June 22. http://arstechnica.com/tech-policy/2011/06/how-france-proved-that-the-internet-is-not-global/

[13] Bogado, D. (2015). No to Internet Censorship in Argentina. *Deep Links (Electronic Frontier Foundation)*, August 11. https://www.eff.org/deeplinks/2015/08/no-internet-censorship-argentina

[14] Radsch, (2015), *op. cit.*

[15] *ibid.*

[16] Brooks, B. and Barbassa, J. (2012). Arrest of Google Brazil head stirs debate over Web. *AP*. http://finance.yahoo.com/news/arrest-google-brazil-head-stirs-debate-over-210814484--finance.html

complies with Turkish law criminalizing defamation of the country's founder Mustafa Kemal Ataturk, or the burning of the Turkish flag, by removing any such content flagged by users.[17]

Other countries have used laws that exist purportedly to combat cybercrime, protect children, or prohibit terrorist content, to pressure platforms to remove politically contentious materials. Russia has been the innovator in this regard. In 2009, Russian law held that website owners are responsible for what users post in the comments on their site. In 2012, they developed a 'blacklist' of sites that include 'forbidden information' (illicit drugs, porn, suicide), requiring Russian ISPs to block these sites. ISPs were forced to respond to requests not only from the court or state regulatory authorities, but also from regular citizens, including the 'Media Guard' youth group, which was targeting gay teen forums and Ukrainian political organizations.[18] In 2014, the Russian government took a bolder step: a new dictate would require transnational platforms that have Russian users to store those users' data on servers located physically in Russia - otherwise the whole platform would be blocked on a national scale.[19] The revelations of NSA surveillance by Edward Snowden were used as justification, but many suspect that housing the data inside of Russia's borders would make it easier for the government to access that data and squelch political speech. As of this writing, the (mostly U.S-based) platforms have refused, and Russia has extended the deadline for compliance. In addition, in 2015 Russia decreed that bloggers with more than 3000 page views per day register as media and follow Russian media laws. This seems to include users with over 3000 daily visitors on Twitter and Facebook.[20]

The United States has by and large stayed true to the safe harbor protections first offered to online intermediaries. But growing concerns about terrorism and extremist content, harassment and cyber bullying, and the circulation of non-consensual images (commonly known as "revenge porn") have tested this commitment. A number of platforms have developed specific policies prohibiting revenge porn,[21] modeled on the notice-and-takedown arrangements in copyright law: platforms are not obligated to proactively look for violations, but will respond to requests to remove them. This involves the kind of adjudicating platforms prefer to avoid: determining whether a complainant (who may not even be a user of that platform) is in fact

[17] http://www.hurriyetdailynews.com/zuckerberg-notes-turkeys-defamation-laws-over-ataturk-as-facebook-updates-rules.aspx?pageID=238&nID=79771&NewsCatID=359

[18] Turovsky, D. (2015). This is how Russian Internet censorship works. *Meduza*, August 13. https://meduza.io/en/feature/2015/08/13/this-is-how-russian-internet-censorship-works

[19] Sonne, P. and Razumovskaya, O. (2014). Russia Steps Up New Law to Control Foreign Internet Companies. *Wall Street Journal*, September 24. http://www.wsj.com/articles/russia-steps-up-new-law-to-control-foreign-internet-companies-1411574920

[20] Luhn, A. (2015). Russia threatens to ban Google, Twitter and Facebook over extremist content. *The Guardian*, May 20. http://www.theguardian.com/world/2015/may/20/russia-threaten-ban-google-twitter-facebook-bloggers-law

[21] Daileda, C. (2015). Social media sites may be better than the law at blocking revenge porn. *Mashable*, March 18. http://mashable.com/2015/03/18/banning-revenge-porn/#E5HZfe5inkqd

the subject of the video or photo, whether the material was posted with or without the subject's consent, who owns the imagery and thus the right to circulate it, and so forth. In early 2016, the Obama administration urged U.S. tech companies to develop new strategies for identifying extremist content, either to remove it or to report it to national security authorities.[22] Around harassment, pressure is coming from users, particularly women and racial minorities, who feel that the abuses leveled upon them by other users have become so unbearable that platforms have an obligation to intervene (Kayyali and O'Brien, 2015; Matias et al, 2015).

