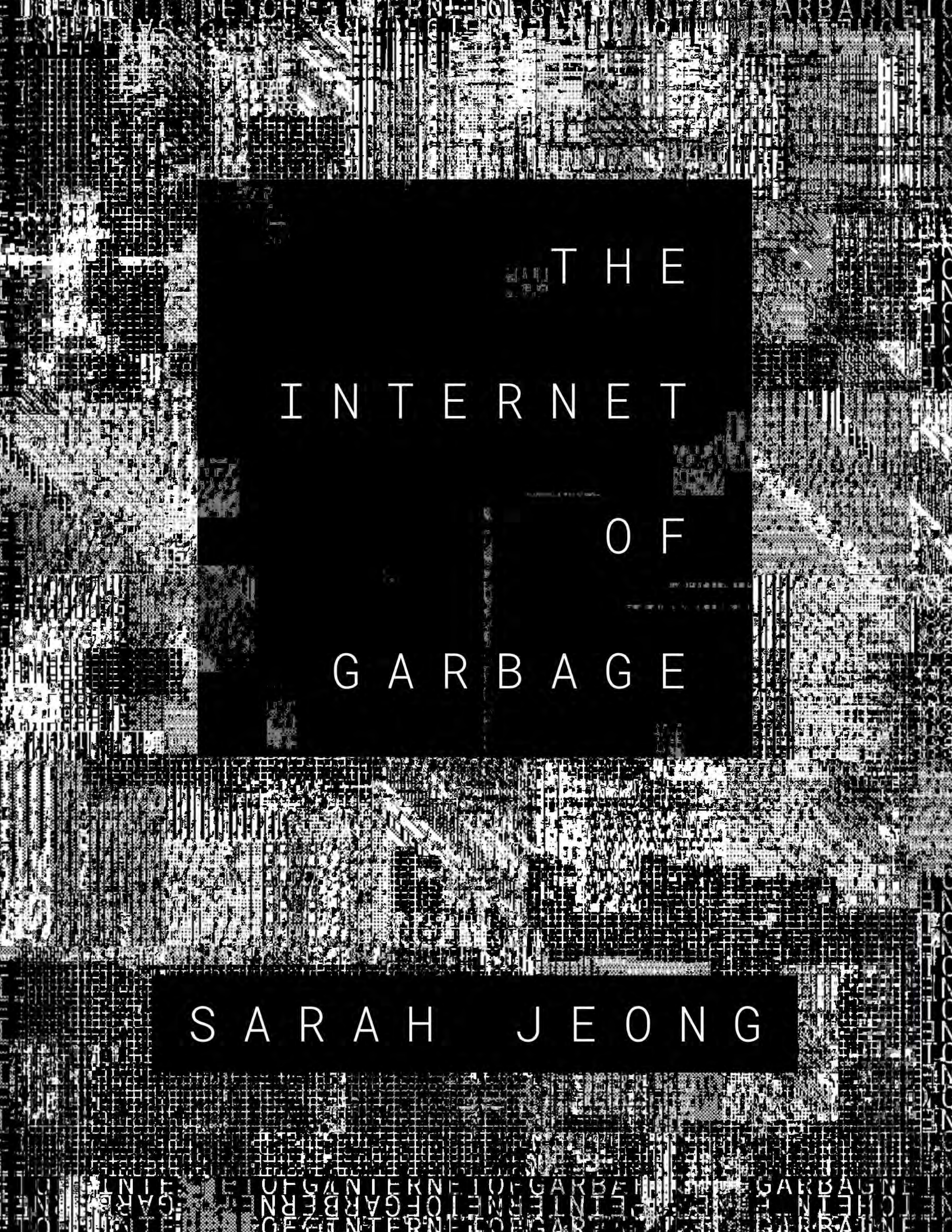


THE INTERNET OF GARBAGE

SARAH JEONG

The Internet of Garbage

Cover and Illustrations by William Joel for *The Verge* , © 2018 Vox Media, Inc.

The Verge Edition 1.5 August 2018

Published by Vox Media, Inc.
www.theverge.com

ISBN: 978-0-692-18121-8

Table of Contents

Preface

In October 2016, my friend Lana Swartz invited me to be a guest speaker in her media studies class at the University of Virginia. *The Internet of Garbage* had been assigned as reading that week for her students, who were mostly undergraduate juniors. I talked to them for a little bit about internet harassment and spam, but at some point, I let loose with what had been weighing on my mind for months. "If I rewrote *The Internet of Garbage*," I told them, "it wouldn't focus on online harassment at all. It would focus on fake news."

Since October 2016, the term "fake news" has been co-opted and stretched beyond recognizability. It was meant to describe a category of viral falsehood churned out and advertised on social networks and search engines to profit from ad dollars—a disturbing phenomenon that some have sought to link to foreign propaganda efforts and election interference. "Fake news" is now the catchphrase of those who would undermine the legitimacy and freedom of the press, a refrain similar to the cry of "*Lügenpresse*" in years past.

But in October 2016, all that was yet to happen. I was fixated on how search engine optimization and ad dollars had fueled an economy of lies. I was particularly glued to the virality of alternate histories—for example, the "Irish slaves" meme, which conflates indentured servitude with chattel slavery by reconstructing an alternate history made of misinterpreted or made-up sources.

I told the UVA class that this had less to do with something evil or broken about the internet and more to do with our collective inability to come to grips with the American legacy of slavery and white supremacy. At the time, I didn't have any answers for what to do about literal fake news or white supremacist mythologies. Two years later, I'm still thinking hard about it.

The year after my guest appearance in that class, two Portlanders were killed on the train stop closest to where I lived, stabbed by a white nationalist who had been harassing two girls, one of whom was in hijab. Months later, Charlottesville—the home of the University of Virginia—was convulsed with white supremacist violence, resulting in the killing of Heather Heyer.

I published *The Internet of Garbage* in 2015, and in its original form, it stands as a testament to an internet that used to be—one that flickered out of existence as platforms solidified into monopolistic silos and the fragile American pact of free speech shattered in the wake of the 2016 election.

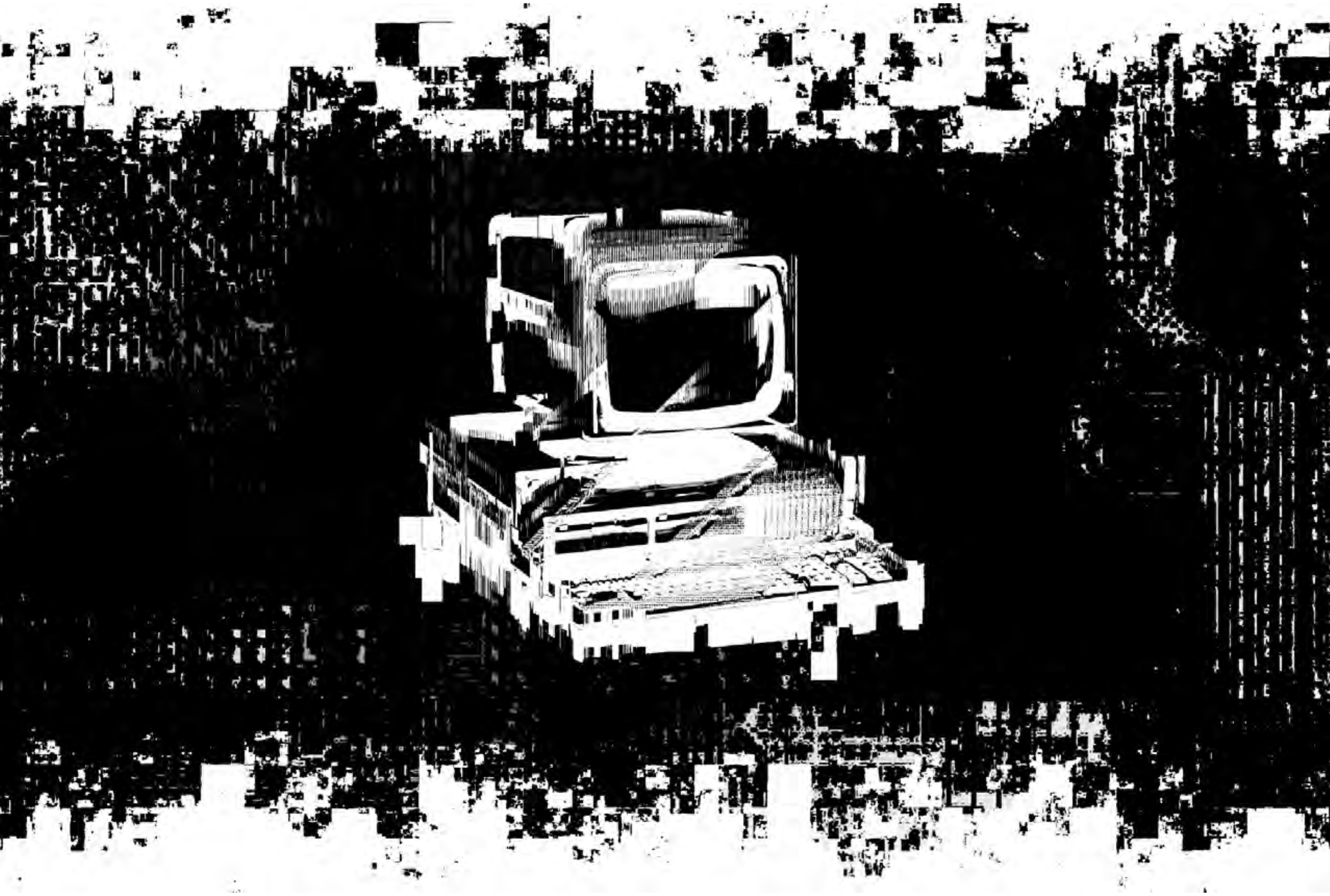
I described the marketplace of ideas in my book, criticizing it as something that has never truly existed for everyone across race and gender. But I did not criticize it enough. In 2018, the marketplace of ideas is sick unto death, close to being replaced by an SEO nightmare populated with the intellectual equivalents of malware, viruses, and banking scams.

I think *The Internet of Garbage* still provides a useful framework to *begin* to talk about our new dystopia, and it continues to be surprisingly relevant in many ways. But I wrote the book with a tone of optimism I did not feel even at the time, hoping that by reaching the well-meaning policy teams across Silicon Valley, I might be able to spark change for the better.

Not only did that change never quite solidify, but the coordinated, orchestrated harassment campaigns of Gamergate that I very briefly touch on in Chapter Two have since overtaken our national political and cultural conversations. These twisted knots of lies, deflection, and rage are not just some weird and terrible online garbage. They shadow executive orders, court rulings, even the newly appointed judiciary. They will haunt us for years to come. We are all victims of fraud in the marketplace of ideas.

I hope that in the very near future, I will be putting out a second edition of *The Internet of Garbage.* In that future edition, I hope to grapple with advertising incentives, engagement traps, international propaganda wars, the American crisis in free speech coinciding with the rise of platform power, and search engine optimization as the new paradigm of speech.

In the meantime, I am putting out *The Internet of Garbage 1.5* as an interim edition. I wish it were more helpful in our present reality. But as imperfect a tool as it is, I figure we all need as much help as we can get.



Chapter One:

The Internet Is Garbage

Introduction

Contract workers in San Francisco, processing thousands of complaints a day. Sweatshops in the Philippines, where outsourced labor decides what's obscene and what's permissible in a matter of seconds. Teams of anti-spam engineers in Mountain View, adapting to the latest wave of bots. An unpaid moderator on Reddit, picking out submissions that violate guidelines.

So much of the internet is garbage, and much of its infrastructure and many work hours are devoted to taking out the garbage. For the most part, this labor is hidden from plain sight. But in recent years, the garbage disposal has broken down. The social media companies have a harassment problem, the pundits have declared.

However, large-scale harassment campaigns are hardly new, and the barrage of crude and cruel messages and undesirable content is only a small part of what makes a targeted campaign a frightening experience for the victim. Yet this part of the equation—the part that is seemingly under the control of Silicon Valley—has received the most attention from the media because it is the most public, visible, and archivable. And as tech companies repeatedly fail to address the problem to everyone's liking, the problem looms ever larger in the public imagination.

The public's understanding of speech online has undergone a serious paradigm shift. Even in tech-centric communities that are generally supportive of "free speech" on the internet, there is a pervasive feeling that harassment must be rooted out and solved. "Anonymity" and "freedom of speech" have become bad words, the catchphrases of an old guard that refuses to open its eyes to a crisis for the internet.

But is there really a crisis, and if so, what is its nature? If the internet is under threat, it is, in essence, under the same threat it's been under from its inception. The internet isn't breaking. Beneath the Wikipedias and Facebooks and YouTubes and other shiny repositories of information, community, and culture, the internet is, and always has been, mostly garbage.

A Theory of Garbage

What do I mean by garbage?

It’s a broad category, one whose boundaries are highly dependent on context. The definition shifts from platform to platform, from year to year, even from week to week.

Garbage is simply undesirable content. It might be content meant to break the code of a site. It might be malware. It might be spam in the specific sense of robotically generated commercial text. It might be a “specific threat” directed toward another user. It might be a vague threat. Or it might be a post sprinkled with a few too many four-letter words. In heavily moderated communities, posts that are deemed to be merely off-topic may be deleted. Posts that might be neither frightening nor offensive nor off-topic can also be deemed to be garbage. On the SomethingAwful forums, postings that are judged to have little value to the community are referred to by the evocative name “shitpost.”

Even in the most anarchic of spaces, there will be content classified as garbage. On 4chan, a site with a reputation for permitting “anything,” “doxing” (posting addresses or other personal information without permission) and “forum raids” (orchestrating a campaign of vandalism or harassment on another site) are forbidden. On the Silk Road, once a Tor-hidden service that allowed people to buy and sell drugs, listings for guns were forbidden. On both sites, child pornography is and was forbidden.

No matter how libertarian, how permissive, and how illegal a site is, there is always content that is deemed to be unworthy of staying on the site. It must be deleted. Perhaps it is because the content is illegal (e.g., child pornography). Perhaps it is dangerous to other users (e.g., malware). And perhaps it simply does not comport with the mission statement of the community—that is, it derails from the purposes and goals of the platform. Whether it is primarily a community of like-minded people (bulletin boards, forums, and mailing lists) or primarily a profit-driven social media company (Facebook, Twitter, Instagram), there is always some content that makes the platform simply *less good*. The standards for “good” differ, but nonetheless, the rule is the same. Some content contributes value, other content detracts. Each corner of the internet works actively to discourage or weed out the trash—otherwise the garbage will choke it up and kill it.

Spam as Garbage

Today, nothing is as uncontroversially garbage as spam. Yet the definition of spam is nebulous. In 1975, Jon Postel, a computer scientist with so much control over the early internet and held in such high regard that he was known as the "God of the Internet" before his death in 1998, wrote the document "On the Junk Mail Problem." The "problem" that the RFC 706 document discussed was actually speculative in nature, and the "junk mail" was described as "undesired," "misbehaving" or "simply annoying" material.

The term "spam," which today is interchangeable with "junk mail," started off as simply denoting annoying behavior. On early chat systems, people would often type "spam, spam, spam, spammity spam"—a reference to a *Monty Python* sketch where a couple's breakfast is repeatedly interrupted by Vikings singing about Spam. If users hit the up arrow, the line would replicate itself, and they could then "spam" the chat room repeatedly with very little effort. Finn Brunton, professor of media, culture and communication at New York University, writes in *Spam: A Shadow History of the Internet*, "In the bandwidth-constrained, text-only space, as you followed exchanges line by line on a monochrome monitor, this was a powerful tool for annoying people. You could push all the rest of the conversation up off the screen, cutting off the other users and dominating that painfully slow connection. . . . The word 'spam' served to identify a way of thinking and doing online that was lazy, indiscriminate, and a waste of the time and attention of others."

