Digital Ethics

Rhetoric and Responsibility in Online Aggression

Edited by
Jessica Reyman and Erika M. Sparby

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Contents

2 Hateware and the Outsourcing of Responsibility

James J. Brown, Jr. and Gregory Hennis

In 1999, Herring described a kind of algorithm for "gendered harassment on-line." Though she never specifically calls it an algorithm, she does offer a clear set of steps that she tracks in two different environments: Internet Relay Chat (IRC) and an e-mail list. In both situations, men harassed women by following a clear pattern, and the responses of women were dealt with in predictable ways: initiation of harassment, resistance to harassment, escalation of harassment, targeted participants accommodate to the dominant group norms and/or targeted participants fall silent (Herring, 1999). Herring links this pattern to the libertarian values baked into many online communities:

> Libertarian values of extreme freedom of expression are also present in both discussions, and benefit the most aggressive participants, who happen (not coincidentally) to be male. By maintaining … that any verbal behavior is authorized, no matter how crude or aggressive, males justify the use of dominating and harassing tactics in the name of free speech.
>
> (p. 163)

Herring's work is as important as it is depressing, since we appear to be living in the exact same world that she described 20 years ago. Contemporary work on online harassment identifies many of the same patterns Herring did, and the end goal remains the same: silencing women and other marginalized groups. Campaigns such as #GamerGate appear to be enacting the same algorithm that Herring saw in IRC and e-mail listservs.

Perhaps it is not fair to say that we are in the *exact* same world, since our current media ecology appears to provide harassers with a more extensive set of tools for coordinated and organized attacks. This has meant that scholars of online harassment have begun to couple the rhetorical analysis of harassment pioneered by Herring and others to the media infrastructures that enable and support such actions (Massanari, 2015; Phillips, 2015; Tarsa & Brown, Jr., 2018). Furthermore, the line between gendered harassment online and off is no longer

even clear, and the campaigns we now witness are not confined to our screens or devices. These media ecologies no longer only involve software and screens, since abusers deploy schemes that involve an extensive communications' infrastructure. Targets are swatted by abusers who inform law enforcement that they should send SWAT teams to a target's home, and search algorithms can be gamed to manipulate one's online identity. In short, the harasser has much more than language at his disposal if he wants to exercise his "free speech" to silence others. Following Herring, we see justifications of certain behavior in the name of free speech as misguided in at least two senses. First, the digital platforms we discuss are not operated by U.S. government entities and are thus not constrained by first amendment protections. Second, the use of free speech arguments to support certain abusive behaviors (and also used to justify the hands-off approach of software companies) seems concerned with the free speech of abusers without considering the free speech of abuse targets. Silencing people by way of harassing behavior shows little regard for the free speech rights of targets of abuse.

The assumption that online infrastructures must uphold free speech protections is part of a long and complex history, but our focus in this chapter is on how that assumption has created a useful set of tools for those who want to abuse and harass women, people of color, and LGBTQ groups. In particular, we see Section 230 of the United States Code as being a major contributing factor to our current predicament, a law that we discuss in more detail later in the essay (Protection for screening of offensive material, 1996). Section 230 freed websites of liability for content published by third parties, gave them the ability to make decisions about the publication of content without fear of being labeled a publisher of that content (and thus making them responsible for it), and provided no clear motivation for websites to police third-party content. By pushing this responsibility to users, the rhetoric of libertarianism has simultaneously empowered abusers and asked victims to fix the problem themselves.

This *outsourcing* of responsibility is one of the key features of Discord, the platform we discuss in this chapter. Our goal in this chapter is to describe some of the features of software platforms that enable harassers while disempowering targets and to describe those platforms with a provocative but, we think, useful term: hateware. This concept does not describe a stable category but rather a sliding scale on which we might place a range of platforms when trying to evaluate how or whether their features enable abuse and harassment. Some of these platforms appear at first glance to be general tools or utilities and others are explicitly created to help abusers and harassers. We recognize the problems with placing Twitter and Facebook in the same category as a platform like Hatreon, a short-lived crowdfunding platform for those who wanted to

avoid the hate speech policies of sites like Patreon or GoFundMe. While large platforms like Twitter have certainly pushed responsibility to users when it comes to dealing with harassment, it would be counterproductive to collapse the difference between it and sites that openly espouse racist and misogynistic ideologies. Still, by creating a continuum and understanding hateware as part of a broader infrastructural problem, we can begin to track the key features of software that props up and supports abuse and harassment, intentionally or not.