Together, these calls to hold platforms liable for specific kinds of abhorrent content or behavior, and the increasing challenges posed by governments seeking to use platforms as a way to constrain political speech and activism, are undercutting the once sturdy principle of safe harbor articulated in Section 230 and elsewhere. As these platforms multiply in form and purpose, become more and more central to how and where users encounter each other online, and involve themselves in the circulation not just of words and images but of goods, money, services, and labor, intermediary liability seems more and more insufficient. Platforms face both more vocal calls to permit contentious speech and more compelling reasons to curate it – not just under pressure from laws, but of their own accord.

governance BY platforms

Social media platforms are eager to keep the safe harbor protections enshrined in section 230. But at this point, all of them are taking advantage of the second half of its protection: nearly all platforms impose their own rules, and police their sites for offending content and behavior. In fact, their ceaseless and systematic interventions cut much deeper than the law requires. Both in terms of their impact on public discourse, and for the lived experience of its users, the rules these platforms impose themselves probably matter more than the legal restrictions under which they function. So while part of the question must be how platforms are governed, an equally important question is how platforms govern (Citron, 2014; Denardis & Hackl, 2015; Gillespie, 2015; Grimmelmann, 2015; Humphreys, 2013; Jeong, 2015; MacKinnon et. al., 2014; Matias et al, 2015; Obar & Wildman, 2015; Reagle, 2015; Roth, 2015; Stein, 2013; van Dijck, 2013; Wagner, 2013).

There are clear reasons why social media platforms, though not legally required to do so, police the content of their sites and the behavior of their users – mostly economic reasons, though not exclusively so.

[22] Geller, E. (2016). White House and tech companies brainstorm how to slow ISIS propaganda. *The Daily Dot*, January 6. http://www.dailydot.com/politics/white-house-tech-companies-online-extremism-meeting/

Troubling content like pornography and graphic violence may scare off wary advertisers, who are not keen to see their products paired with an X-rated video or a xenophobic rant. Platforms worry about users leaving if they're overwhelmed by porn or trolls. This is especially true as platforms seek to expand their user base: platforms typically begin with users who are more homogenous, share the goal of protecting and nurturing the platform, and can solve some tensions through informal means; as their user base broadens, platforms find themselves hosting users and whole communities with very different value systems, and who look to the platform to police content and resolve disputes. And, the content and behavior users may find perfectly acceptable does not always fit neatly with the platforms' effort to protect its public brand. Revisions of site policies often occur when a new company purchases a platform, and struggles to incorporate its permissive ethos amid its other services.[23] But economic considerations are always intertwined with other kinds: the deeply felt commitment of the platform operators for nurturing a healthy community or encouraging the best creative output of their users; a sense of public obligation, especially as a platform grows and exerts greater influence on the public landscape; and certainly attention to criticisms leveled by angry users, journalists, or activists.

These platforms must constantly police the pornographic, the harassing, and the obscene. There is no avoiding it entirely. But doing so can be a politically fraught exercise, particularly when the politics of visibility is involved (Bakardjieva 2009; Couldry, 2015; Dahlberg 2007; Gray, 2009; Gross, 2002; Thompson 2005). I mean "visibility" in the sense that groups seeking legitimacy struggle to simply be seen against the wishes of those who would marginalized and silence them, such as gay rights or public breastfeeding; visibility in the sense that some kinds of antagonism between groups goes unnoticed or uncommented on, such as the culture of violence against women; visibility in the sense that one group's speech is seen as potentially dangerous to others, as in fundamentalist Islamic propaganda; and visibility in the sense that some kinds of images are seen as potentially dangerous to those who choose to consume them, such as "self-harm" images that may support anorexic, cutting, or suicidal behavior. Sometimes visibility is not just a political accomplishment, but one that must also overcome the mechanics and governance of the medium (Bucher 2012; Milan, 2015; Thompson 2005).

With these unavoidable and perhaps unsolvable contentious politics increasingly inhabiting their sites, social media platforms have not only had to develop and refine their rules, and develop more sophisticated means of policing their sites. They have also had to develop their own logics that underpin how and why they intervene. This is not to suggest that their policies are always conceptually coherent, in principle or in application; most have developed over time, often in an ad hoc fashion, often after having to

[23] Gillespie, T. (2013). Tumblr, NSFW porn blogging, and the challenge of checkpoints. *Culture Digitally*. January 14. http://culturedigitally.org/2013/07/tumblr-nsfw-porn-blogging-and-the-challenge-of-checkpoints/

face a contentious issue they were unprepared for. But out of each site composing this or that rule out of this or that thought process, there are certain kinds of approaches that seem to have coalesced.