In early years, in fact, it wasn't all that clear that spam should be proactively deleted or filtered. The debate was framed as a free speech issue, where the most libertarian standpoint, according to Brunton, was one that permitted "any speech except that which actively interferes with Usenet's ability to function—that is, that which would restrict the speech of others." For Brunton, this is summarized best in Dave Hayes' 1996 "An Alternative Primer on Net Abuse, Free Speech, and Usenet":

> *Example of Net Abuse:*
>
> - *Posting articles that directly crash the news server that is to inject the post into the news stream.*
> - *Posting articles that contain control messages designed to crash news servers.*
> - *Directly hacking into a news server to disable it.*
>
> *Examples of Things that Are Not Net Abuse:*

- *Volumnous [sic] posting*
- *SPAM*
- *Excessive cross-posting*
- *Off-topic posting*
- *Flaming or arguing*

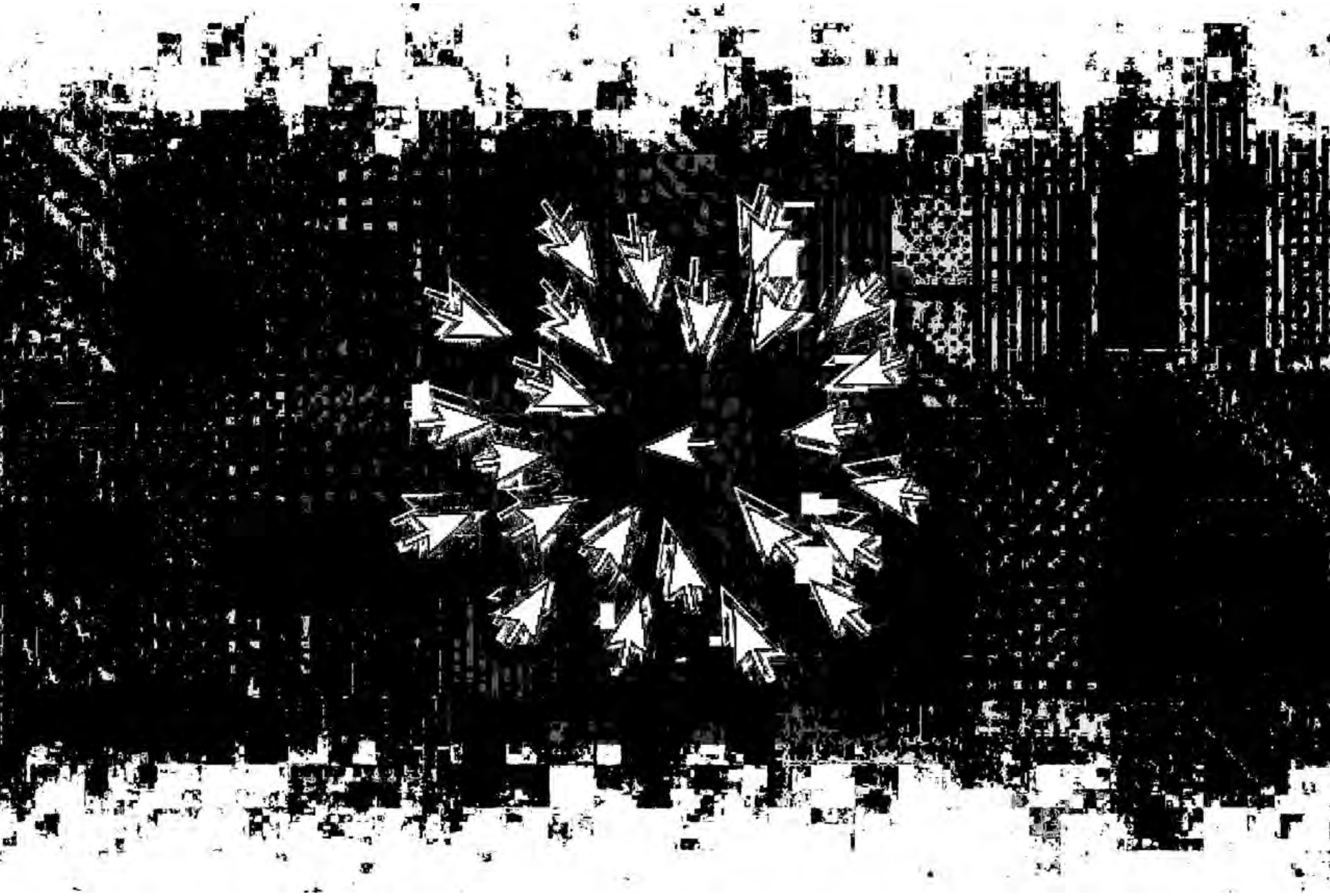
By the 2000s, commercial spammers like Laura Betterly would defend themselves on the basis that sending out millions of unsolicited messages was "what America was built on. Small business owners have a right to direct marketing." In fact, they would characterize the vitriolic backlash against spam as coming from "hate groups that are trying to shut down commercial email."

But today spam is largely understood as robotically generated text issued from "botnets" comprising computers that have been unknowingly recruited into transmitting mind-bogglingly large amounts of unwanted messages advertising Viagra, genital enhancements, and Nigerian get-rich-quick schemes, or linking to malware in order to steal passwords or simply recruit yet another computer into the mechanical zombie horde. Spam has become the realm of Russian crime rings (as documented by Brian Krebs in many places, including his book *Spam Nation*), a multimillion-dollar industry that is combated in turn by *billions* of dollars in anti-spam technology.

Of course, the old definition of spam still lingers. For example, someone might be chided for "spamming a mailing list" when they are not a robot attempting to evade a filter, nor a commercial mailer advertising a product or a service. But by and by, spam is beginning to have a relatively narrow definition. It is the stuff that lands in the spam filter that you want in the spam filter, the garbled poetry text from strange addresses full of misspelled words and suspicious links.

The deep ambiguity in the word "spam" in the early days echoes how nebulous the word "harassment" is today. While the media focuses on discrete, uncomplicated campaigns of hatred against women like Caroline Criado Perez in 2013, Anita Sarkeesian in 2012, or Kathy Sierra in 2007, the worst harassment often occurs in deeply complicated circumstances. When complex internet pile-ons like Gamergate get heated, the term "harassment" is flung back and forth like an accusation, with each side convinced that the other side is the real harasser, and that their opponent is now using the term in bad faith to apply to mere criticisms or mildly unpleasant language.

I don't mean to say that there is no such thing as harassment, no more than I believe there is no such thing as intimate partner violence, even though it is common for domestic abusers to claim that their victims are the ones who are abusing *them*. But the word, when used, is often not grounded with any specificity. As “spam” once was, it merely means an undesirable message. We are in the early days of understanding “harassment” as a subcategory of garbage. Just like spam used to be a catchall for many different kinds of garbage, harassment, too, has become a catchall. But if we are to fight it, the definition must be improved.



Chapter Two:

On Harassment

Harassment in the News

With international attention on the mass of Twitter threats sent to Caroline Criado Perez in 2013 (and the later prosecution of some of the people who sent her those threats), and increasing media coverage of other incidents, “harassment” is a word that is bandied about with increasing frequency. The word remains poorly defined, but it is generally understood in relation to the following high-profile campaigns.

Caroline Criado Perez

In 2013, Caroline Criado Perez called for a woman to be featured on an English banknote. She won, resulting in Jane Austen replacing Charles Darwin on the tenner. Then began the abuse. According to Criado Perez, 50 tweets an hour were being hurled at her, including rape threats. Hot on the heels of one big media story (the banknote change), this online firestorm received massive attention in the news. Two users were imprisoned over the affair. They had tweeted things like:

- "f*** off and die you worthless piece of crap"
- "go kill yourself"
- "Rape is the last of your worries"
- "shut up bitch"
- "Ya not that gd looking to rape u be fine"
- "I will find you [smiley face]"
- "rape her nice ass"

Anita Sarkeesian

Anita Sarkeesian is a cultural critic who successfully raised money through a crowdfunding campaign to do a series of YouTube videos about sexist tropes in video games. Along with a positive reception, there was also a prolonged and clearly disproportionate backlash, which Adi Robertson at *The Verge* described as “an incessant, deeply paranoid campaign against Tropes vs. Women generally and Sarkeesian personally. This includes a flood of violent comments and emails, videos documenting ways in which she's not a ‘real gamer,’ a game in which you can punch her in the face, and a proposed documentary devoted to exposing the ‘lies’ and ‘campaign of misinformation’ from what is, again, a collection of opinions about video games.”

Sarkeesian documented the comments she received on YouTube. The following is a very small selection of a very large number of comments that have been selected at random:

- “I hate ovaries with a brain big enough to post videos.”
- “Fun aside, she completely forgot to mention that every guy in video games has these stereotypes too. Do you see us parading about it? No honey, it’s a video game.”
- “tits or get the fuck out.”
- “Yeah, I can’t wait for the day we get to play “ugly feminist ham planet: the game” That would sell millions of units.”
- “ask money for making a fucking blog? and you made it in a way that women should pledge for not being dominated by man. Smart and evil plan. you are the reason why women’s are the inferior gender for the whole history of mankind”
- “back to the kitchen, cunt”
- “what a stuck up bitch. I hope all them people who gave her money get raped and die of cancer”
- “Ahahahahahahha you stupid IDIOT!!!!!!”
- “I would give her money to STOP making videos. She sucks the joy out of everything, and has this perpetual ‘smelling a fart’ miserable look on her face.”

And amid that flood:

- “I am okay with this and I believe everyone is too. Case dismissed.”

Sarkeesian wrote:

> *In addition to the torrent of misogyny and hate left on my YouTube video (see below) the intimidation effort has also included repeated vandalizing of the Wikipedia page about me (with porn), organized efforts to flag my YouTube videos as “terrorism”, as well as many threatening messages sent through Twitter, Facebook, Kickstarter, email and my own website. These messages and comments have included everything from the typical sandwich and kitchen “jokes” to threats of violence, death, sexual assault and rape. All that plus an organized attempt to report this project to Kickstarter and get it banned or defunded.*

That was in 2012. In August 2014, she would be forced to flee her home after receiving a threat. In October, she canceled a lecture at Utah State University after someone sent a message to the university saying that they would commit “the deadliest school shooting in American history” if Sarkeesian was allowed to speak.

Amanda Hess

Amanda Hess, a journalist who often writes about feminism and women's issues, published a personal essay in *Pacific Standard* magazine about the online abuse she faced. The essay received a high amount of publicity and Hess received an award for the piece. In one example of the harassment she received

> *. . . someone going by the username "headlessfemalepig" had sent me seven tweets. "I see you are physically not very attractive. Figured," the first said. Then: "You suck a lot of drunk and drug fucked guys cocks." As a female journalist who writes about sex (among other things), none of this feedback was particularly out of the ordinary. But this guy took it to another level: "I am 36 years old, I did 12 years for 'manslaughter', I killed a woman, like you, who decided to make fun of guys cocks." And then: "Happy to say we live in the same state. Im looking you up, and when I find you, im going to rape you and remove your head." There was more, but the final tweet summed it up: "You are going to die and I am the one who is going to kill you. I promise you this."*

Hess went to the police, but the officer she spoke to didn't even know what Twitter was, and didn't take any of it seriously. To Hess, this was a failure of the legal system. She knew there was an industry-wide problem, one that was gendered, and she knew it because her colleagues faced the same problems. "None of this makes me exceptional," she wrote. "It just makes me a woman with an internet connection."

Zoë Quinn

Zoë Quinn sits at the center of an enormous conflagration across the entire games industry known as Gamergate. Gamergate as a phenomenon is marked by incessant, low-grade harassment on social media, punctuated by loud, malicious doxes (the nonconsensual publication of private information such as physical addresses, Social Security numbers, and so on). Many people—game developers, game journalists, and ordinary people in the community—have been dragged into the mess, simply by expressing opinions on one side or the other. Both sides have claimed to have been harassed, and both sides have also accused the other of making up that harassment.

It's not immediately obvious to the casual observer what Gamergate is about, or why Quinn matters so much to it. The tagline of Gamergate, which is, by now, a source of dark humor on the internet, is "ethics in journalism." Quinn, an independent game developer, is accused of many things, including exchanging sexual favors for positive reviews of her game Depression Quest. (The man she supposedly slept with never even reviewed her game.) The "ambient hum of menace" in Quinn's life is less centered around how her small indie game ruined the entire genre of video games and more around supposed sexual activity (a theme that comes up again and again when women are harassed).

There's much to be said about Gamergate as the template of a new kind of culture war, and a signifier of a new era in games, one that has been called by some "the death of the gamers." The phenomenon has forced some women out of game development and games journalism. It's brought into new prominence the term "social justice warrior," or SJW for short. The SJW moniker seems to come from the belief that people who criticize video games for a lack of diversity are the enemy—a kind of cultural juggernaut with a supposed chokehold on the media, which must be forcefully opposed. Gamergate as a force is aligned against everyone they perceive to be SJWs. What any of this has to do with Zoë Quinn is not particularly obvious.

Gamergate is complicated. It's also fairly simple: it's a harassment campaign instigated by Zoë Quinn's ex-boyfriend, Eron Gjoni. Quinn was already being harassed before Gjoni, but her ex amplified it many times over: "Before Gjoni's post, she had received 16 megabytes of abuse. When she stopped saving threats last December [2014]—because she couldn't keep up with the bombardment—she had 16 gigabytes: 1,000 times more."

Quinn and Gjoni dated for five months. After the relationship ended, he created "The Zoe Post," a blog post alleging that she had been unfaithful to him during their relationship. A *Boston Magazine* profile of Gjoni states, "By the time he released the post into the wild, he figured the odds of Quinn's being harassed were 80 percent."

He was right. Quinn received a barrage of threatening messages, like

> *If I ever see you are doing a pannel [sic] at an event I am going to, I will literally kill you. You are lower than shit and deserve to be hurt, maimed, killed, and finally, graced with my piss on your rotting corpse a thousand times over.*

Quinn was doxed—her personal information, including her address and Social Security number, was published. She moved. The harassment continued—and with some patient investigation, Quinn was able to document Gjoni egging on her harassers from behind the scenes. What Gjoni was doing was both complicated and simple, old and new. He had managed to crowdsource domestic abuse.

Kathy Sierra

After the previous examples, Kathy Sierra's story will begin to sound redundant. But what happened to Sierra, an author and tech blogger, happened in 2007, long before this current wave of interest in gendered harassment. At first she only received messages, messages that read just like the kinds received by Quinn, Criado Perez, Sarkeesian, and Hess.

- "Fuck off you boring slut . . . i hope someone slits your throat and chums down your gob."
- "Better watch your back on the streets whore. . . . Be a pity if you turned up in the gutter where you belong, with a machete shoved in that self-righteous little cunt of yours.[1]"
- "The only thing Kathy Sierra has to offer me is that noose in her neck size."

And then came the doxing, a pseudonymous post that published her Social Security number and address. It was accompanied by a fabricated history of Sierra's life that would be echoed in the claims in "The Zoe Post"—claims about her getting plastic surgery, about her cheating on her former husband, about her turning to prostitution. The threats ramped up. Sierra moved across the country.

There are huge swaths of her story that Sierra won't talk about publicly, and it's understandable. It is, in fact, deeply unusual for Sierra to have gone as public as she did with the harassment she faced. Targets of harassment tend to silence themselves for fear of further abuse. Their fears are not unfounded—a consistent pattern is that the harassment ramps up the more you talk about it. And so the worst of what they receive remains hidden from the public eye. But when someone is so publicly doxed, it's easy to guess what they're dealing with.

About That Media Narrative . . .

The first mention of Kathy Sierra in *The New York Times*, in 2007, doesn't talk much about how harassment upended her life. It focuses, rather, on the "online heckling," the "anonymous comments," the "vitriol," and "threats of suffocation, rape and hanging."

In the media narrative, harassment amounts to words—rape threats and bomb threats—from anonymous strangers, to women who have done "nothing to deserve it." Indeed, for people who don't engage in this kind of behavior, the fact that it happens at all is deeply perplexing.

For Sierra, she and other women are targeted because they have visibility at all:

> *The real problem—as my first harasser described— was that others were beginning to pay attention to me. He wrote as if mere exposure to my work was harming his world.*
>
> *. . .*
>
> *I now believe the most dangerous time for a woman with online visibility is the point at which others are seen to be listening, following, liking, favoriting, retweeting. In other words, the point at which her readers have (in the troll's mind) drunk the Kool-Aid. Apparently, that just can't be allowed.*

What happened to Sierra, to Quinn, to Hess, to Sarkeesian, to Criado Perez is frightening, absurd, and unconscionable. But it's also a very small microcosm of online harassment. These are the cleanest, simplest, most straightforward examples of harassment. The women have names and reputations and audiences who listen. Their attackers are anonymous. The assault is documentable. And the brutality of the onslaught doesn't seem to be "warranted." Because their stories fit this pattern, their narratives receive the most media play.

Sierra today is introspective about this phenomenon. She sees sustained abuse everywhere, hidden from sight and out of mind, simply because it "makes sense" to the public that the target is being harassed. When interviewed by Benjamen Walker on his *Theory of Everything* podcast, she said:

> *Probably my most surreal moment in my life was sitting in a restaurant with my daughters and there's CNN playing in this sports bar, and there is Michelle Malkin, and she's*

> *saying, "Well, where the hell was everyone when I was getting all my death threats? One little tech blogger gets death threats and oh my god." And I thought, Yeah, but, what do you expect? It's not surprising, because of the things she's saying. Of course, now I'm horrified that I thought that. But it's a natural reaction. And again, I think, the fact that people couldn't do that with me is exactly why it became a story that caught so many people's attention. Because people kept asking, What did she do? What did she do? What did she talk about to piss them off? And then they couldn't figure it out. Because there wasn't anything.*

The way the media framed the harassment campaign against Sierra also reflects a second bias in the public imagination. *The New York Times* has mentioned the word "doxing" three times. All three examples are about men, and in at least one, the word is misused. In the media narrative, harassment becomes unruly words, not Social Security numbers. It becomes rape threats, but not the publication of physical addresses. It becomes floods and floods of frightening tweets, not a SWAT team knocking on your door because someone on the internet called the police with a fake threat.