There are platforms that actively attempt to prevent abusive behavior, and these would likely sit outside the hateware spectrum. For instance, an ex-reddit employee named Dan McComas established a social media site called Imzy in 2016. Imzy was an attempt to address, at the level of software design and community norms, some of the problems McComas and others saw at reddit. As Tarsa and Brown, Jr. (2018) argue, Imzy developers recognized that the interface could be complicit in the problem of harassment, and its designers attempt to take a different kind of approach (Tarsa & Brown, Jr., 2018). Unfortunately, the site failed to gain a critical mass of users and shut down in 2017 (Buhr, 2017). The hateware spectrum ranges from "explicitly encourages harassment" to "implicitly encourages harassment," and software that combats harassment is (theoretically) doing neither. Software designed to openly attempt to combat harassment could conceivably find itself on the spectrum because certain design decisions have unintended consequences. However, this does not change the fact that some platforms can avoid being labeled as hateware.

Our hope is that the term "hateware" provides scholars with a theoretical approach to understanding the deep and sometimes invisible infrastructures of harassment. To this end, we start by explaining the term and how we're using it, then move to an extended analysis of the platform Discord. Discord was used to organize the Unite the Right event in Charlottesville, Virginia, which resulted in the death of Heather Heyer as well as two law enforcement officers. Those events put Discord in the spotlight and led to the company banning certain users from the service. However, this attempt to address bad actors and bad behavior after the fact ignores the ways the platform's design actually enables that same behavior. Because Discord uses similar structure and design techniques to other platforms, but also because of these recent events, we found it best to use Discord as our case study. We describe how some of Discord's structure and design decisions make it an example of hateware and how the crowdsourcing of community management (a libertarian infrastructure) creates fertile ground for abusive activity. We close by detailing how the concept of hateware might aid future research on abuse and harassment. In that light, it should also be mentioned that our study is best viewed as a template for analysis of software, rather than a final say on the matter of hateware.

Hateware

Much recent research addressing online harassment has attempted to bring to light how design is contributing to the problem of abuse and harassment, representing a broader interest in design that has long been at the core of digital rhetoric scholarship (Arola, 2010; Jeong, 2015; Kaufer & Butler, 1996; Selfe & Selfe, 1994). Sparby's (2017) work is a key example of this as she examines the "memetic rhetoric" that drives aggression and abuse in spaces like 4chan. The circulation of memes—the cultural units circulated through a community—helps to establish and fortify the identity of a community, and members of that community imitate behaviors in order to participate in identity formation. Sparby offers a detailed account of how this process happens, and she also argues that interface design is an integral part of that story: "users in online collectives often engage in memetic behavior influenced by the interface's technological design, ethos, and collective identity" (Sparby, 2017, p. 86). The design of 4chan and its community norms directly contribute to the community's collective ethos, one that lauds outrageous and shocking behavior. Sparby argues that the platform's anonymity, ephemerality, lack of user registration requirements, short-lived threads, and lack of an archive all contribute to the platform being a haven for hate (Sparby, 2017, p. 88). Interestingly, these design choices are actually put forth as a kind of *lack of design* by the 4chan community itself. This is especially clear when we examine that 4chan articulates its "rules" by saying that there are, for the most part, no rules (Sparby, 2017). As Sparby's research shows, there are indeed many rules, some of which are informed by humans (community norms) and others that are enforced by computation (software design), that establish a platform for a range of uses and abuses.

Massanari has conducted a similar analysis of reddit, a platform that is perhaps second only to 4chan in terms of its fame as a platform for trolling, abuse, and harassment. Massanari uses the term "platform politics" to describe "the assemblage of design, policies, and norms that encourage certain kinds of cultures and behaviors to coalesce on platforms while implicitly discouraging others" (Massanari, 2015, p. 8). She describes how reddit's karma system incentivizes bad behavior, including reposting material across multiple subreddits and making comments that rally behind the community's shared ethos of a "cyber/technolibertarian bent, gender politics, and geek sensibilities" (p. 9). Such activities increase one's karma score, affording more influence in the community. The platform also makes it easy to create multiple accounts, even after a user has been banned. In addition to the karma score and the behaviors it incentivizes, reddit joins many major platforms by providing very little recourse for those who are being harassed (p. 10). Even when moderators are empowered to step in and address abusive behavior, reddit's

hands-off ethos often wins the day: "Reddit administrators are loathe to intervene in any meaningful way in content disputes, citing Reddit's role as an impartial or 'neutral' platform for discussion" (p. 11). This idea that a neutral stance somehow stands outside of the fray is prevalent and is a direct result of a regulatory and cultural environment that insists on protecting the free speech rights of users at the expense of the safety of marginalized populations.