where the lines are drawn

Considered together, the guidelines at the prominent, general purpose platforms reveal striking similarities. This should not be surprising, as they encounter many of the same kinds of questionable content and behaviors, often look to each other for guidance on how to address them, and are situated together in a longer history of speech regulation that offers well-worn signposts on how and why to intervene. Most have some rule prohibiting or limiting the following:

- sexual content and pornography
- representations of violence and obscenity
- harassment of other users
- hate speech
- representations of or promotion of self-harm
- representations of or promotion of illegal activity, particularly drug use

Additionally, some platforms have rules about using a "real" identity, or about what can and cannot be done under the cloak of anonymity. Some include advice or pointers for ensuring the smooth working of the site and the quality of its offerings. And some prohibit certain forms of commercial activity and self-promotion. Platforms differ on how they draw each of these lines, what kind of caveats they're willing to consider, and what kinds of consequences are leveled against offenders.

One could dismiss these guidelines as mere window dressing — as a performed statement of coherent values that do not in fact drive the actual enforcement of policy on the site, which can often be more slapdash, strategic, or hypocritical. I find it more convincing to say that these are statements of policy and principle that are struggled over at times, are deployed when they are helpful and can be sidestepped when they're constraining, and that do important discursive work beyond simply guiding enforcement. These guidelines matter, not only when they are enforced, and not only simply to lend strength to the particular norms they represent. Platforms adjust their guidelines in relation to each other, and smaller sites look to the larger ones for guidance, sometimes borrowing language and policies wholesale. They perform, and therefore reveal in oblique ways, how platforms see themselves as public arbiters of cultural value. They are also by no means the end of the story, as no guidelines in the abstract could possibly line up neatly with how they are enforced in practice.

Looking at these guidelines together, it is clear that these platforms develop their rules not just in anticipation of inappropriate content, but in response to it. This can be an internal process, where a content

policy team notices the emergence of a category of content they would like to curtail, or a surge in flags from users. Unanticipated categories of content may be formalized into new guidelines or attached to existing rules. And these adjustments can also come in response from outcries and public controversies, often unexpected ones. In these guidelines, we can see the scars of past challenges.

enforcement and the problem of scale

While the law invited early platforms to enjoy a hands-off safe harbor like ISPs and other conduits, the operators of early platforms were also steeped in the tradition of online community management. In the early days of the web, while ISPs were fending off liability for pornography and copyright infringement, online communities were discovering and addressing the challenges of interpersonal conflict and obscene speech, developing forms of moderation that attempted to protect their community and embody a spirit of governance that they hoped best captured their values and the values of their users. Community management was often the work of volunteers, either the webmaster or site manager, or participants in the communities who took on the role of moderation themselves (Postigo, 2009). Sometimes moderation emerged in response to a shock to the community: the first troll to dramatically disrupt a community that had, perhaps naively, assumed that everyone wanted the same things, and required no governance at all (Dibbell, 1998). Moderation took many forms, perhaps as many forms as there were online communities: from the benevolent tyranny of a webmaster, to public arbitration among the entire community, to ad hoc councils chosen to do the work of determining policies and doling out punishments. As communities grew and changed over time, new members and new conflicts challenged these forms of governance (Bergstrom, 2011; Lampe et. al., 2014; Kerr & Kelleher, 2015; Shaw and Hill, 2014); sometimes they adjusted, sometimes the sites died and people moved on.

Such moderation persists, in the contemporary equivalents of online communities. But the key reason these approaches are ill-suited to social media platforms is scale. The moderation of online communities depended on community members who knew the webmaster, regulars who knew each other, and a history of interactions that provided the familiarity and trust necessary for a moderator to arbitrate between aggrieved parties.[24] Tough cases could be considered together, policies could be weighed and adjusted together. This was the scale of the forum, rather than the *demos* (Forsyth, 2016). Some social media platforms that began at this scale continue to pursue forms of community moderation by retaining some structural form of groups within the platform. For instance, Reddit depends on volunteers to moderate

[24] Many thanks to Kevin Driscoll for this observation.

particular subreddits, and Facebook expects the managers of a Facebook Group to moderate it (although the site will also respond to complaints about Groups just as it does any other kind of content). But as these platforms have grown in scale and ambition, the scale necessary for community moderation has become increasingly untenable.