And lastly, the harassers are almost always depicted as anonymous strangers. Never mind that Kathy Sierra's most prominent harasser, Andrew Auernheimer, has always had his legal name connected to his online pseudonym, "weev." Never mind that the campaign against Zoë began with and was egged on by an ex-partner.

Despite the growing public concern expressed for the targets of gendered online harassment, harassment is still depicted as an ephemeral harm from ghostly entities. But it's not. It encompasses a vast spectrum that includes intimate partner violence, stalking, invasion of privacy, hacking, and even assault.

Is Harassment Gendered?

I focus here on gendered harassment for good reason. There is a considerable amount of documentation indicating that online harassment disproportionately impacts women. In the following section, I will review some of the literature, most of which was laid out in Danielle Citron's *Hate Crimes in Cyberspace*. These studies, to some extent, match up with anecdotal accounts and with my own personal experience. But there are good reasons to look out for future research on the topic, which I will elaborate below.

The US National Violence Against Women Survey estimates that 60 percent of "cyber stalking victims" are women. The National Center for Victims of Crimes estimates that women actually make up 70 percent. Working to Halt Online Abuse (WHOA) collected information from over 3,000 reports of "cyber harassment" between 2000 and 2011 and found that 72.5 percent of reporters were female. The Bureau of Justice Statistics reports that 74 percent of individuals stalked online or offline are female. From 1996 to 2000, the majority of the NYPD Computer Investigation and Technology Unit's aggravated cyber harassment victims were female.

In 2006, researchers placed silent bots on Internet Relay Chat. The bots with female names received 25 times more "malicious private messages" (defined as "sexually explicit or threatening language") than the bots with male names.

Casual, non-academic, less-controlled experiments bear out similar results. As described in the next section, Twitter users have experimented with changing their avatars to reflect different genders and/or races. Male users masquerading as female users were rudely awakened to find that they were now subjected to a constant buzz of malicious or just plain obnoxious remarks. One man created a female profile on the dating app Tinder that did nothing but silently match. In less than a few hours, men had called the bot "a bitch" and also told "her" to "go kill [herself]."

Women regularly report anecdotal evidence that changing their avatar to something other than a photo of their face (even if it's a cartoon of their face) dramatically decreases the hostile messages sent to them.

Intersections of Harassment

But these studies, statistics, and anecdotes need to be more thoroughly interrogated. There is one notable dissenting finding, though it can't be taken seriously—a 2014 study (the Demos study) that looked a sample of 65 British celebrities over two weeks and found that people tweeted more rude things to the men than the women. The Demos study should be disregarded purely on the basis of the sample—too small, taken from an unusual minority group (celebrities), and "calibrated" for unsound reasons (the celebrities were selected so the same number of tweets were aimed at men and women).

The findings of the Demos study can't in good faith be extrapolated more broadly, but on the other hand, it's a study that doesn't rely on self-reporting. The above examples, aside from the IRC bot study, all do. Self-reporting can lead to skewed results, not because women are whiny or more sensitive, but because the framing of the issue may or may not resonate equally across the board, or because particular victims actively avoid interacting with the entity that is collecting the data.

For example, Danielle Citron notes that a 2009 study of 992 undergraduate students found that "nonwhite females faced cyber harassment more than any other group, with 53% reporting having been harassed online. Next were white females, at 45%, and nonwhite males, at only 31%." But Citron then goes on to say, "There is no clear proof that race is determinative," since according to both the Bureau of Justice Statistics and cases closed by the New York City Police Department's CITU, the majority of cyber harassment victims are white women. But it takes no stretch of the imagination to believe that people of color are less likely to go to the police. Citron decries how victims are unwilling to report to the police due to the (not unreasonable) perception that their concerns will not be taken seriously. But she does not go into whether this perception might vary from person to person along lines of race or even gender, or whether this may even be tied to whether the person identifies an instance of online harassment under the phrase "online harassment" (or whatever wording the study or the law enforcement agency uses).

Researchers Alice Marwick and danah boyd, for instance, have found that teenagers will identify instances of online bullying as "drama"—the term "cyberbullying" fails to "resonate" with them. "Cyberbullying" is a term that adults use. Similarly, it's possible that the term "online harassment" is one that is more prevalently used by a particular group to describe their experiences, even if other groups are equally impacted by the exact same behavior.

Given the overall consistency of studies regarding online harassment and women, it would be quite the surprise to find that men are, after all, impacted equally. While more studies are needed, they will likely continue to support the proposition that women are impacted disproportionately. But future studies will have much to add about other aspects of harassment that are, at the moment, taken as gospel. For example, Citron reports that most harassers have no "personal connection to" their victim, based on a WHOA study. But this finding seems exactly like the kind of thing would be affected by the "cyberbullying"/"drama" effect. Future studies will also likely flesh out how other axes of oppression affect harassment.

I submit to you that harassment is amplified by many things, including gender identity, race, and sexual orientation. I make this claim from some empirical evidence and a great deal of personal observation. I've chosen to discuss gender and online misogyny at length because narratives about (white, heterosexual, cisgender, respectable) women have made the debate about online harassment more visible. And as part of that, the literature on harassment as a socially oppressive force is quite robust when it comes to gender, and less robust when it comes to other intersections of identity and oppression.

For the sake of the future of the internet, more studies should be conducted on this topic, and soon. But until more literature is published, I will simply describe the "Race Swap" experiment.

In 2014, the writer Jamie Nesbitt Golden—who is black, female, and visibly so in her Twitter avatar—changed her avatar on Twitter to a picture of a white man. She noticed an immediate change in how much harassment she received. Writer Mikki Kendall (also black, female, and recognizably so in her avatar) followed suit, and suggested that other people experiment with changing their avatars to reflect different races and genders—to essentially walk a mile in someone else's shoes. For Kendall, who typically receives dozens of vitriolic tweets a day, the change was marked. When discussing controversial topics, she was less likely to be the target of abusive comments, and instead, would receive "reasonable," "calm," and "curious" responses, even from the same people who had trolled her viciously in the past. Kendall said to the radio show *On The Media*, "I still got the occasional angry comment, but none were gendered, none were racialized. It was actually more, more likely to be something like 'oh you're a jerk, or you're an asshole.' That kind of thing."

When white male writers used avatars of white women, or people of color, they were dismayed by the change. At least one was forced to change back after two hours.

I want to make two points about intersectionality and harassment:

First, the internet is experienced completely differently by people who are visibly identifiable as a marginalized race or gender. It's a nastier, more exhausting internet, one that gets even nastier and even more exhausting as intersections stack up. It's something to keep in mind, particularly since media narratives of the "worst" kinds of harassment rarely feature people of color.

Second, intersections make a difference in how to craft policy. Anti-harassment has to be aimed at protecting the most vulnerable. What, for example, is the point of prioritizing educating police if the most vulnerable internet users (say, transgender people and/or sex workers and/or people of color) are the least likely to actually call the police? How does one mitigate sexual shaming through nude photos if the targeted individual is a sex worker?

These are considerations I hope readers can carry with them throughout the book. A one-size-fits-all approach based on the favored media narrative will certainly have unintended consequences for those whose stories don't make it into the headlines.

On Doxing

Doxing, mentioned briefly above, is a subset of harassment. And like “harassment,” the word “doxing” (sometimes spelled “doxxing”) is ill-defined. Bruce Schneier, an expert on security, defines it as following:

> *Those of you unfamiliar with hacker culture might need an explanation of “doxing.”*
>
> *The word refers to the practice of publishing personal information about people without their consent. Usually it's things like an address and phone number, but it can also be credit card details, medical information, private emails—pretty much anything an assailant can get his hands on.*
>
> *Doxing is not new; the term dates back to 2001 and the hacker group Anonymous. But it can be incredibly offensive. In 2014, several women were doxed by male gamers trying to intimidate them into keeping silent about sexism in computer games.*

The strict “hacker” definition of “dropping dox,” as it was initially phrased, involves the publication of documentation (or “docs”/“dox”). As Schneier points out, these can be addresses, phone numbers, financial information, medical records, emails, and so forth. The part where the definition of “doxing” gets murky is that the word’s prominent appearances in the media haven’t involved dropping dox at all. Rather, it’s come (sometimes!) to signify the unmasking of anonymous internet users without their consent. The word burst into the mainstream in 2012 (although it had been used in previous articles in 2011 in the paper), as documented by *The New York Times*' “Words of 2012,” which included the following:

> ***DOX:*** *To find and release all available information about a person or organization, usually for the purpose of exposing their identities or secrets. “Dox” is a longstanding shortening of “documents” or “to document,” especially in technology industries. In 2012, the high-profile Reddit user Violentacrez was doxed by Adrian Chen at Gawker to expose questionable behavior.*

In 2012, Adrian Chen published an article exposing the legal name of Reddit user Violentacrez, moderator of a number of subreddits like r/creepshots and r/jailbait (associated with photos of women taken without their consent, or photos that might actually be illegal). Although his article gave the name, occupation, and town of residence for Michael Brutsch, a.k.a. Violentacrez, nothing written in the article actually "dropped dox" on Brutsch. Neither his address, nor phone number, nor Social Security number was exposed. Yet the response to Chen's article, particularly on Reddit, was deeply vitriolic. Many subreddits still impose an embargo on *Gawker* articles as links, and one of the long-standing Reddit-wide rules imposes a ban on posting personal information:

> *NOT OK: Posting the full name, employer, or other real-life details of another redditor*

The idea that a "full name" can constitute a dox represents an understanding that in some contexts, the publication of a legal name serves as an incitement to drop a full dox. Through Google and other databases, a full name can lead to other details—a revealing social media account, pictures of one's children, a work address, even a home address. Information like addresses, phone numbers, and place of work has been publicly available for a long time. But as sociologist and writer Katherine Cross points out, "the unique force-multiplying effects of the internet are underestimated. There's a difference between info buried in small font in a dense book of which only a few thousand copies exist in a relatively small geographic location versus blasting this data out online where anyone with a net connection anywhere in the world can access it."

The context in which the publicly information gets posted matters. When the dox is posted "before a pre-existing hostile audience," the likelihood that malicious action follows from it is much higher. Cross calls it "crowd-sourcing harassment." In the words of Kathy Sierra:

> *That's the thing—it's not so much the big REVEAL, it's the context in which that reveal happens—where someone is hoping to whip others up into a froth and that at least some of them will be just angry and/or unbalanced enough to actually take action. The troll doesn't have to get his hands dirty knowing someone else will do it for him.*

Sierra was fully doxed, meaning that her Social Security number was posted. Yet the term "dox" was not associated with what happened to her until many years later. The person largely thought to have been responsible for the initial disclosure of information, Andrew "weev" Auernheimer, used the term "dropped docs" to a *New York Times* journalist in 2008 when interviewed about

Kathy Sierra. But the word only became associated with her much later on, as "dox" emerged into the mainstream, with its meaning diluted to include the mere unmasking of anonymous individuals.

Unmasking an identity can have terrible consequences for a person—particularly if their full name is tied to more sensitive publicly available information. However, when doxing includes home addresses and Social Security numbers, the consequences are, obviously, much weightier. A doxing can wreck your credit and leave you vulnerable to much more visceral threats, like letters and packages being sent to your home, or worse, assault, whether at the hands of a real-life stalker or the police.

SWATting

A dox can turn into assault by proxy when it progresses to SWATting, where a target's name and home address are used in a false emergency call that instigates a SWAT raid on the target. Katherine Cross writes, "They may accuse the dox victim of building a bomb, holding hostages, hosting a drug lab, or any number of other things likely to prompt a SWAT raid on the target's home. This has been a popular use of dox in the gaming community in particular."

Understandably, victims of SWATting are often reluctant to speak out. But this skews media commentary on the phenomenon. Most mainstream coverage of SWATting has focused on men, with no real reason for the SWATting being given. Meanwhile, in January 2015, three people were SWATted by Gamergate, the phenomenon that is at its core an extended harassment campaign against Zoë Quinn. On January 3rd, the former home of Grace Lynn, a games developer and critic of Gamergate, was SWATted. On January 9th, Israel Galvez, also a critic of Gamergate, was visited by police officers. Katherine Cross reported, "The lack of a more aggressive response was due to Galvez having warned his local police department that [an internet board] had doxed him and his family."

At the moment, the most prominent media coverage of SWATting has focused on young men SWATting other young men, particularly when the SWATting is livestreamed over the internet. In the most common narrative, viewers of livestreamed video games will call in SWAT teams on the livestreamer. The way the story is framed is young male video gamers playing extreme pranks on other young male video gamers. But SWATting also happens to women, as a reactive, ideological backlash against perceived feminism gone too far.

In her documentation of the phenomenon of doxing, Cross writes, "I am at pains to remind the reader that all of this traces back to opinions about video games, a seething hatred of feminists who play and write about them, and a harassment campaign that began as an extended act of domestic violence against developer Zoë Quinn by a vengefully abusive ex-boyfriend."

Doxing Women

Bruce Schneier notes that doxing has existed since 2001. Others recall seeing the term come up in IRC channels in the mid-2000s, particularly regarding retaliation against *New York Times* writer John Markoff, who had written a controversial exposé of the hacker Kevin Mitnick. Markoff is credited (or blamed) by some with having helped with Mitnick's later arrest and imprisonment. (The journalist's email account was compromised in 1996.) In 2011, dox were dropped on HBGary Federal, a company that claimed to be able to out members of Anonymous and LulzSec (infamous internet vigilante groups composed entirely of anonymous or pseudonymous members) by gathering information on social media. Dropping dox is how the internet retaliates against those who threaten it—but it's not just a substantive retaliation; it is a policing of the internet as a public space.

As of writing, of the half-dozen times "doxing" has been mentioned by *The New York Times* (and the one time "SWATting" has been mentioned), none of the instances reported involve women. The upsurge in the use of the term, as seen through the Google Trends graph, is tightly correlated instead with the rise of LulzSec and the Adrian Chen article about Michael Brutsch/Violentacrez. News stories and trend pieces about doxing and SWATting focus tightly on men, even now, in a moment where online harassment against women is receiving growing media attention.

Yet there is a clear and well-documented pattern of using doxing to punish women for being visible on the internet. Doxing is a tactic that hackers have used to "protect" other hackers (e.g., lashing out at the man who helped imprison Mitnick, or going after a company that seeks to unmask hackers). It says, "We're from the internet, and we don't like you." It says, "You don't belong here."

Doxing originated as vigilante retaliation by hackers against their perceived enemies. It makes less sense when it is performed against individuals like Kathy Sierra, Anita Sarkeesian, Zoë Quinn and so on, unless you understand the motivation as one of deep misogyny, one that says, "These women don't belong on the internet." Doxing is an intimidation tactic in what its practitioners clearly believe is a war for online spaces.

Like online verbal abuse, doxing is a tactic to dominate the voice of the internet. Everyone has his own understanding of what does or does not belong on the internet—in other words, what garbage needs to be taken out. In the case of misogynists, women are that garbage.

With this background in place, I have two points I want to make about doxing as a phenomenon, and why doxing should inform solutions to online harassment.

Concluding Thoughts on Doxing

First, one of the most obvious and yet most missed points is that *anonymity is not the problem*. It's quite apparent that anonymity can be a shield against this most extreme variant of online abuse. Twitter's head of Trust & Safety, Del Harvey, is in fact pseudonymous because at the age of 21, she started volunteering with the site Perverted Justice, which would catch predators on the internet by posing as children. Her work put people in jail, and her pseudonym is one of several protective measures taken because of it. In an interview with *Forbes*, Harvey stated, "I do a lot in my life to make myself difficult to locate."

When seeking to curb online abuse, reducing anonymity can actually exacerbate it. Any anti-harassment policy that looks to "unmask" people is not just a threat to, say, anonymous political speech in countries with repressive governments; it's actually counterproductive as an anti-harassment measure in the first place.

Second, something that is often missed with respect to this issue is that regulatory and legislative reforms that would mitigate or limit doxing were proposed almost 10 years ago to Congress by the Electronic Privacy Information Center (EPIC). EPIC prevailed in one way, but the harms they drew attention to have persisted. In this time when a very specific type of online harassment is in full swing, now would be the time to press forward with an enhanced privacy agenda that springs directly from EPIC's 2006 push against pretexting.