In addition, on large-scale platforms there is simply too much content and activity being posted to support a proactive review process, where a moderator would examine each contribution before it appeared on the site. Apple is a notable exception, in that it reviews every iPhone app before making it available in their app store; but Apple fields hundreds of submissions a day, not millions. And it has certainly come under fire for failures in judgment, both for apps they rejected and ones they approved (Hestres, 2013). Most other platforms must embrace a "publish-then-filter" (Shirky, 2008) model, meaning enforcement is by necessity reactive rather than proactive.[25] This means that even heinous content may get published, at least briefly, and criminal behavior may occur (and have its intended impact) before anything is done in response. Plenty of content that violates site guidelines remains online for days, or years, because of the sheer challenge of policing platforms as immense as these.

This raises a legal question, in that it is arguably impossible for a platform to assure that no illegal content or behavior will appear there. Section 230 answers this very well; any regime that replaced it would have to grapple with this challenge. And it raises an ethical challenge, in that users cannot avoid obscenity or be protected from harassers with complete certainty. I say it is arguably impossible, because the resources that could be put toward this effort are limited only by convention. We can't imagine a platform employing enough people to review every piece of content before it is posted, but in principle they could; and given current expectations, users would probably be unwilling to accept the delay this would impose on their status updates and shared photos. These constraints are, in fact, movable. China, for instance, employs hundreds of thousands, maybe as many as two million, to scour social media for political criticism, and blocks some websites and keyword searches automatically.[26] I'm not suggesting that this is an ideal approach, only noting that it is not. The political and cultural reality, in the West, is that we accept that platforms cannot review content before it is posted, and we reject the delay that would impose, and yet we also demand that platforms respond quickly and consistently to our complaints.

[25] This excludes automated mechanisms for identifying problematic content. But such automatic filters, at this point, are generally only useful for identifying spam (based on its format and origins), child pornography (based on comparison to a collected database of examples), and profanity (based on simple language identification. Such tools have not yet been successfully extended to the recognition of pornography, hate speech, or harassment.

[26] Hunt, K. and Xu, C. (2013). China 'employs 2 million to police internet'. *CNN*, October 7. http://www.cnn.com/2013/10/07/world/asia/china-internet-monitors/

the labor of content moderation

Large-scale social media platforms have developed intricate and complex systems for conducting content moderation at scale. This requires immense human resources, if not at quite the Chinese scale.[27] These people generally labor in obscurity, some set at a distance from the platforms and its internal aims, and often with little oversight. And each layer of moderation introduces an element of ambiguity and potential bias into what remains a largely opaque process (Roberts, 2016).

At the top, most platforms have an internal policy team charged with overseeing moderation. They set the rules, oversee their enforcement, adjudicate the particularly hard cases, and craft new policies going forward in response. These are, by and large, small teams, often just a handful of full-time employees; sometimes they are an independent division, while in other cases they sit under the umbrella of "trust and safety", "community outreach," customer service, or technical support. These groups are obscure to users, by design and policy. They are difficult for users to reach, and the statements and policy changes they generate are often released in the voice of the company itself. Together they are a very small community of people, based overwhelmingly in the San Francisco area, and individuals tend to move from platform to platform in their career.[28] At the scale at which most platforms operate, these internal teams would be insufficient by themselves, but they have an outsized influence on where the lines are drawn, what kinds of punishments are enforced, and the philosophical approach the team and the platforms are to governance itself.

At many companies, there is a substantially larger group of people who provide a frontline review of specific content and incidents beneath the internal moderation team. These might be employees of the platform, at the home office or in satellite offices around the world, but in more and more cases they are employed on a contract or freelance basis by the platforms: hired as independent contractors through third-party "temp" companies, or clickworkers employed through crowdwork services such as Amazon's Mechanical Turk, Upwork, or TaskUs -- or both, in a two-tiered system.[29] These clickworkers are obscure by circumstance. Many work outside of the United States in places with cheaper labor, especially the Philippines

[27] Chen, A. (2014). The Laborers Who Keep Dick Pics and beheadings out of Your Facebook Feed. *Wired*, October 23. http://www.wired.com/2014/10/content-moderation/

[28] Rosen, J. (2013). The Delete Squad: Google, Twitter, Facebook and the new global battle over the future of free speech. *New Republic*, April 29. https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules

[29] Chen, *op. cit.* Buni, C. and Chemaly, S. (2016). The Secret Rules of the Internet: The murky history of moderation, and how it's shaping the future of free speech. *The Verge*, April 13. http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech

and India, distanced from the platform and the users they are moderating; they are also distanced from the company through contract labor arrangements and the intervening interfaces of the crowdwork platforms.[30]

These clickworkers are used as a first response team, fielding complaints from users and making quick decisions as to how to respond. Quick can mean seconds per complaint, which means each user is getting very little attention, and clickworker moderators are facing a torrent of atrocities.[31] Most complaints are now fielded as a matter of course; only the most difficult to judge are directed up to the internal content team for further deliberation. Moderators are generally unaware of the identity of the user, and given very little context from which to reach their decision. Decoupling the images and posts in question from the site and its context, and demanding that review happen at breakneck speed, serves as one solution to the problem of scale, and makes it possible for platforms to respond so quickly, often in less than 24 hours. Many are concerned about the psychological toll of this work.[32]

As previously mentioned, if the platform is structured in a way conducive to it, some sites continue to depend on community moderators. Reddit, for instance, has smaller, persistent "subreddits" that exist within a larger platform, each overseen by moderators, and Wikipedia has an "administrator" class of superusers who do other kinds of back-end work on the platform. These community moderators are usually volunteers and thus independent of the platform itself, with their own commitment to the site or group for which they are responsible. In most cases, these moderators are given tools by the platform that make possible the enforcement of group policies, such as the ability to delete content or suspend users. But the approaches vary by site and by individual moderator. The people responsible for moderating groups may be expert or ignorant at the dynamics of doing so, even-handed or tyrannical. They are often people invested in the group's success, though this can be an asset or a detriment. That commitment often appears as a rigid adherence to first principles and an unwillingness to adapt to new members, new uses, and new challenges. In a tenuous position in relationship to the platform, such moderators are often overworked but undersupported, provided a weak mandate for the role they're expected to play, sometimes invited backstage and sometimes held at arms-length.[33]

Platforms where small-scale moderation is impossible generally turn to the users to assist in the policing of the site. This comes in two forms. Either the site invites users to flag problematic content and behavior, thus turning those complaints over to the platform (and its army of moderators and clickworkers)

---

[30] Gray, M. (2016). Your job is about to get 'taskified'. *Los Angeles Times*, January 8. http://www.latimes.com/opinion/op-ed/la-oe-0110-digital-turk-work-20160110-story.html

[31] *ibid.*

[32] Stone, B. (2010). Policing the Web's Lurid Precincts. *New York Times*, July 18. Chen, 2014, *op. cit.*

[33] Matias, J. N. (2015). What Just Happened on Reddit? Understanding The Moderator Blackout. *Social Media Collective*, July 9. https://socialmediacollective.org/2015/07/09/what-just-happened-on-reddit-understanding-the-moderator-blackout/

to adjudicate, or the site gives users tools to rate and block content, and designs mechanisms into the system to filter content towards those who want to see it, and away from those who don't. The implications of these two models are important, and I will address them in a moment. But in terms of a labor force, both depend on the crowd to police itself, though the populations on which they depend and the dilemmas each must grapple with are different.

Flagging is now widespread across social media platforms. A small icon or link beneath a post, image, or video offers the user a pull-down menu facilitating their complaint, often with submenus to classify the nature of the offense. In the earliest days of contemporary platforms, such mechanisms were unavailable, or buried in the help pages; in recent years, platforms have made it easier and easier to find these flags, though in some cases criticism has dogged specific platforms for inadequate mechanisms. On one hand, flagging puts the work of complaining right at the point of offense, in front of those most motivated to complain. On the other hand, it is optional. This means that the population of users who deputize themselves to flag content are those most motivated to do so, which raises certain questions. What motivates them, and how well do their values and concerns align with that of the larger user community? Are they in harm's way, or are they politically motivated to try to get something removed from the site? Are they acting on behalf of and in concert with the site policies, or are they articulating their own gut response? The flag is a thin form of expression, leaving the platform with only a vague sense of the nature and motivation of the complaint – and this may be to the platform's advantage. And, although platforms have data on who flags, to what degree, and what proportion of the entire user base they represent, this is not information they readily share (Crawford and Gillespie 2014).