In February 2006, Marc Rotenberg, the executive director of EPIC, gave a prepared statement to the House Committee on Energy and Commerce, drawing attention to the issue of pretexting—obtaining personal information through fraudulent means, such as pretending to be someone else over the phone. Rotenberg noted that in one case in New Hampshire (*Remsburg v. Docusearch*), a woman's stalker had hired a data broker, which had then contracted a private investigator. The PI pretexted the woman, pretending to be her insurance company, and was able to procure her location. The information was handed over to the stalker; later, he shot the woman and then killed himself. "The availability of these services presents serious risks to victims of domestic violence and stalking," Rotenberg said in his prepared statement. "There is no reason why one should be able to obtain these records through pretexting."

As a result of these hearings, a federal law was passed banning pretexting to acquire someone else's personal phone records. But the risks posed by data brokers, to victims of domestic violence and stalking, remain. Physical addresses abound through data brokers. An address can be a simple Google search away. This is the primary method through which people get doxed on the internet.

The essential nature of the problem is tied to forms of violence against women like stalking, assault, and intimate partner violence. Doxing is not an Internet of Garbage problem the same way abusive words and pictures are. While the same things that can mitigate abuse in general can also mitigate doxing (e.g., healthy moderation of posts that contain dox, cultivation of norms against doxing, strictly enforced platform-wide rules on doxing), the consequences of doxing cannot be addressed by the same strategies. If the policy proposals in this book seem too little and too weak in the face of the kind of outrageous harassment documented in the media, it is because they aren't meant to fully solve the worst kind of online harassment. The absolute extremes of online harassment manifest from the same behavioral patterns that produce the overall grinding tedious malice directed at women, but they cannot be addressed through the same strategies.

A Taxonomy of Harassment

Harassment as a concept is a pretty big bucket, one that ranges from a single crude tweet to the police knocking down your front door. In order to craft reasonable policies, the bucket must be analyzed and broken down, but it is nonetheless all there in a single bucket. Targets of harassment, particularly members of marginalized groups, may view a single comment differently than an outsider might, because they recognize it as part of a larger pattern.

Harassment exists on two spectrums at once—one that is defined by *behavior* and one that is defined by *content*. The former is the better way to understand harassment, but the latter has received more attention in both popular discourse and academic treatments.

When looking at harassment as **content**, we ultimately fixate on "death threats" as one end of the spectrum, and "annoying messages" at the other end. Thus the debate ends up revolving around civil rights versus free speech—where is the line between mean comments and imminent danger? Between jokes and threats?

Behavior is a better, more useful lens through which to look at harassment. On one end of the behavioral spectrum of online harassment, we see a drive-by vitriolic message thrown out in the night; on the other end, we see the leaking of Social Security numbers, the publication of private photographs, the sending of SWAT teams to physical addresses, and physical assault. Saying that these behaviors are all on the same spectrum does not mean that they merit the same kind of censure and punishment. Similarly, catcalling does not merit criminal liability, but for recipients it nonetheless exists on a spectrum of sexist behavior—that is, the consistent male entitlement to a woman's attention that they receive—that runs right up to assault and other terrible physical consequences. I don't think that placing these behaviors side by side means it's impossible to differentiate between one act and the other for the purposes of post hoc punishment. Seeing these behaviors on the same spectrum becomes illuminating because it teaches us not how to punish, but how to design environments to make targets feel safe.

Harassing content and **harassing behavior** of course overlap. Internet postings are both content and behavior. But changing the lens can completely change your understanding. "Someone called me a bitch on their blog" is different from "Someone has posted on their blog about how much they hate me, every day for about three months." Both of these statements can be about the same situation, but one speaks through the lens of content, and the other through the lens of behavior.

Harassing content can be divided right along its spectrum from least extreme to most extreme:

- **Sub-Threatening Harassment**. Harassment that cannot be construed as a threat. Being called a bitch or a racial slur is sub-threatening harassment.
- **Colorably Threatening Harassment.** Harassment that is not overtly threatening, but is either ambiguously threatening such that an objective observer might have a hard time deciding, or is clearly intended to make the target fearful while maintaining plausible deniability. For example, "I hope someone slits your threat," is colorably threatening harassment. Likewise, so is sending a picture of a burning cross to an African American.
- **Overtly Threatening Harassment**. "I know where you live; I'm going to kill you tonight."

Harassing behavior, on the other hand, can be sorted in two ways, by **investment** and by **impact**. When sorted by investment, harassing behavior is defined according to the investment that the harasser makes in their efforts.

- **Sustained Hounding**. This is, more or less, stalking—a person acting solo who doggedly goes after one or more individuals, whether by just sending them horrible messages, obsessively posting about their personal lives, or even sending them physical mail or physically following them around.
- **Sustained Orchestration**. Orchestration is crowdsourced abuse—the recruitment of others into harassing someone in a sustained campaign. Orchestration may happen by simply posting dox, or by postings that incite an audience to go after someone for whatever reason.
- **Low-Level Mobbing**. This is the behavior of those who are recruited into a sustained campaign, but never themselves become orchestrators of the campaign. They amplify the harassment, but may not themselves obsess over the targets. They would not be harassing that individual without the orchestrator to egg them on.
- **Drive-By Harassment**. Just some random person being terrible as a one-off.

When classified by **impact**, harassing behavior is defined by the long-term effect on the target. In this sorting, the classifications overlap and interact.

- **Emotional Harm**. Emotional harm can run the gamut from being mildly put off, to being driven to serious distress and even depression.
- **Economic Harm**. Economic harm can be done by going after a target's job, making them unemployable by Google-bombing search results for their name, posting a Social Security number, and destroying their credit.
- **Personal Harm**. This includes assault, assault by proxy (SWATting), and compromising the target's privacy through, for example, doxing or hacking

All of these things absolutely overlap. But I would argue that by separating these aspects out, we can better understand how we need to design both legal systems and technical systems. Legal systems should address the most extreme kinds of harassment: the overtly threatening harassing content, the personal harm, some types of economic harm, and perhaps even some forms of sustained orchestration.

The technical architecture of online platforms, on the other hand, should be designed to *dampen* harassing behavior, while *shielding* targets from harassing content. It means creating technical friction in orchestrating a sustained campaign on a platform, or engaging in sustained hounding. For example, what if, after your fifth unanswered tweet within the hour to someone who didn't follow you, a pop-up asked if you really wanted to send that message?

It also means building user interfaces that impart a feeling of safety to targets. Code is never neutral, and interfaces can signal all kinds of things to users. For example, what if Caroline Criado Perez had been able to hit a "panic button," one that prompted her with a message that Twitter Trust & Safety was looking into the issue and, until then, messages from strangers would be hidden from her?

I've used examples that are specific to Twitter, because I want it to be apparent that these decisions have to be tailored to platforms. Although platforms can learn from each other and adopt similar methods, no rule or tactic can be universal. The important thing to take away is that simply deleting or filtering offending content is not the end goal. Deletion can be a form of discouragement toward harassers and safety for the harassed, but it's only one form.

On Modern-Day Social Media Content Moderation

I will acknowledge that there is a very good reason why the debate focuses on content over behavior. It's because most social media platforms in this era focus on content over behavior. Abuse reports are often examined in isolation. In an article for *Wired* in 2014, Adrian Chen wrote about the day-to-day job of a social media content moderator in the Philippines, blasting through each report so quickly that Chen, looking over the moderator's shoulder, barely had time to register what the photo was of. Present-day content moderation, often the realm of US-based contractors or even outsourced abroad (as in Chen's article), is set up to operate on an assembly line model, with discrete repetitive tasks given to each worker that operate along consistent, universal rules. (Whether consistency and efficiency are actually achieved is up for debate.) With the massive amount of reports that these teams must process, they don't have the time to deliberate as though they were a judge or jury. Easy, bright-line rules are the best. Tracking behavior over time or judging the larger context takes up time and energy.

Does this mean it's economically or technically impossible to incorporate more consideration for larger patterns of behaviors when moderating content? Absolutely not. But most importantly, even if the focus on creating bright-line rules specific to harassment-as-content never shifts, looking at harassing behavior as the real harm is helpful. Bright-line rules should be crafted to best address the behavior, even if the rule itself applies to content.

Beyond Deletion

The odd thing about the new era of major social media content moderation is that it focuses almost exclusively on deletion and banning (the removal of content and the removal of users, respectively).

Moderation isn't just a matter of deleting and banning, although those are certainly options. Here are a range of options for post hoc content management, some of which are informed by James Grimmelmann's article, "The Virtues of Moderation," which outlines a useful taxonomy for online communities and moderation:

- **Deletion.** Self-explanatory.
- **Filtering.** Filtering makes something hard to see, but doesn't remove it entirely. It could mean user-specific filtering (like Facebook's News Feed algorithm), or it could mean mass-filtering throughout the entire platform (on Reddit, negative-rated comments are grayed-out and automatically minimized, although they can be expanded if necessary). It

could also mean client-side filtering, such as an individual being able to block other users from interacting with them.

- **Editing.** Editing alters the content but doesn't necessarily remove it entirely. Editing can build on top of content productively. If a post is deemed to be worthy of being there, but has problematic parts that violate the rules, editing can remove the parts that are objectionable. For example, say that a journalist posts a series of public record request responses on social media as part of his reporting, and one of the documents accidentally reveals the physical address of a person. The best response here would be to edit, if possible, to remove the address, rather than ban the journalist or delete his postings. Editing can verge on being deletion in practice. See, for example, forums or comment threads where moderators are empowered to replace entire offending posts with humorous caps-lock commentary, such as "I AM A WHINY BABY."
- **Annotation.** Additional information is added. Grimmelmann offers the examples of the eBay buyer/seller feedback system, the Facebook "Like" button, and Amazon's reviews. Posts on many forums are annotated with ratings (indicating the goodness or badness of a post according to users). Annotation can be the handmaiden of filtering or blocking. Used in a certain way on certain platforms, it could also give users insight into whether another particular user is known to be an abusive personality.
- **Amplification and Diminution.** Amplification/Diminution is a hybrid of annotation and filtering. Technically, the Facebook "Like" button falls into this category. (Posts with more likes tend to appear in feeds more often!) Another example is how Reddit or Quora operate on an upvote/downvote system. The UI decision to gray-out negative-rated Reddit answers is technically a form of diminution. Yahoo Answers allows the asker to select the "best" answer, which then floats up to the top. The SomethingAwful forums maintain a Goldmine (where good threads are archived for posterity) and a Gaschamber (where bad threads are consigned to a deserved afterlife).

These are what you can do to **content**. Post hoc processes also include options levied against **users** for their offending behavior.

- **Banning.** A user account is deactivated.
- **IP Bans.** The Internet Protocol address the user is posting from is banned. IP bans are circumventable and can have unintended consequences. (For example, IP-banning a university student posting from a university connection could mean wiping out the posting privileges of many other people.) However, some platforms have found IP bans to be nonetheless warranted and effective.
- **Suspension.** This is assumed to be a temporary ban. It may be a suspension for a set time period, or it may be a suspension pending certain actions the user has been asked to take.
- **Accountability Processes.** This is a new form of post hoc moderation processes directed toward users, one that holds a great deal of promise. An accountability process pulls a user aside, not only to put a check on their behavior, but also to rehabilitate them. Laura Hudson reported on how the online game League of Legends created a successful accountability process:

> League of Legends *launched a disciplinary system called the Tribunal, in which a jury of fellow players votes on reported instances of bad behavior. Empowered to issue everything from email warnings to longer-term bans, users have cast tens of millions of votes about the behavior of fellow players. When Riot [the company] asked its staff to audit the verdicts, it found that the staff unanimously agreed with users in nearly 80% of cases. And this system is not just punishing players; it’s rehabilitating them, elevating more than 280,000 censured gamers to good standing. Riot regularly receives apologies from players who have been through the Tribunal system, saying they hadn’t understood how offensive their behavior was until it was pointed out to them. Others have actually asked to be placed in a Restricted Chat Mode, which limits the number of messages they can send in games—forcing a choice to communicate with their teammates instead of harassing others.*

Harassers’ motivations are ill-understood. It may be that harassers are simply misguided people. It may be that they are incurable sociopaths. (It may be both.) But accountability processes work not only because they give people a chance to have a genuine change of heart; they also shine sunlight into people’s faces, letting them know their community does not condone their behavior. A user doesn’t have to have a real change of heart to decide to simply go along with the norms that are being enforced.

What Happens Before: Setting Norms

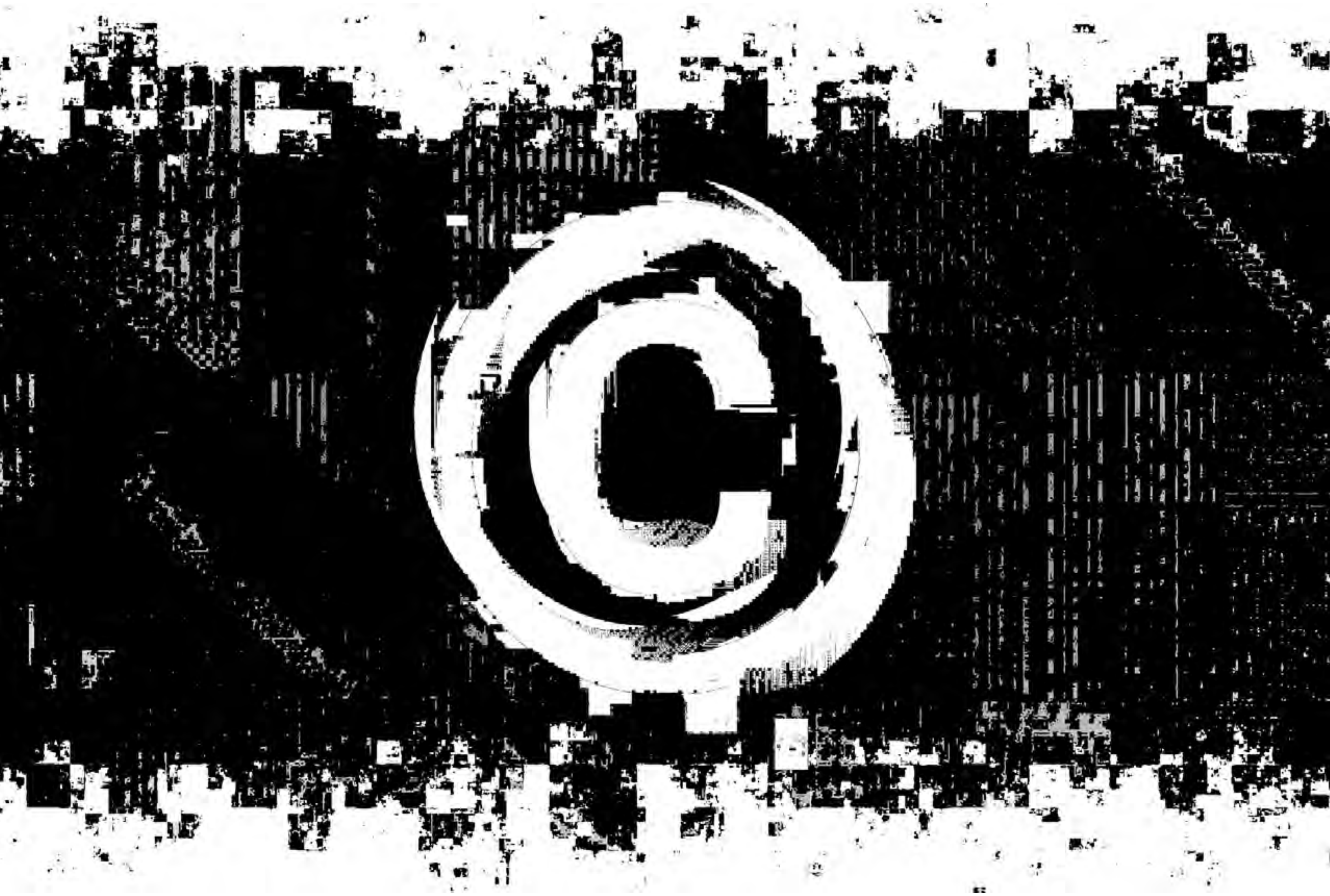
All of the above methods of ex post moderation also operate on the ex ante level—when users see how their community is being moderated, they conform to avoid being moderated. (In many places, when a new user makes a misstep or posts something undesirable or boring, they will often be told hostilely to “lurk more”—in essence, asking them to absorb the norms before bothering to contribute.)