Recently, some platforms have experimented with granting some users the status of "superflagger," prioritizing their flags over others. These users might be law enforcement organizations, activist organizations concerned with a specific kind of violation or protecting a specific population of users, or long-time users who are recognized as reliable. Generally, who they are remains opaque to users. While platforms can gain insight over time into the reliability and evenhandedness of a particular group of flaggers, this can be extremely taxing on the people and groups who sign on to play this role. (Matias *et. al.*, 2015)

The alternative is for platforms to ask users to rate their own content when it is first posted, and then provide filtering mechanisms so that users can avoid content they want to avoid. Unlike flagging, this enlists all users, which distributes the work more equitably and diminishes the concern that those doing the flagging do not represent the whole community. The challenge with this approach is achieving full participation and consistency. Platforms are wary about introducing too many steps at the moment a user posts, worried that an unwieldy and multi-click interface could discourage participation. So the rating process must either be lean and depend heavily on defaults, or it must happen less often. On Tumblr, for example, users are asked

to rate their entire blog, rather than each post, and the default rating is "safe". While this makes the interface quite simple, the rating can only serve as a blunt instrument: a Tumblr user who rarely posts risqué content and a user who regularly posts pornography are rated the same, as “NSFW” (Not Safe For Work). Users will inevitably have different interpretations of what is "adult" or "violent" or "not for children," especially regarding their own posts, leaving the platform with limited means for ensuring consistency across users. Many platforms penalize users for failing to rate adult material, or for doing so in ways that wildly differs from the platform moderator's opinion of it, or from users who come across it.

These layers of laborers, from the internal team setting the rules and adjudicating the hardest cases, to clickworkers reviewing each bit of flagged content, to volunteer moderators overseeing groups within a platform, to flaggers lodging their complaints, to all users enlisted to rate content, represent a set of tiered solutions to the problem of scale. Many platforms use a combination of some or all, and the workflow that moves questionable content from one tier to the next may differ. But it leaves us with (at least) three levels of vagary about the process and the possible biases, consequences, and side effects that might intrude along the way: who rates and flags and according to what criteria; who does the first line of review and how do they adjudicate different kinds of content; and who are these internal teams and how do they guide this system of governance all the way down. Many of the concerns about these systems of platform governance live in the uncertainties of this multi-tiered system. And they breed in the shadow of a process that remains distinctly opaque to public scrutiny.

to remove or to filter

Platforms also have two choices for what to do with offending content: remove it, or mark it as such and help users avoid it. In practice, most platforms do some combination of both. Even permissive platforms remove the most heinous and illegal material, and even the most sensitive platforms often have a category of content that has an age warning or rating. The difference tends to be where the balance is drawn between the two approaches, and how that balance is justified. But the two approaches have different implications as forms of governance.

It is commonplace for platforms to remove content deemed offensive and users deemed harassing. The advantages of this approach are numerous: content that offends one user is likely to offend others, so the removal addresses multiple points of offense; with it gone, it cannot offend again. Publicly, removal demonstrates a decisive commitment to protection, allows the platform to celebrate that it does not tolerate such content or behavior, and avoids associating the company brand with something offensive. And removal saves human resources later, having to adjudicate on the same content or user down the road.

On the other hand, removal is a blunt instrument, an all-or-nothing determination. It removes that content for everyone, not just the ones who were offended. It runs counter to the principles promised by so many platforms: open participation, unencumbered interaction, and the protection of speech. And removing content from a platform altogether represents a deeper cut in terms of the protection of speech. The U.S. tradition of First Amendment jurisprudence has long established that preempting speech entirely is a more problematic intervention than imposing penalties for it after the fact, because it silences that speech in the process (Armijo, 2013; Balkin, 2014; Meyerson, 2001). Removing users does more than limit speech, it interrupts their ability to participate on that platform, and removes all of their future speech as well. At the same time, users whose content is deleted or account is suspended often simply create a new profile and post again, leading platform governance into an endless game of whack-a-mole, where content reappears in slight variation, under new names, or from dummy accounts that can only be identified in hindsight.

Removals can also feel like—or be criticized as—a judgement of the user themselves. While the platform may have merely determined that the content in question was statistically similar to other deleted content, the person who posted it may feel that its deletion is a judgment of them. This is especially problematic when what was deleted was, from the user's perspective, a positive expression of themselves: for example, when users, especially women, post pictures that, while they do expose their bodies, represent a moment of physical triumph – giving birth, breastfeeding their newborn, surviving a mastectomy or other surgery – only to have those photos deleted as inappropriate or pornographic.[34]

Finally, because removal is so blunt an instrument, it opens platforms up to charges of subjectivity, hypocrisy, political conservatism, and self-interest. For the largest platforms, content moderation will never be complete or consistent, which means that any user who feels their deleted post or image was fine can easily find content still on the site that they think is worse. Explanations for why that more egregious content remains rarely give the platform the benefit of the doubt: in that inconsistency, aggrieved users see subjectivity, hypocrisy, and bias. And removals that seem to benefit the platform in some way can look self-interested -- or, to put it less generously, platforms may remove content they want to do away with under the guise of content moderation, on behalf of their community.