But norms can also be set outside of ex post action on offending content. For example:

- **Articulation.** This is just setting clear rules. Anil Dash writes in a blog post titled “If Your Website’s Full of Assholes, It’s Your Fault,” that community policies or codes of conduct should be “short, written in plain language, easily accessible, and phrased in flexible terms so people aren’t trying to nitpick the details of the rules when they break them.”
- **Positive Reinforcement.** Moderation can demonstrate to users which kinds of posts are bad, but it can also demonstrate what kinds of posts are good and worth striving toward. Grimmelmann uses the example of moderators selecting “new and noteworthy” posts. Reddit offers a “best of Reddit” spotlight, and also gives users the options to “gild” each other—buy a premium account for a user they think has written a good post. A post that has been “gilded” is annotated with gold star.
- **The Aura of Accountability.** Good ex post moderation means an aura of accountability is preserved within the community—that is, there are consequences for bad behavior. But there are ex ante considerations as well. Anil Dash suggests forcing users to have “accountable identities,” which could mean real names, or it could just mean having a consistent pseudonym over time.

The Aura of Accountability doesn’t only go one way. If users feel that the platform is accountable to them, they are more invested in the platform being good and less likely to trash it. Nearly all platforms use blogs to update the community on policy changes. The blogs often try to use language that indicates that the platform is “listening” or that it “cares.” This is pretty *pro forma*—can you imagine a platform that straightforwardly admits that its priority is pleasing shareholders? What’s more interesting is how some platforms and communities have adopted user-oriented governance structures. Wikipedia makes for an interesting example, but one particularly compelling (though perhaps not replicable) instance is how the game EVE Online has a Council of Stellar Management (CSM), composed of representatives elected from their user base. The CSM is actually regularly flown out to the company headquarters to meet with staff and discuss the game and its features.

Creating user investment in their community, what Grimmelmann calls a "sense of belonging and their commitment to the good of community," is a matter of both moderation and architecture. It's not just a touchy-feely thing—it's also a technical problem, in that the code of the platform must be able to stop or adequately deter bad actors from constantly registering new accounts or overwhelming the platform with unwanted messages. It's cyclical: when people are invested in the community, the community will police and enforce norms, but when unrepentant bad actors are never banished or are able to reproduce their presence at an alarming rate (sockpuppeting), community trust and investment will evaporate.



Chapter Three:

Lessons from Copyright Law

The Intersection of Copyright and Harassment

On December 15th, 2014, an *en banc* panel of 11 judges of the Ninth Circuit Court of Appeals sat for oral arguments in *Garcia v. Google*. Cris Armenta, the attorney for the plaintiff, began her argument:

> *Cindy Lee Garcia is an ordinary woman, surviving under extraordinary circumstances. After YouTube hosted a film trailer that contained her performance, she received the following threats in writing:*
>
> *Record at 218: "Are you mad, you dirty bitch? I kill you. Stop the film. Otherwise, I kill you."*
>
> *Record at 212: "Hey you bitch, why you make the movie Innocence of Muslim? Delete this movie otherwise I am the mafia don."*
>
> *Record at 220: "I kill whoever have hand in insulting my prophet."*
>
> *Last one, Record at 217. Not the last threat, just the last one I'll read. "O enemy of Allah, if you are insulting Mohammed prophet's life, suffer forever, never let you live it freely, sore and painful. Wait for my reply."*

At this point, Armenta was interrupted by Judge Johnnie Rawlinson. "Counsel, how do those threats go to the preliminary injunction standard?"

Indeed, her opening was an odd way to begin, and the observers—mostly lawyers deeply familiar with copyright who had followed the case with great interest—were confused by it. Wasn't *Garcia* a case about copyright law and preliminary injunctions?

For Cindy Lee Garcia, of course it wasn't. It was a case about her right to control her exposure on the internet. But in her quest to end the barrage of hate aimed at her, she ended up in a messy collision with copyright doctrine, the Digital Millennium Copyright Act (DMCA), the Communications Decency Act (CDA), and the First Amendment.

The Ninth Circuit had released an opinion earlier that year, written by then Chief Judge Alex Kozinski. *Garcia* may have made few headlines, but it caused a wild frenzy in the world of copyright academia. In short, Kozinski's opinion appeared to break copyright law as had been understood for decades, if not a century.

The case was a hard one. The plaintiff was sympathetic, the facts were bad, and the law was—Kozinski aside—straightforward. Cindy Garcia had been tricked into acting in the film *The Innocence of Muslims*. Her dialogue was later dubbed over to be insulting to the prophet Mohammed. Later the film's controversial nature would play an odd role in geopolitics. At one point, the State Department would blame the film for inciting the attack on the Benghazi embassy.

Meanwhile, Garcia was receiving a barrage of threats due to her role in the film. She feared for her safety. The film's producers, who had tricked her, had vanished into thin air. She couldn't get justice from them, so she had to settle for something different. Garcia wanted the film taken offline—and she wanted the courts to force YouTube to do it.

Garcia had first tried to use the DMCA. YouTube wouldn't honor her request. Their reasoning was simple. The DMCA is a process for removing copyrighted content, not offensive or threatening material. While Garcia's motivations were eminently understandable, her legal case was null. The copyright owner of the trailer for *The Innocence of Muslims* was Nakoula Basseley Nakoula, not Garcia. Garcia pressed the theory that her "performance" within the video clip (which amounted to five seconds of screen time) was independently copyrightable, and that she had a right to issue a DMCA takedown. YouTube disagreed, and their position was far from unfounded—numerous copyright scholars also agreed. (In the December 2014 en banc hearing, Judge M. Margaret McKeown would comment, "Could any person who appeared in the battle scenes of *The Lord of the Rings* claim rights in the work?")

Garcia went to court. She lost in the district court, and she appealed up the Ninth Circuit. To nearly everyone's surprise, then Chief Judge Kozinski agreed with her that her five-second performance had an independent copyright, a move that went against traditional doctrinal understandings of authorship and fixation.

A strange thing then unfolded there. It wasn't merely a decision that Garcia had a copyright inside of a work someone else had made. If it had been, Garcia could go home and reissue the DMCA request. But instead, the court ordered YouTube to take down the video—thus creating an end-run around the DMCA, even though the DMCA notice-and-takedown procedure had been specifically designed to grant services like YouTube "safe harbor" from lawsuits so long as they complied with notice-and-takedown. (Cathy Gellis, in an amicus brief written for Floor64, additionally argued that an end-run around CDA 230 had also been created.) Kozinski had broken copyright law *and* the DMCA.

Google/YouTube immediately appealed the decision, requesting an *en banc* hearing—essentially, asking the court of appeals to rehear the case, with 11 judges sitting instead of only three. Their petition was accompanied by 10 amicus briefs by newspapers, documentarians, advocacy groups, industry groups for technology companies and broadcasters, corporations like Netflix and Adobe, and law professors by the dozen.

Nobody liked the *Garcia* ruling. What did it mean for news reporting casting interview subjects in an unflattering light? And what did it mean for reality television shows? For documentaries? What did it mean for services like Netflix that hosted those shows and documentaries? The first Ninth Circuit opinion had created a gaping hole in copyright and had pierced through the well-settled rules that governed how copyright liability worked on the internet.

In May 2015, the first ruling was reversed by the *en banc* panel. “We are sympathetic to her plight,” the court wrote. “Nonetheless, the claim against Google is grounded in copyright law, not privacy, emotional distress, or tort law.”

Garcia is a case that may even go up to the Supreme Court, though until then, interest in *Garcia* will likely be confined to copyright academics and industry lawyers. Yet lurking beneath the thorny legal and doctrinal issues is the great paradigm shift of the present digital age, the rise of the conscious and affirmative belief that women should have, must have, some kind of legal recourse to threats online. It’s how Cris Armenta wanted to frame her argument, and it is no doubt an important motivating factor to the 2014 Kozinski decision. Cindy Lee Garcia is a woman stuck between a rock and a hard place. Nonetheless, the 2014 *Garcia* decision is wrongly decided. *Garcia* is not just a weird copyright case; it’s a case that speaks volumes about popular attitudes toward online harassment and about the dead end that will come about from the focus on content removal.

How the DMCA Taught Us All the Wrong Lessons

Cindy Garcia went straight to the DMCA because it was the "only" option she had. But it was also the "only" option in her mind because 16 years of the DMCA had trained her to think in terms of ownership, control, and deletion.

When you assume that your only recourse for safety is deletion, you don't have very many options. It's often very difficult to target the poster directly. They might be anonymous. They might have disappeared. They might live in a different country. So usually, when seeking to delete something off the web, wronged individuals go after the platform that hosts the content. The problem is that those platforms are mostly immunized through Section 230 of the Communications Decency Act (described in detail below). The biggest gaping hole in CDA 230, however, is copyright. That's where most of the action regarding legally required deletion on the internet happens, and all of that is regulated by the DMCA.

The Digital Millennium Copyright Act

The Digital Millennium Copyright Act, among other things, provides "safe harbor" to third party intermediaries so long as they comply with notice-and-takedown procedures. So if a user uploads a Metallica music video without permission, Warner Bros. cannot directly proceed to suing YouTube. Instead, Warner Bros. would send a DMCA notice. If the notice is proper, YouTube would be forced to take down the video, or otherwise they would no longer be in their "safe harbor."

The safe harbor provision of the DMCA is largely touted with encouraging the rise of services like YouTube, Reddit, WordPress, and Tumblr—services that are considered pillars of the current internet. These sites host user-generated content. While there are certainly rules on these sites, the mass of user-generated content can't be totally controlled. Without DMCA safe harbor, these sites couldn't cope with copyright liability for material that slipped through the cracks. Although today YouTube uses a sophisticated Content ID system that does manage to automatically identify copyrighted content with surprisingly accuracy, Content ID was developed later in YouTube's history. This extraordinary R&D project couldn't have existed without the early umbrella of protection provided by DMCA safe harbor. Theoretically, DMCA safe harbor protects the little guys, ensuring that the internet will continue to evolve, flourish, and provide ever-new options for consumers.

The DMCA is also one of the handful of ways you force an online intermediary to remove content.

The Communications Decency Act, Section 230

Under present law, DMCA works in lockstep with Section 230 of the Communications Decency Act, which generally immunizes services from legal liability for the posts of their users. Thanks to CDA 230, if someone tweets something defamatory about the Church of Scientology, Twitter can't be sued for defamation.

There are very few exceptions to CDA 230. One notable exception is federal law banning child pornography. But the big one is copyrighted material. Copyright infringement is not shielded by CDA 230—instead, any violations are regulated by the provisions of the DMCA instead.

CDA 230 was created in response to *Stratton Oakmont, Inc. v. Prodigy*, a case where the web service Prodigy was sued for bulletin board posts that "defamed" Wall Street firm Stratton Oakmont. (Today, Stratton Oakmont is best known as the subject of the Martin Scorsese film *The Wolf of Wall Street*, a film adaptation of a memoir.)

At the time, Prodigy received 60,000 postings a day on its bulletin boards. The key was that Prodigy did enforce rules, even if it couldn't control every single posting. By taking any sort of action to curate its boards, it had opened itself up to liability. Strangely, the *Stratton Oakmont* decision discouraged moderation and encouraged services to leave their boards open as a free-for-all. Legislators sought to reverse *Stratton Oakmont* by creating CDA 230.

Changing CDA 230?

CDA 230 was a shield intended to encourage site moderation and voluntary processes for removal of offensive material. Ironically, it is presently also the greatest stumbling block for many of the anti-harassment proposals floating around today. CDA 230 can seemingly provide a shield for revenge porn sites—sites that purportedly post user-submitted nude pictures of women without their consent. Danielle Citron in *Hate Crimes in Cyberspace* proposes creating a new exception to CDA 230 that would allow for liability for sites dedicated to revenge porn, a smaller subset of a category of sites for which Citron adopts Brian Leiter's label: "cyber-cesspool."

CDA 230 has no doubt been essential in creating the internet of 2015. Any changes to the status quo must be carefully considered—how much of the internet would the new exception take down, and which parts of the internet would it be? What kind of exception would there be for news sites and newsworthy material? Crafting the perfect exception to CDA 230 is not theoretically impossible, but then there is an additional practical aspect that muddies the waters.

Any legislation laying out a new exception, no matter how carefully crafted from the start, will likely suffer from mission creep, making the exception bigger and bigger. See, for example, efforts to add provisions to outlaw "stealing cable" in a 2013 Canadian cyberbullying bill. Anti-harassment initiatives become Trojan Horses of unrelated regulation. It is rhetorically difficult to oppose those who claim to represent exploited women and children, so various interest groups will tack on their agendas in hopes of flying under the cover of a good cause.

At the time of writing, CDA 230 remains unaltered. But new considerations are in play. Many of the major revenge porn sites have been successfully targeted either by state attorneys general or by agencies like the Federal Trade Commission. One operator, at least, was not blindly receiving submissions as a CDA 230–protected intermediary, but was actually *hacking* into women's email accounts to procure the photos. Other operators were engaging in extortion, charging people to "take down" the photos for a fee. Revenge porn websites have demonstrated a long and consistent pattern of unlawful conduct adjacent to hosting the revenge porn itself. These sites, which Danielle Citron calls the "worst actors," never quite evade the law even with CDA 230 standing as-is. It turns out that these worst actors are, well, the worst.

A new exception to CDA 230 aimed at protecting the targets of harassing behavior stands in an uncanny intersection. A narrow exception does not officially make criminals out of people who were acting badly; it rather targets people who have consistently demonstrated themselves to be engaged in a host of other crimes that are prosecutable. But a broad exception, targeted just a step above the "worst actors," could be disastrous for the internet.

Turning Hate Crimes into Copyright Crimes

When her *Hate Crimes in Cyberspace* went to print, Citron outlined a proposal for a limited and narrow exception to CDA 230, meant to target these "worst actors." But she also took great pains to explain how it was not targeted at other, more mainstream sites, citing Reddit as an example of a site that would not be affected.

Shortly after *Hate Crimes in Cyberspace* was published in September 2014, Reddit became ground zero for the distribution of nude photos of celebrities that had been hacked from their Apple iCloud accounts. "Leaked" nudes or sex tapes are nothing new in Hollywood, but in an era of increasing awareness of misogyny on the web, this mass nonconsensual distribution of photos struck a new chord. Jennifer Lawrence called what happened to her a "sex crime," and many pundits agreed.

Reddit was slow to remove the subreddit that was the gathering place for the photos. But eventually it did, with the reasoning being that the images being shared there were copyrighted. A tone-deaf blog post by then CEO Yishan Wong announced that they were "unlikely to make changes to our existing site content policies in response to this specific event," explaining:

> *The reason is because we consider ourselves not just a company running a website where one can post links and discuss them, but the government of a new type of community. The role and responsibility of a government differs from that of a private corporation, in that it exercises restraint in the usage of its powers.*

The title of the post was, incredibly, "Every Man is Responsible for His Own Soul." Yishan Wong resigned in November 2014 (supposedly over an unrelated conflict). In February 2015, under new CEO Ellen Pao, Reddit implemented new policies on nonconsensually distributed nude photos. By May 2015, Reddit implemented site-wide anti-harassment policies.

As of writing, Reddit is now in a very different place than it was in 2014—but its actions in September of that year are a fascinating case study in the worst way for a platform to handle harassment. Reddit is not a "worst actor" in the hierarchy of platforms, and its relative prominence on the internet likely did end up influencing its eventual policy changes, despite initial resistance. What's striking about the September 2014 incident is that in removing the offending subreddit, Reddit did not appeal to morals, the invasion of privacy, Reddit's pre-existing rule against doxing (the nonconsensual publication of personal information), or the likely crime that had occurred in acquiring the photos in the first place. Instead, Reddit cited DMCA notices, effectively placing copyright as a priority over any of those other rationales.

The affair doesn't cast Reddit in a particularly good light, but the bizarre entanglement between the DMCA and gendered harassment on the internet isn't new. Regardless of their motivations, both Reddit and Cindy Lee Garcia fell into the same trap: They turned a hate crime into a copyright crime.

When people are harassed on the internet, the instinctive feeling of those targeted is that the internet is out of control and must be reined in. The most prominent and broad regulation of the internet is through copyright, as publicized in the thousands of lawsuits that the Recording Industry Association of America (RIAA) launched against individual downloaders, the subpoenas the RIAA issued to the ISPs to unmask downloaders, and the RIAA and MPAA's massive lawsuits against the Napsters, Groksters, and even YouTubes of the world. In our mass cultural consciousness, we have absorbed the overall success of the RIAA and the MPAA in these suits, and have come to believe that this is how one successfully manages to reach through a computer screen and punch someone else in the face.

Online harassment, amplified on axes of gender identity, race, and sexual orientation, is an issue of social oppression that is being sucked into a policy arena that was prepped and primed by the RIAA in the early 2000s. The censorship of the early internet has revolved around copyright enforcement, rather than the safety of vulnerable internet users. And so we now tackle the issue of gendered harassment in a time where people understand policing the internet chiefly as a matter of content identification and removal—and most dramatically, by unmasking users and hounding them through the courts.

Yet an anti-harassment strategy that models itself after internet copyright enforcement is bound to fail. Although the penalties for copyright infringement are massive (for example, statutory damages for downloading a single song can be up to $150,000), and although the music and movie industries are well-moneyed and well-lawyered, downloading and file-sharing continue.

Content removal is a game of whack-a-mole, as Cindy Lee Garcia learned. Shortly after the first Ninth Circuit decision in her favor, she filed an emergency contempt motion claiming that copies of *The Innocence of Muslims* were still available on the platform, demanding that Google/YouTube not only take down specific URLs but also take proactive steps to block anything that came up in a search for “innocence of Muslims.”

From Garcia’s point of view, if her safety was at stake, then only a total blackout could protect her. But copyright law was not created to protect people from fatwas. Her case, already a strange contortion of copyright law, became even messier at this moment, as her lawyer asked for $127.8 million in contempt penalties—the copyright statutory damages maximum of $150,000 multiplied by the 852 channels that were allegedly “still up.” At that moment, Cindy Garcia, who had so far been a sympathetic plaintiff laboring under extraordinarily difficult circumstances, suddenly became indistinguishable from a copyright troll—plaintiffs who abuse copyright law in order to make substantial financial profits.

Google’s reply brief clapped back: “Garcia’s fundamental complaint appears to be that *Innocence of Muslims* is still on the internet. But Google and YouTube do not operate the internet.”

The Illusive Goal of Total Control

Garcia may have been right that removing or disabling most or even some instances of the video could have mitigated her circumstances. But it’s hard to say, especially once the cat was out of the bag. Indeed, during the December 2014 oral arguments, Judge Richard Clifton chimed in with, “Is there anyone in the world who doesn’t know your client is associated with this video?” Garcia’s attorney stumbled for a bit, and Judge Clifton interrupted again, musing, “Maybe in a cave someplace, and those are the people we worry about, but. . . ”

In many circumstances, when online content continues to draw attention to a target of harassment, the harassment is amplified, and once the content falls away out of sight, the interest disappears as well. But at the same time, Garcia wasn’t seeking to merely mitigate the harassment; she wanted to wipe the film off the internet simply because she had appeared in it.

Garcia was chasing a dream of being able to completely control her image on the internet. It’s an echo of the same dream that the record industry has been chasing since the 1990s. It’s not that you *can’t* impact or influence or dampen content in the digital realm. But there’s no way to control every single instance, forever.

Any anti-harassment strategy that focuses on deletion and removal is doomed to spin in circles, damned to the Sisyphean task of stamping out infinitely replicable information. And here, of course, is the crux of the issue:

harassing content overlaps with harassing behavior, but the content itself is only bits and bytes. It’s the consequences that echo around the content that are truly damaging—threats, stalking, assault, impact on someone’s employment, and the unasked-for emotional cost of using the internet. The bits and bytes can be rearranged to minimize these consequences. And that’s a matter of architectural reconfiguration, filtering, community management, norm-enforcement, and yes, some deletion. But deletion should be thought of one tool in the toolbox, not the end goal. Because deletion isn’t victory, liberation or freedom from fear. It’s just deletion.

Chapter Four:
A Different Kind of Free Speech

Stuck on Free Speech

As mentioned earlier, when we focus on content over behavior, there is a terrible tendency to get stuck on terms like “threats,” “true threats,” “imminent harm,” and “hate speech.” These are all terms borrowed from the American tradition of First Amendment jurisprudence.

The specter of American constitutional law looms large over the landscape of extralegal social media rules. It is imbued throughout the wording of the terms of service of various sites, in both the official rules of platforms and the justifications they give for them. See, for example, how Reddit’s Yishan Wong spoke of “imminent harm” in 2014 (no doubt invoking the Brandenburg test), or how Twitter’s Tony Wang called the company the “free speech wing of the free speech party” in 2012, or how Facebook changed its policies in 2013 to prohibit what they describe as “hate speech.”

The adoption of American constitutional jargon likely has a lot to do with the American origin of most English-speaking social media platforms, and may even be a cultural carry-over from the birth of ARPANET and Usenet (the early predecessors of the web we know today) in the States.

Nonetheless, the language of First Amendment jurisprudence online is thrown around without much regard for their actual use in case law. While this is not the place to give a full summary of First Amendment doctrine, the following are useful points to keep in mind:

- **The First Amendment does not apply to online platforms.** Facebook, WordPress, Twitter, and Reddit are private entities. The First Amendment only applies to government action. So the First Amendment would bar an *American law* (state or federal) that banned rape jokes. It would not bar Facebook from banning rape jokes.
- **The “shouting ‘Fire!’ in a crowded theater” analogy is not completely true.** The First Amendment does not protect certain forms of “dangerous” speech, but the danger must be “directed to inciting, and likely to incite, imminent lawless action.” This is known as the Brandenburg Test. For example, shouting “Fire!” in a crowded theater is protected by the First Amendment if the shouter really believes there is a fire, even when there isn’t.
- **A “true threat” doesn’t mean that the threatener actually intends to carry out the threat.** True threats are presently ill-defined. The most recent case involving true threats, *Elonis v. United States*, has not offered any elaboration on the subject. But we do know that true threats are separate from the Brandenburg Test. A true threat doesn’t have to be factual in order to be true; it’s true because it makes the recipient fear serious violence and is intended to make them afraid. *Virginia v. Black* summarized the law:

> *“True threats” encompass those statements where the speaker means to communicate a serious expression of an*

> *intent to commit an act of unlawful violence to a particular individual or group of individuals.* ***The speaker need not actually intend to carry out the threat****.* *Rather, a prohibition on true threats “***protect[s] individuals from the fear of violence***” and “***from the disruption that fear engenders***,” in addition to protecting people “from the possibility that the threatened violence will occur.” [emphasis added]*

- **Hate speech *is* protected under the First Amendment.** Of course, not *all* hate speech. Hate speech that runs afoul of the Brandenburg Test, or turns into a true threat, is not protected. But hate speech is, as Scalia would put it, a viewpoint, and discriminating against a viewpoint does not comport with the First Amendment.

The Marketplace of Ideas

If the First Amendment doesn't apply to social media platforms, why should we care about free speech in the first place?

It's a good question to ask. But before I attempt to answer that, let's talk about where the First Amendment comes from.

First Amendment doctrine is actually relatively new, born over a century after the actual text of the Amendment was written. It begins with a series of cases clustered around the first two World Wars and the onset of the Cold War, where American socialists, communists, anarchists, and anti-war activists were prosecuted for activities ranging from printing anti-conscription pamphlets to speaking at socialist conventions. These prosecutions, and some of the decisions behind them, were motivated by the fear that radical speech would result in national destruction, whether through demoralization in a time of war, or through violent overthrow of the United States by communists. While the United States probably didn't have anything to fear from presidential candidate Eugene Debs (convicted and imprisoned for speaking at a state convention of the Ohio Socialist Party), it was a time when rhetoric felt far from "mere words." With the rise of fascism in Europe and the violent, ideologically motivated overthrow of governments overseas, radical political activities were policed ever-heavily in the States.

Present-day First Amendment doctrine is born out of *Abrams v. United States*, a 1919 case where anarchists were convicted for pro–Russian Bolshevik pamphlets that included exhortations like, "Awake! Awake, you Workers of the World! Revolutionists" and "Workers, our reply to this barbaric intervention has to be a general strike!" and "Woe unto those who will be in the way of progress. Let solidarity live!" The Supreme Court upheld their conviction, in a decision that is now considered to be defunct. It is rather the dissent by Oliver Wendell Holmes that lives on:

> *But when men have realized that time has upset many fighting faiths, they may come to believe even more than they believe the very foundations of their own conduct that the ultimate good desired is better reached by free trade in ideas—that the best test of truth is the power of the thought to get itself accepted in the competition of the market, and that truth is the only ground upon which their wishes safely can be carried out. That, at*

> *any rate, is the theory of our Constitution. It is an experiment, as all life is an experiment. Every year, if not every day, we have to wager our salvation upon some prophecy based upon imperfect knowledge.*

Holmes here makes an implicit reference to political philosopher John Stuart Mill's concept of the "marketplace of ideas." For Mill, speech is best left unhindered because the truth will emerge out of the free competition of ideas. Even opinions that seem obviously wrong should be left alone. Not only is there always the chance that the seemingly wrong opinion might turn out to be true, but the existence of a wrong opinion helps people to better see the truth. Seeing a wrong idea next to a right idea creates "the clearer perception and livelier impression of truth, produced by its collision with error."

The idea of this marketplace appears in other political theorists' writing, including Hannah Arendt's *The Human Condition*, where she theorizes the *polis*—the Greek democratic city-state—as springing from the *agora* (literally "marketplace"), where a man can distinguish himself as a "doer of great deeds and the speaker of great words" by participating in the discussions occurring in the public spaces of the city.

Arendt's *polis* is "the organization of the people as it arises out of acting and speaking together." The capacity for speech, the opportunity to speak, and the greatness of the words were fundamental to politics and to Greek democracy, since society was held together by persuasion rather than force. And persuasion could only be achieved by speech.

The idea that the freedom of speech, the marketplace of ideas, and the agora are the precondition of democracy persists today, inherent in American law. It can also be found, like a lingering vestigial structure, in the codes of conduct and terms of service of various internet platforms first founded in America.

Social Media is a Monopoly

So here we get to why free speech is important even when we're dealing with private entities like Facebook or Twitter.

A social network is not like another service. It may derive value from your data, but it doesn't do anything useful with your data for you. Facebook is valuable because of one's Facebook friends. In order to really "move" from one service to another, a user has to move their entire network of friends with them. Of course, there were social networks before Facebook, and it is somewhat comforting to believe in a kind of a generational cycle for the web. Just as Facebook replaced Myspace and Myspace replaced LiveJournal, so too will some yet-unbranded entity supersede Facebook.

This belief, of course, glosses over the many ways in which Facebook has fostered and entrenched its own ubiquity. The Like button may have indeed achieved the "seamless integration" between Facebook and the internet that the company sought to create. Then there's the company's push in Asia, Africa, and Latin America to provide free Facebook access ("Facebook Zero") to users of WAP-enabled feature phones (cellphones that are not smartphones). In countries like the Philippines or Myanmar, where people primarily access the internet through feature phones, Facebook is the internet. This is known as "zero-rating"—Wikipedia is also zero-rated in many countries, but Facebook has been the most aggressive about it. Zero-rating violates the principle of net neutrality, which explains why zero-rating happens abroad (even though for some years, net neutrality was not quite the "law of the land").

Ubiquity often trumps the flaws that Facebook has been repeatedly criticized for. Even when changes in privacy endanger activists, they stay on Facebook. In 2012, Katherine Maher, now chief communications officer at Wikimedia, was quoted as saying:

> *Traveling around Tunisia, I kept being introduced to bloggers . . . [but] it turns out very few of them were actually blogging. . . . Instead, the "bloggers" were people who were active on Facebook. . . . People are on Facebook not because they prefer it for any ideological reason, but because it offers the ability to reach the right people, with minimal effort, and maximum replicability.[b]*

Social media researcher danah boyd wrote in 2010, "Facebook may not be at the scale of the internet (or the internet at the scale of electricity), but that doesn't mean that it's not angling to be a utility or quickly become one." In the present day, it is no secret that Facebook is aiming for "zero-rating" throughout the developing world, looking to capture the "next billion" on their network.

But even looking just at the United States, social media companies entrench themselves, because social networks are a natural monopoly. Individual users stay even if they don't like the platform or the UI, because *it's where the conversation is happening*. A social media platform is like the only mall or the only church in a small town. You might not like the mall, and you might not be a Christian, but you have to go meet with your neighbors somewhere.

Social media has a huge distorting effect on public discourse. In 2014, Facebook drove 3.5 times more traffic to BuzzFeed Partner Network news sites than Google did. Facebook clicks are vital to ad-based journalism. A story that simply doesn't look very interesting to Facebook users won't be read as widely. In 2014, in the wake of the Ferguson protests, sociologist Zeynep Tufekci noticed that discussion of Ferguson was absent from Facebook, although Twitter was ablaze with pictures, videos, commentary, links to bail funds, and outpourings of support, grief, anger, hate, and vitriol. Had Facebook's algorithmic filtering eaten up posts about Ferguson?

It might have been as simple as a difference between what people use Twitter for, as opposed to Facebook, but this example should be enough to give pause. A person who primarily uses Twitter would be far more likely to be aware of Ferguson than a person who primarily uses Facebook. Social media has an enormous distorting effect on what we perceive as civic participants.

If "blind" algorithmic filtering can have that kind of a disproportionate effect on us, we should take all the more care with content-based rules about speech on social networks. In the wake of a policy change where Facebook began to take down misogynistic content, the ACLU responded with a blog post that acknowledged that although the First Amendment did not technically apply, its size and ubiquity should be of concern. "In short, when it comes to the vast realm it oversees, Facebook is a government."

One critic scathingly responded, "Shorter ACLU: 'Facebook is so big it's LIKE a government so the 1st Amendment DOES apply!'"

But the ACLU's point should be taken seriously. The internet is presently siloed off into several major private platforms. Each silo is maintained by, in the words of the ACLU, "a near-absolute ruler that can set any rule, squelch any speech, expel any ‘citizen’ for any reason, with only the due process protections it sees fit to spend money on.”

It doesn’t mean that the First Amendment must be blindly appropriated or applied to Facebook as though it were indeed the United States government. After all, content-based rules are inevitable, because every platform inevitably has rules about what is too worthless to keep around. But these rules are the expression of how the platforms actively decide what kind of societies and communities they are cultivating. If equal civic participation and democratic societies are what we seek, then these content-based rules have to be designed with that in mind.

Agoras and Responsibilities

Online platforms have often co-opted First Amendment language in ways that don't make much sense. It's not just that the First Amendment technically doesn't apply to them. Rather, the platforms that claim to uphold "free speech" are actually proactively engaged in moderation models that are not just mildly inconsistent with, but are deeply adverse to, the freedom of speech.

In 2011, after Adrian Chen "doxed" Reddit user Violentacrez on *Gawker*, links to *Gawker* articles were banned on several major subreddits on Reddit. The ban remains partially in place today. It was not only a punitive response to the speech of an individual on a separate platform, but a long-term embargo on a *press* organization for doing something that certain Reddit moderators disagreed with. Regardless of one's views on whether Chen should have outed Violentacrez/Michael Brutsch, this result does not exactly generate a free marketplace of ideas. Reflect back on Reddit's response to the hacked celebrity nude photographs posted in 2014. "[W]e consider ourselves . . . the government of a new type of community. The role and responsibility of a government differs from that of a private corporation, in that it exercises restraint in the usage of its powers."

In the same infamous "Every Man is Responsible for His Own Soul" blog post, Yishan Wong also added, "You choose what to post. You choose what to read. You choose what kind of subreddit to create and what kind of rules you will enforce. We will try not to interfere."

It was a strange series of laissez-faire pronouncements. Posters, readers, and moderators exist on completely different levels of control. The only thing that makes them the same is that they are all *users* of Reddit—both consumers and unpaid laborers all at once.

The key to parsing the discrepancy between Reddit's actual model and its claims to free speech is that Reddit runs on the free labor of unpaid subreddit moderators, with each moderator or group of moderators cultivating their own little fiefdom where they enforce their own rules. Reddit's supposed commitment to free speech is actually a punting of responsibility. It is expensive for Reddit to make and maintain the rules that would keep subreddits orderly, on-topic, and not full of garbage (or at least, not hopelessly full of garbage). Only by giving their moderators near absolute power (under the guise of "free speech") can Reddit exist in the first place.

Other platforms are not trapped in such a vicious catch-22, but the same cost-cutting attitude can be spotted in early-stage platforms. For example, Twitter implemented the “Report Abuse” button in 2013 shortly after the Caroline Criado Perez story made waves in the media. The implementation of more extensive content moderation had been in the works, but this had to be rushed out in response to the backlash. For many social platforms, moderation is an afterthought, tacked on top of the technology.

The New Agoras

Communities have a vested interest in discouraging behavior in the general category of harassment, exactly because ignoring the little things means implicitly condoning the rest of it, and creating an atmosphere of fear for potential targets. Encouragingly, many platforms and services are beginning to acknowledge this. Just in the early half of 2015, both Twitter and Reddit, notorious for their “free speech” stances, announced new policies on harassment. Vijaya Gadde, general counsel for Twitter, wrote in an op-ed for *The Washington Post* that “Freedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up.” Some months after that, Reddit announced a policy change prohibiting harassment. Their explanation: “Instead of promoting free expression of ideas, we are seeing our open policies stifling free expression; people avoid participating for fear of their personal and family safety.” Effective anti-harassment can make a freer marketplace of ideas, rather than inhibiting it.