Allowing obscene content or problematic users to stay, but rating them so that users who care not to encounter them will be automatically filtered away, is arguably a less invasive approach. It allows platforms to proclaim their commitment to protecting the speech of their users, though it also opens them to criticism, ranging from being too permissive to harboring pornographers and terrorists. If the right balance is struck,

[34] Collins, P. (2013). Why Instagram Censored My Body. *Huffington* Post, October 17. http://www.huffingtonpost.com/petra-collins/why-instagram-censored-my-body_b_4118416.html; Peters, L. (2014).What you need to know about Facebook and Instagram's war on motherhood. *The Daily* Dot, July 15. http://www.dailydot.com/opinion/why-we-need-stop-censoring-motherhood/

the platform can enjoy the traffic and revenue both generated by users seeking illicit content, and by users who want a "clean" experience of the platform. For permissive platforms that have developed a sturdy community around adult interests, or pride themselves on allowing unfettered debate, or position themselves as hands-off when it comes to what users do, it offers a mode of governance aligned with these aims.

This is not unlike how adult content has sometimes been handled before: the adult movies in the back room of the video store, the magazines on the top shelf at the newsstand, the pornographic cable channels encrypted. But instead of a cashier looking at a driver's license at the point of sale, social media platforms must patrol users algorithmically. Some form of "safesearch" mechanism must recognize which users are in a safe mode and refuse to deliver to them the unsafe material. Instead of blocking content at the point of sale, they're blocked at the point of search, which means the very same mechanism we expect to help us find content is also being used by the platform to prevent us from finding it. This can have real cultural and political consequences, like when Tumblr blocked the term "#gay" because it is commonly associated with pornographic images, and thereby blocked all other non-pornographic content similar tagged.[35]

This kind of technical choreography can be harder for users to see, and are not always where one might expect it. For instance, all of the major search engines allow users to opt into some form of a safesearch -- if I don't want to see pornographic results, I tick a box and never receive them. But the major search engines go one step further. Even if I am not in safe mode, i.e. I am consenting to potentially receive explicit links, and I then conduct a search that the site deems to be not adult, like "movies" or "toys," it will deliver only non-explicit results, using the same algorithmic delineation as if I were in safe mode. The reasoning is that, given my generic search, I probably don't want links to adult movies or sex toys. Reasonable, but the intervention is a hidden one, and in fact runs counter to my stated preferences.

Finally, these algorithmic approaches can offer the platform a compelling response to the legal demands of specific nations, but in ways that may differ from the intent of the law in question. For instance, when Germany, Singapore, and South Korea all complained to Flickr that, by allowing explicit content so long as the user rates it as such, it was violating their laws restricting access to pornography for minors, Flickr responded, not by designing an age barrier, but by making it such that users from those countries can only use the site in safe mode. In other words, laws protecting minors become technical measures restricting adults. When the government of Pakistan complained to Facebook about a particular Facebook group encouraging people to draw the image of Mohammed, Facebook removed the group only from the search

[35] Baker-Whitelaw, G. (2013). New NSFW content restrictions enrage Tumblr users. *The Daily Dot*, July 18. http://www.dailydot.com/lifestyle/tumblr-nsfw-content-tags-search/

results of users located in Pakistan. For those users, the offending page was simply not there; even its removal was invisible. Technical measures that keep some users away from some content, while continuing to display it to others, are a convenient solution, but raise troubling questions about the power of social media platforms to offer different media to different publics, in ways that are hard to discern or criticize.

conclusion: the question of responsibility

We tend to defend platforms as free conduits of speech until we are too troubled by something that freely moves through their system. When the government, or the aggrieved user, or the culture at large, demands that the platform "do something" about the problem, that request generally lies somewhere between a genuine belief in the platform's responsibility and the practicality of looking to them to intervene. It may be a convenience or a strategy: platforms do have the means to intervene in the circulation of abhorrent content and at the moment of abhorrent behavior. Chasing individual bad actors is difficult, consumes time and resources, and makes little impact: getting a platform to intervene systematically promises to have a much broader impact. But platforms also make human behavior highly visible, leading to what Mueller calls a "fallacy of displaced control": since the problem is most obvious there, we tend to assign blame to the platform itself. When this comes in the form of a legal imposition, it can appear to some as a displacement of accountability: "Instead of punishing bad behavior, we strive to control the tool that was used by the bad actor(s). Instead of eliminating illegal materials or activities, we propose to eliminate internet access to illegal materials or activities." (Mueller 2015, 807) Platforms get the burden and the blame for what users say and do.