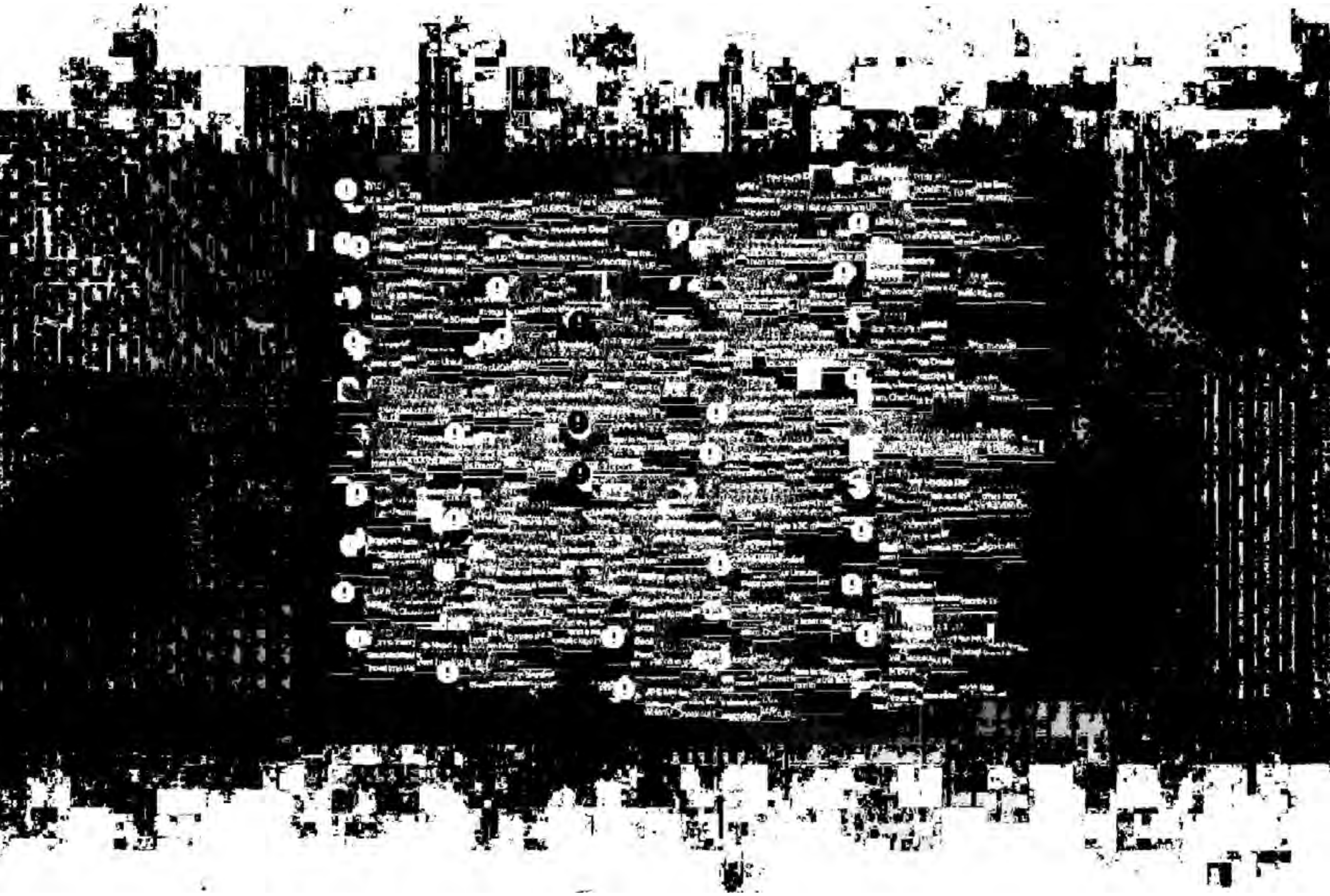
Promoting user safety doesn’t mean mass censorship is the answer. Of course, different kinds of platforms have different kinds of obligations to their users. In 2008, Heather Champ, the director of community at Flickr, was quoted as “defending the ‘Flickrness of Flickr,’” while saying, “We don’t need to be the photo-sharing site for all people. We don’t need to take all comers. It’s important to me that Flickr was built on certain principles.” Flickr is not Facebook, Facebook is not Twitter, Twitter is not Reddit, Reddit is not 4chan, 4chan is not a forum to discuss chronic illness and that forum is not a private mailing list.

For small and intimate communities, the question of balancing speech and user safety is relatively null. But large-scale platforms are different. Although they are technically private property and not subject to First Amendment speech protections even when their users and servers are based in the US, they are beginning to resemble public squares of discussion and debate, the main staging grounds of the kind of speech that connects people to other people and forms the foundation of democracy. Platforms like Facebook, Twitter, or Wikipedia might not have a legal obligation to protect free speech, but failure to do so would have serious consequences for culture itself.

Communities that do purport to be for everyone have an obligation to cultivate a community of inclusive values simply because they should put their money where their mouths are. Free speech doesn't just mean the ability for anyone to say anything. When free speech forms the foundation of democracy, free speech is more than a libertarian value. It indicates a more complex, difficult obligation: the obligation to create a space for everyone. It means a space where participants aren't silenced by fear, or shouted down.

In 2012, Twitter CEO Dick Costolo called his company a reinvention of the agora. There's a bit of deep irony there. The agora isn't the agora just because anyone can say anything. The agora is the place where equals meet to discourse with each other: The agora is where Greek democracy begins. And Greek democracy by definition excluded women, slaves, and foreigners. When we seek to build truly equal platforms and marketplaces of ideas fit for the 21st century, we are trying to create things that have never existed and cannot be constructed by mindlessly applying principles of the past.

Free speech is an innovation we are constantly striving toward, not something that can be achieved with a total hands-off punting of responsibility. We see in the next section why that is—even though John Stuart Mill might have thought that even bad or wrong ideas had value in the marketplace of ideas, John Stuart Mill never dealt with legions of junk-transmitting botnets. Unbeknownst to most users of the internet, we constantly live on the brink of being swallowed up by garbage content. We breathe free because of the occasionally overzealous efforts of anti-spam engineering.



Chapter Five:

Spam: The Mother of All Garbage

What Is Spam?

Spam is the mother of all garbage.

As mentioned earlier, spam gained its name from the practice of "spamming" early chat rooms with "spam, spam, spam, spammity spam," thus rendering the chat room unusable. Spam shuts down communications and renders communities unable to function. It overwhelms the human attention span, it threatens technical structures, it even exacts an economic cost to the internet at large. For Finn Brunton, whose history of spam spans from 1970s to the present day, spam is "very nearly the perfect obverse of 'community.'"

On Usenet in the 1980s, spam had not yet come to specifically signify machine-generated commercial text. Brunton writes, "Spamming was taking up more than your fair share of that expensive and precious data transmission every night as departments paid to pull in megabytes of data over their modems and consuming the scarce disk space with duplicate text so sysadmins [system administrators] would need to delete the whole batch of messages sooner."

The days of Usenet are gone, but spam still burns up dollars, server space, man-hours, and other resources. Spam is the internet's ever-present sickness, permanently occupying the attention of a vast and hyper-intelligent immune system.

Yet most internet users hardly ever have to think about spam. Anti-spam, particularly email anti-spam, is at a point where abuse/spam engineers like Mike Hearn can declare "the spam war has been won. . . for now, at least."

But at what cost? Some sources estimate that the email security market is about $7 billion. In 2012, Justin Rao and David Reiley estimated that "American firms and consumers experience costs of almost $20 billion annually due to spam." (The worst part is that they also estimated that spammers worldwide only gross about $200 million annually.) No clear figures could be found on the cost of anti-spam efforts on social media platforms, or how many personnel work on the problem full-time across the industry.

It's clear, however, that a massive effort is holding back a deluge. A Gartner Report in July 2013 estimated that about 69 percent of all email is spam. And 69 percent is a vast improvement! Email spam is actually on the decline as spammers shift their efforts toward social media—in 2010, Symantec estimated that 89 percent of all email was spam.

Without anti-spam, the internet simply wouldn't be functional. And at the same time, spam is speech. Low-value speech, perhaps, and speech that *ought* to be censored, but speech regardless. Debates over spam and censorship have gone back and forth from the early days of Usenet all the way to the 2000s. But today, as we face a very similar debate with harassment, comparisons to spam are rarely made, even though it's almost nearly the same problem—that is, the problem of garbage removal.

The Anti-Spam War Machine

"When you look at what it's taken to win the spam war with cleartext [unencrypted mail], it's been a pretty incredible effort stretched over many years," wrote Mike Hearn, a former engineer on Gmail's abuse team, in a mailing list email on cryptography and spam that has since become widely read. "'War' is a good analogy: There were two opposing sides and many interesting battles, skirmishes, tactics, and weapons."

The war involved intense volunteer efforts, experiments and innovations, research and development, and the rise of a specialized private sector. It took international, coordinated hubs like the organization Spamhaus. It now eats up billions of dollars a year.

Hearn's own brief history of spam begins with the "regex" or "regular expression," defined as a sequence of characters that forms a pattern to match. For example, why not simply block every email that contains the word "Viagra"? (Hearn recounts that not only did spammers adapt to avoid blacklists, innocent bystanders were affected, like an Italian woman named "Oli**via Gra**dina" who had all of her emails "blackholed.")

Of course, it wasn't as simple as just filtering for a blacklist of words picked out by people. (Or at least, it wasn't quite that simple for very long.) Bayesian filers looked through massive amounts of emails over time and gradually determined that certain words or combinations of words were associated with spam, thus allowing for a filter to bounce emails that matched that profile. We are of course familiar with what spam morphed into in response to Bayesian filters: when spam breaks through our current highly effective filters, it comes in the form of litspam—strange broken poetry, nonsensical sentences that have been arranged to evade filters. The arms race accelerated, with anti-spammers working to create filters that could detect litspam.

The current filter deployed by Gmail works by calculating reputations for the sending domain of the email. A reputation is scored out of 100, and is a moving average that is calculated by a mix of manual feedback (users pushing the

“Report Spam / Not Spam” buttons) and automatic feedback (presumably a form of Bayesian filter).

The success of this filter is a non-insignificant accomplishment. Reputations have to be calculated quickly. Because a mail with an unknown reputation is always let through, spammers will simply try to “outrun” the system. The reputation calculating system eventually was able to calculate scores “on the fly,” thanks to a “global, real-time peer-to-peer learning system . . . distributed throughout the world and [able to] tolerate the loss of multiple datacenters.”

But once this incredible piece of engineering was up and running, Gmail had to battle a new problem that it had set up itself. Now that spammers were burning addresses, and sometimes entire sending domains, on the reputation system, they had to get new ones.

Because Gmail accounts are free, spammers were free to sign up for Gmail accounts. (Hearn’s own work involved creating programs that would resist automated signup scripts, making it difficult for spammers to automatically acquire new Gmail addresses.)

Spammers also sought to hijack email addresses and domains by stealing passwords. In 2013, Gmail declared victory against the hijackers in a blog post describing how enhanced security measures, such as sending a verification text message to a phone number when suspicious activity was detected on the account, had put an end to, in Hearn’s words, “industrial-scale hacking of accounts using compromised passwords.”

Gmail may have presently won its war against spam after going through extraordinary measures, but anti-spam is an ongoing project elsewhere. Companies can be quite close-lipped about their anti-spam efforts. In the middle of a war, after all, you don’t want to leak intelligence to the enemy.

But what’s clear is that the fight against garbage is one that enacts a significant economic cost. It employs brilliant minds and it generates extraordinary new technologies. It’s a raging war that the average user gives little thought to, because, well, the anti-spammers are winning. Emails go through (mostly). Inboxes aren’t hopelessly clogged. The internet, overall, is operational.

Spam and Free Speech

Those who are spammed are often completely and utterly convinced that the spam is *garbage*, that it is trash that should have never flickered before their eyes. The debate about spam and free speech has mostly fizzled out in this decade, but for the entire history of spam and anti-spam, the censorship of speech by anti-spam measures has been an ongoing concern.

The very first "spam" message, according to Finn Brunton, was in fact political speech: an unsolicited mass mailing sent out on the Compatible Time-Sharing System (CTSS) network that had been developed at MIT, a network similar to but different from the internet's direct predecessor, ARPANET. The message was a long, impassioned antiwar speech, beginning with: "THERE IS NO WAY TO PEACE. PEACE IS THE WAY."

The "spammer" was Peter Bos, a sysadmin who had used his special privileges on CTSS to mass-message everyone. When his superior told him it was "inappropriate and possibly unwelcome," he replied that it was important. Sending the message out through CTSS meant it would reach an audience of scientists and engineers likely to be working on defense-related projects—for Bos, his unwanted mass mailing was a matter of conscience.

This "first" spam message has a strange correlation to the pamphlets and speeches at issue in the earliest First Amendment cases in US law. Later "spam" doesn't so much resemble political speech, but spammers have long cited their freedom to speak. "It's what America was built on. Small business owners have a right to direct marketing," said Laura Betterly, a commercial spammer, in a 2004 interview.

In 2005, the Electronic Frontier Foundation (EFF) published "Noncommercial Email Lists: Collateral Damage in the Fight Against Spam," a white paper describing their free speech concerns with anti-spam technology and outlining principles to limit overzealous spam blocking. "Our goal is to ensure that internet users receive all of the email they want to receive, without being inundated with unwanted messages. At the same time, we want to preserve the ability to send bulk email to lists of people who have chosen to receive it—something spam-blocking technologies and policies threaten to burden, if not eliminate."

The EFF cited the repeated difficulties that MoveOn.org, a progressive activist organization that often sent out action alerts through email, faced due to anti-spam.

> *Often, these alerts will ask subscribers to send letters to their representatives about time-sensitive issues, or provide details about upcoming political events. Although people on the MoveOn.org email lists have specifically requested to receive these alerts, many large ISPs regularly block them because they assume bulk email is spam. As a result, concerned citizens do not receive timely news about political issues that they want.*

For the EFF, this was "free speech being chilled in the service of blocking spam." It's hard to argue with them, particularly since people on the MoveOn.org email lists had opted in to receive the emails. This problem has a lot to do with the nature of server-side anti-spam—centralized filtering that has a tendency not to take into account the individual preferences of the recipients. In many cases, this is filtering the recipients couldn't opt out of. Today, the server-side/client-side distinction when it comes to email spam is much more nuanced. As discussed in the last section, the Gmail anti-spam reputation system does take into account the recipients' preferences (i.e., whether they marked as spam or not spam), thus allowing the definition of "spam" to be significantly determined by subjective preferences.

What client-side spam filtering does is—depending on your point of view—either give freedom to the user to decide or punt responsibility for garbage disposal to the user. Being responsible for defining spam on your own or opting in to certain blacklists is simultaneously better and worse for you. It is both freeing and burdensome. In contrast, server-side filtering has enormous benefits in terms of efficiency and network effects, even if it takes agency away from the end-user.

The speech versus spam concerns have never gone away. They remain with us even if they don't circulate much in mainstream public discourse, and they should certainly be taken seriously. However, even a hard civil libertarian organization like the Electronic Frontier Foundation acknowledges that spam filtering is an essential part of the internet.

Harassment as Spam

The thing is that harassment is not that different from spam—and not just in that the loosest definition of both is "unwanted messages." Mikki Kendall, when discussing how RaceSwap mitigated the daily harassment she faced, said, "One of the things that's really nice is not waking up to see 62 comments calling me everything but a child of god."

For large-scale sustained campaigns and short-term pile-ons, harassment is harmful not just because of threats or the general emotional impact; it is also harmful because it makes the internet completely useless. Inboxes fill up. Social media feeds flood with hate.

Behind the scenes, the general advice circulating when these things happen is to turn off your phone and hand over social media accounts to trusted parties who will look through the messages for anything imminently threatening, and to *simply stop being on the internet*, because the internet is momentarily broken for you.

Large-scale sustained campaigns also resemble tiny, crude, handmade botnets. At the center is an orchestrator, who perhaps engages in harassment from multiple "sockpuppet" accounts—the corollary to the burnable domains and email addresses that are so highly sought after by spammers. But the really bizarre phenomenon is all the low-level mobbers, who have little-to-no real investment in going after the target, and would not manifest any obsessions with that particular target without the orchestrator to set them off. Here they resemble the zombie nodes of spam botnets, right down to the tactics that have been observed to be deployed—rote lines and messages are sometimes made available through Pastebin, a text-sharing website, and low-level mobbers are encouraged to find people to message and then copy/paste that message.