But, in principle, there might be reason to think of platforms as bearing some responsibility. Copyright law point to at least two ways in which intermediaries could be held responsible for the activity of their users. First, if they gain financially from the illicit transaction, and second, if they have some material effect on the transaction, making it easier or expanding its scope.[36] In the case of offensive content and behavior, similar questions could be asked: does a platform pair advertising with offensive content? Do they materially enhance the ability of a bad actor to harass or threaten another user?

Then we might consider other versions of platform responsibility as well. Are they responsible because, by their very existence, they connect people who would not be connected otherwise? Do they have

[36] U.S. copyright jurisprudence added a third, in the *MGM v. Grokster* decision, where the court held Grokster liable because it encouraged or "induced" copyright infringement by advertising how easy it was to attain pirated music and movies through their software.

a responsibility, though this runs counter to Section 230, once they make a promise to intervene in particular ways? Do they have a greater responsibility as they grow larger, as they displace other central venues of public life, or if they gain monopolistic power in a particular genre of services? Some of these questions depend on the role we think platforms play in shaping the activity and discourse they host, materially and institutionally. Some depend on what kinds of public obligations we are willing to impose on private institutions of any kind.

This question must be split into two because, as I have argued, there are two phases of governance in question, the governance of platforms and the governance by them. One question revolves around legal mandates requiring platforms to intervene. A second question is what kind of responsibility do platforms once they begin to moderate users and content on their own accord: how are they accountable for how they do so, or for how they meet whatever promises they make about their larger public role.

While Section 230 may have tried to provide both sides of a safe harbor from liability for user content and behavior – safe harbor from being held accountable for it, and the legal freedom to intervene on users' behalf without being then held accountable for how extensively they do so – platforms are in some ways hamstrung between these two positions. They are indeed intermediaries, stuck in the middle in both the legal and practical sense: halfway between users with different values, halfway between policymakers and the people they seek to regulate, halfway between a conduit and a curator, and halfway between an array of internal aims and external demands. But they also get to play both sides, where they enjoy all the right to intervene, but with little responsibility about how they do so and under what forms of oversight.

The language of the impartial conduit is still powerful, though it seems to be diminishing in the glare of the most alarming content and egregious behavior being circulated through and perpetrated on these platforms, and in light of the different legal approaches around the world. Even in the West, with a robust safe harbor principle, we oblige platforms to remove illegal content like child pornography, and are considering at other kind of obligations, such as revenge porn and extremist content – governments and publics are not only willing to make exceptions, they are beginning to reconsider their starting assumptions about whether platforms should be responsible for what happens on them. And users, faced with direct harms coming at them on their chosen platform, quickly adjust their understanding of that platform, from an unfettered space in which to play, to a responsible guardian failing to intervene. These are not just questions about the proper legal rules for intermediaries, they are broader societal questions about how bad something has to be to justify adjusting a general principle. This question falls heavily on platforms, sometimes asked not to intervene, sometimes required to, and shapes how we think about how to govern them, and how they govern.

In addition, the policies of the major social media platforms have themselves become a terrain for longstanding debates about the content and character of public discourse. That our dilemmas about terrorism and Islamic fundamentalism, about gay sexuality, about misogyny and violence against women, each so heightened over the last decade, should erupt here too is not surprising. The controversies these sites face can be read as a barometer of our society's pressing concerns about public discourse more broadly: which representations of sexuality are empowering and which are explicit, and according to whose judgement; what is newsworthy and what is gruesome, and who draws the line; how do we balance freedom of speech with the values of the community, with the safety of individuals, with the aspirations of art, and with the wants of commerce.

For both reasons, it is high time to reconsider the responsibilities of platforms. This should include reasserting a principle of safe harbor tailored for social media platforms, not borrowed whole cloth from a law designed for ISPs and search engines. It should include articulating positive expectations for what platforms are -- legally, culturally, and ethically – such they platforms can shift their fundamental approach: from being nominally impartial conduits that in fact intervene, to being the architects of public spaces of discourse, that depend on specific rules of play that they then obviously have the right to enforce. And it should include a new standard of transparency and accountability for how they do so: more information about the inner workings of the moderation process, more data provided about who flags and how those complaints are adjudicated, more transparency about the labor forces involved, and more public accountability about how and why the rules are made.