In late 2014, I reported on a bizarre occurrence where Twitter had apparently begun to blacklist slurs in @-replies to a UK politician of Jewish origin who was being piled on by white supremacists. The blacklist was a crude, ham-fisted one that resembles those from the earliest days of email anti-spam, the kind of regex-filtering that had "blackholed" emails for poor Olivia Gradina. In response, the white supremacists had merely gotten riled up and began to call on each other to harass her even more, recommending that the others start putting asterisks or dashes into their slurs, or even use images of text so that the text filter couldn't spot what they were doing—a miniature version of the earliest years of the anti-spam arms race.

Patterning harassment *directly* after anti-spam is not the answer, but there are obvious parallels. The real question to ask here is, *Why haven't these parallels been explored yet?* Anti-spam is huge, and the state of the spam/anti-spam war is deeply advanced. It's an entrenched industry with specialized engineers and massive research and development. Tech industries are certainly not spending *billions of dollars* on anti-harassment. Why is anti-harassment so far behind?

A snide response might be that if harassment disproportionately impacts women, then spam disproportionately impacts men—what with the ads for Viagra, penis size enhancers, and mail-order brides. And a quick glance at any history of the early internet would reveal that the architecture was driven heavily by male engineers.

But that is the snide response, and is, of course, a gross overgeneralization. Yet it's puzzling. Harassment isn't a new problem in the slightest. Finn Brunton's own history of spam, which of course focuses on *spam*, nonetheless reveals the evolution of both spam and harassment in tandem.

For example, in 1988, a "grifter" began to post scam messages on Usenet begging for money. As anger over the messages, his service provider—placed under enormous pressure—ended up posting his real name and phone number. With that, the denizens of Usenet found his physical address and began to hound him.

In 1992, one member of the online community, the WELL, posted "An Expedition into Nana's Cunt," a long and hateful tirade against his female ex-partner who was also active on the WELL. Brunton writes, "As the argument about whether to freeze or delete the topic dragged on, other users began bombarding the topic with enormous slabs of text, duplicated protests, nonsense phrases—spam—to dilute Mandel's hateful weirdness in a torrent of lexical noise rendering it unusable as a venue for his breakdown."

And in the late 1990s, a vigilante hacker posted the personal emails and files of a notorious spammer, including "photographs . . . in various states of undress in her office and at her home." This likely is not the first instance of revenge porn, but it's a rather early one.

These are only some early incidents that intersected with the history of spam, but they're revealing. They involve the same tactics and sometimes the same misogyny that we see today. Harassment has always been with us, and yet we do not have many more tools than the people on Usenet or the WELL did.

Architectural Solutions to Harassment

I certainly don't mean to say that the solution to harassment is the simple, brute application of anti-spam technology to it. First of all, the free speech concerns regarding anti-spam are still with us. Calibrating anti-spam technology *per se* to make sure political speech isn't stifled is an ongoing process, and would be an even more difficult task when applying that technology to harassment. Secondly, it wouldn't really work. Spam as a phenomenon looks similar to harassment, but it's on a larger scale, motivated by different reasons, run by different kinds of people, and propped up by a different technical architecture.

Nonetheless, anti-spam is an important lesson. Garbage can be mitigated by and disposed of through architectural solutions—in other words, by code. Manual deletion, manual banning and legal action are not the only tools in the toolkit.

Architectural solutions are on the rise. On Twitter, users have taken to deploying self-help mechanisms like The Blockbot, the GG Autoblocker, and Blocktogether. The Blockbot is a centralized blocklist with multiple tiers maintained by a few people with privileged access to adding accounts to the list. GG Autoblocker (GGAB) is a blunt algorithmic predictive tool that determines the likelihood that an account is part of the "Gamergate" phenomenon, and accordingly blocks them. (GGAB is counterbalanced by a manual appeals process that wrongfully blocked persons can go through.) Blocktogether is a tool for sharing blocklists with other people. It allows people to subscribe to their friends' blocklists—a rough-and-tumble form of the same work that Spamhaus does. (GGAB uses Blocktogether to share its auto-generated list.) In June 2015, Twitter implemented sharable blocklists, no doubt taking their cue from Blocktogether. At time of writing, Twitter's in-house blocklist sharing functionality is much more rudimentary than the features available through Blocktogether.

All of these tools are opt-in. They are a rough form of client-side filtering, meant to address what is perceived to be the laxness of Twitter on the server-side. (Notably, the Blocktogether tool was created by a former Twitter anti-spam engineer who is now working at the Electronic Frontier Foundation.) This kind of filtering doesn't delete the messages of harassers; it merely removes an unwilling audience, thus balancing out speech concerns and the needs of the harassed.

None of these garbage-removal tools can actually stop stalkers, doxers, hackers. They do not change the root behaviors of harassers. They are, in fact, incredibly blunt, and with the exception of Blocktogether—which is more of an add-on feature and less of a filter on its own—should never be adopted server-side. But they provide their users with peace of mind and a better, enhanced social media experience. They are a UI tweak that takes away the low, angry buzz that comes with being a target on the internet.

These kinds of technical solutions are non-trivial improvements to the everyday lives of many individuals. Dealing with garbage is time-consuming and emotionally taxing. That's why social media companies pay people to do it full time, and why those employees often feel the need to stop after a few years. In his article for *Wired*, Adrian Chen quoted a former YouTube content moderator as saying, "Everybody hits the wall, generally between three and five months. You just think, 'Holy shit, what am I spending my day doing? This is awful.'"

In 2014, the news blog *Jezebel*, a satellite of the Gawker network, posted what can only be described as a revolt against their management. "We Have a Rape Gif Problem and Gawker Media Won't Do Anything About It," read the headline.

The open discussion platform on the Gawker sites (known as Kinja) allowed for anonymous, untracked comments, the rationale being that whistleblowers ought to be protected. Meanwhile, the *Jezebel* site, a feminist-oriented women's interest blog, was being bombarded by anonymous comments containing graphic animated gifs of women being raped. According to *Jezebel*, "because IP addresses aren't recorded on burner accounts, literally nothing is stopping this individual or individuals from immediately signing up for another, and posting another wave of violent images."

Jezebel writers were expected to moderate the comments and delete them so the readers didn't have to see them. But what of the *Jezebel* staff? "Gawker's leadership is prioritizing theoretical anonymous tipsters over a very real and immediate threat to the mental health of *Jezebel*'s staff and readers," they wrote.

Here, Gawker Media had made the mistake of seeing this as an intractable tradeoff between harassment and free speech. Whistleblowers must be protected, ergo, *Jezebel* staffers must see the rape gifs. A furious debate erupted in the media world. Meanwhile, in more technical circles, the bemused question was raised—why hadn't Gawker Media just disabled gif embedding on anonymous burner accounts?

This is one example where architecture can operate in tandem with moderation. Code is never neutral; it can inhibit and enhance certain kinds of speech over others. Where code fails, moderation has to step in. Sometimes code *ought* to fail to inhibit speech, because that speech exists in a gray area. (Think:

emails in the Gmail system that have not yet received a reputation rating.) But it’s delusional to think that architecture never has any effect on speech whatsoever. Technical and manual garbage-removal are two sides of the same coin, and must work together if garbage is to be taken (out) seriously.

The thing about beginning the technological arms race against harassment is that even if it’s different from spam in tricky ways, the arms race will simply never reach the scale of the spam war. It’s not just that there’s no economic incentive to harass; it’s also that harassment is harassment because it’s meant to have an emotional impact on the recipient. Harassment can’t evolve into litspam because then it wouldn’t be harassment anymore.

On the Size of the Internet

Why is online harassment so scary anyways?

This may be an odd question to throw out at this juncture, but while we're talking about user interfaces, we should talk about how content can be digitally packaged to amplify harassing behavior.

Think about it: what is it about a tweet that contains "@yourusername" that becomes a personal offense? An aggressive comment made to my face could easily escalate. And an angry letter mailed to my physical address can become an implicit threat. But what kind of harm is happening when strangers are shouting at me from miles away? If someone graffitis a graphic sexualized comment on a wall in another city? If someone unleashes a long, disturbing rant about me in the privacy of their own home?

As with other policy debates about the internet—whether it's about downloading movies, or disabling a server with a distributed denial of service attack—arguments for and against regulation of harassing speech rely on analogies to real-world behavior. "You wouldn't steal a car, would you?" asks the MPAA. Copyright activists might reply that although stealing a car leaves one less car for the owner, downloading a movie means a second copy in the world. "Breaking windows is illegal, why not breaking websites?" one might argue in favor of criminalizing DDoS attacks. But Computer Fraud & Abuse Act reformists will point out that a DDoS closely parallels calling a telephone number over and over again.

Debates about the internet are battles of analogies. The debate over online harassment isn't any different, but the nature of the analogy is intrinsically different.

Other internet policy issues have to do with the *bigness* of the internet, its Wild West nature. Copyright enforcement is a game of whack-a-mole because the internet is so "big" and information "wants to be free." The number of infringing links and torrents approaches the infinite, as does the number of viruses and malware loose on the web. Black markets on the dark net continue to proliferate despite law enforcement crackdowns, and cultivate their reputations as places where "anything" can be bought. When it comes to these issues, the internet looks like an eternal frontier, a never-ending expanse with room for an infinite amount of data, information, and gathering places. (The internet's premier impact litigation group, the Electronic Frontier Foundation, implicitly references that aspect in its own name.)

But harassment isn't a "Wild West" problem. Harassment doesn't happen because the internet is "too big"—it happens because it's too small. Part of this has to do with thoughtless user interface design. When it comes to Twitter, an @-reply feels offensive because it "invades" your online space; it's a message delivered straight to your face because the Twitter user interface is designed to make @-replies visible.

More importantly, examination of sustained harassment campaigns shows that they are often coordinated out of another online space. In some subcultures these are known as "forum raids," and are often banned in even the most permissive spaces because of their toxic nature. In the case of the harassment of Zoë Quinn, Quinn documented extensive coordination from IRC chat rooms, replete with participation from her ex-boyfriend. Theoretically, sustained harassment can take place entirely on a single platform without having to receive reinforcement from an outside platform, but I have come across no such instances.

When looking through the lens of online harassment, the internet is simply too small. When one platform links to another platform in these cases, it creates a pipeline of hate with very little friction. Even if the targeted platform maintains certain norms, the oncoming invaders ignore them, operating only under the norms of their originating platform. A simple Google search can connect together all the disparate aspects of a person's digital life, allowing bad actors to attack each and every part even without knowing them particularly well to begin with.

For persecuted individuals, there is no eternal frontier to flee to. Certainly one could retreat by deleting one's entire online presence, but this is not the promise of a boundlessly big internet. For targets of sustained online harassment, the internet is a one-room house full of speakers blaring obscenities at them.

Anti-harassment can take the form of smashing the speakers or turning off the electricity. Or it could take the form of turning down the volume, throwing a blanket over the speakers, giving people noise-canceling headphones, or even building new rooms in the house. Anti-harassment is about giving the harassed space on the internet, and keeping the electronic frontier open for them.

Conclusion:

The Two Futures of Anti-Harassment

Building a community is pretty tough; it requires just the right combination of technology and rules and people. And while it's been clear that communities are at the core of many of the most interesting things on the internet, we're still at the very early stages of understanding what it is that makes them work. - Aaron Swartz, September 14th, 2006

Online anonymity isn't responsible for the prevalence of horrible behavior online. Shitty moderation is. - Zoë Quinn, March 21st, 2015

I've discussed the shape of the problem—harassment as a spectrum of behaviors; harassment as a spectrum of content; and the effect of harassment on larger ecosystems of speech. I've also discussed the anti-spam industry as a useful comparison to anti-harassment.

I've laid out this picture in order to underscore the importance of addressing these issues. But throughout I've also engaged a pettier, more practical way to understand online harassment: Online harassment makes products unusable. Harassment blows up phones with notifications, it floods inboxes, it drives users off platforms. Harassment is the app-killer.

Put aside the very real harms of sustained harassment campaigns—the SWAT team visits, the bomb threats, the publication of addresses, Social Security numbers, or medical records. Even a low-level explosion of sub-threatening harassment should be of concern to tech companies, especially social platforms that rely on user-generated content, because these services have three dictates:

- Attracting users
- Building architecture to attract content from those users
- Removing or hiding the inevitable accumulation of garbage content.

Online harassment as content simply falls into the larger, broader, pre-existing category of garbage. Throughout the history of the internet, communities, open-source projects, and standards groups have grappled with the problem of garbage. Now corporations grapple with it as well, and have unfortunately defaulted to the worst stratagems of capitalism—first, by punting responsibility to consumers, and second, by outsourcing to underpaid contractors or, worse, to overseas sweatshops.

There are *decades* of collective experience out there on platform cultivation and community moderation from which the industry can draw. There are two futures for social media platforms. One involves professional, expert moderation entwined with technical solutions. The other is sweatshops of laborers clicking away at tickets.

Garbage collection should not be an afterthought. As outlined above, garbage collection that adequately addresses harassment is integral to a more egalitarian internet. But every social media company should take platform cultivation seriously. Rather than understanding it as non-technical support tacked onto a technical product, platform cultivation should be understood as a multidisciplinary effort that is integral to the product itself. The basic code of a product can encourage, discourage, or even prevent the proliferation of garbage. An engineer's work can exacerbate harassment, or it can aid a community moderator. Community moderation is not just about ex post removal of garbage—it is also about the ex ante dissemination of norms, as well as the collection of information that will best inform engineers on how to build out technical architecture in the future.

Right before Twitter rolled out the new Block button in 2013, Trust & Safety vociferously objected on the basis that it would magnify abuse. Block Version 2 worked the same way that the Mute button does now. Instead of blocking people from following and retweeting your account, it would simply make it so you couldn't see them. The experts strongly dissented internally, but it rolled out anyways. After public outcry, Block Version 2 was reverted back to the original Block within just 12 hours. It was later reintroduced as a new option: the Mute button.

The recommendations of Twitter Trust & Safety should never have been ignored. Moderators must have valued input in technical changes, and technical changes must be made in order to aid moderators. For example, one of the other things that Riot Games, the publisher of *League of Legends*, did to mitigate in-game harassment was to turn chat into an opt-in function. Players could still use it if they wanted, but only if they wanted. Hudson writes, "A week before the change, players reported that more than 80% of chat between opponents was negative. But a week after switching the default, negative chat had decreased by more than 30% while positive chat increased nearly 35%. The takeaway? Creating a simple hurdle to abusive behavior makes it much less prevalent."

What led the reforms that Riot Games instituted was a "player behavior team" of people with "PhDs in psychology, cognitive science, and neuroscience to study the issue of harassment by building and analyzing behavioral profiles for tens of millions of users." Riot Games assembled a panel of experts to design bespoke solutions for their product; their experts delivered.

What made *League of Legends* better wasn't an army of contractors in the Philippines, mass-banning, mass-deletion, the stripping of anonymity, or the pursuit of legal action. It was a handful of architecture tweaks and a user-run system of user accountability, designed by a dedicated team.

I can't dismiss the impact that an overseas warehouse of people answering tickets can have, but the drawbacks are obvious. Low investment in the problem of garbage is why Facebook and Instagram keep accidentally banning pictures of breastfeeding mothers or failing to delete death threats. Placing user safety in the hands of low-paid contractors under a great deal of pressure to perform as quickly as possible is not an ethical outcome for either the user or the contractor. While industry sources have assured me that the financial support and resources for user trust and safety is increasing at social media companies, I see little to no evidence of competent integration with the technical side, nor the kind of research and development expenditure that is considered normal for anti-spam.

Anil Dash wrote in 2011:

> *You should make a budget that supports having a good community, or you should find another line of work. Every single person who's going to object to these ideas is going to talk about how they can't afford to hire a community manager, or how it's so expensive to develop good tools for managing comments. Okay, then save money by turning off your web server. Or enjoy your city where you presumably don't want to pay for police because they're so expensive.*

People will never stop being horrible on the internet. There will never not be garbage. But in a functioning society, someone comes to collect the trash every week. If private platforms are to become communities, collectives, agoras, tiny new societies, they have to make a real effort to collect the garbage.

About the Author

Photo by James Bareham for The Verge

Sarah Jeong is a journalist and lawyer who serves on *The New York Times* editorial board. She was formerly a senior writer at *The Verge*, and has bylines at *The Atlantic*, *The Washington Post*, *The New York Times Magazine*, *Motherboard*, *Forbes*, *The Guardian*, and more. In 2017, she was named as one of Forbes's 30 under 30 in the category of Media.

Jeong graduated from Harvard Law School in 2014. As a law student, she edited the *Harvard Journal of Law & Gender*, and worked at the Electronic Frontier Foundation and at the Berkman Center for Internet & Society. She was a Poynter Fellow in Journalism at Yale for 2016, and also currently a fellow at the Internet Law & Policy Foundry.