4chan and reddit are key examples of sites that are designed to invite certain kinds of behavior, and while they are often held up as a key example of toxic online communities, our analysis in this essay is an attempt to identify how all networked platforms encourage and deter certain kinds of actions. In fact, we aim to provide scholars with a theoretical approach that can be used to examine a broad range of software platforms and not just those that are so obviously filled with what Jane (2016) calls "e-bile." We use the term hateware to describe software that employs policies, algorithms, and designs that enable, encourage, and/or directly participate in abuse or harassment. This definition certainly accounts for sites like 4chan, but it also helps us examine less obvious examples of software that enables harassment. Software platforms can exist on a sliding scale, ranging from sites and services that openly promote harassment and abuse to those that enable bad actors in more pernicious ways.

Platforms can be positioned along the hateware spectrum and a platform's position on that spectrum can shift depending on changes in functionality, terms of service, or general design. Sites like 4chan and Twitter exist on the extremes, the former being a haven for hate, ignorance, and griefing, and the latter taking an extremely lax, inconsistent, and mostly user-driven approach to policing user behavior and enforcing community standards. Platforms that take any stance on abuse and begin to employ safer techniques can move from a status like 4chan's toward a similar status to Twitter, and some software may even graduate away from the spectrum and join the likes of Imzy. Software that continues to outsource responsibility (like Twitter) can begin tumbling down that slope, and head toward a status similar to 4chan. So, what are the key features of hateware? How do we know where software lies on the hateware spectrum? Given that this chapter is an attempt to open up this question rather than establish a final answer, we offer some provisional ideas about these key features in the hopes that future work will extend and revise our findings here.

In fact, there are many factors that may contribute to a piece of software inching toward hateware status. Our focus in this chapter is on how many platforms allow groups to govern themselves and have few broadly stated community norms. However, there are a number of other features common in hateware. For instance, some platforms selectively allow hypervisibility and simultaneous near invisibility to users, adopting

a haphazard approach to anonymity or pseudonymity (an especially complicated issue, given that anonymity can be wielded as a protective tool by targets of harassment). Some have unclear, lax, or vague policies and rules, employ algorithms that are easily gamed and manipulated, and take little initiative to police their user-base. However, hateware's key feature is the outsourcing of responsibility, and this design choice is largely due to the current regulatory environment. Such platforms advertise a hands-off approach, but commonly take steps to at least minimally moderate and filter content. Because they do practice some form of moderation, they appear to recognize their role in managing and addressing online harassment. However, they too often avoid the difficult task of addressing abusive and harassing behavior.

These platforms have begun to dominate the Internet thanks to Section 230 of the United States Code, which was established by the Communications Decency Act of 1996. Section 230 is a landmark portion of the U.S. Code that, among other things, grants immunity to online platforms when one user is harassed by another, whether or not the provider of the platform intervenes. Section 230 states that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider" (Protection for screening of offensive material, 1996). This language protects websites and social media services (as well as their users) from being treated as publishers. While publishers are at risk of lawsuit due to the publication of libelous or otherwise illegal content, many social media sites, such as Facebook, are not. More than this, the law allows such providers of "interactive computer services" to attempt moderation without opening themselves up to legal action:

> No provider or user of an interactive computer service shall be held liable on account of—
>
> A any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or
> B any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).
>
> (Protection for screening of offensive material, 1996)

Providers are allowed to moderate content, but they are not required to. Prior to the establishment of Section 230, websites *could* be held liable if they attempted to moderate anything at all, and this was thanks to the fact that traditional defamation law was applied to such cases prior

to the introduction of Section 230 in 1996. Section 230 was meant to provide an incentive for platform providers to moderate their content, by allowing them to retain immunity even when they moderate. In reality, this law has created fertile ground for abuse and harassment. Platform providers continued to neglect moderation, because what Section 230 failed to do was to provide a *disincentive for not moderating*.

This brings us to today, a moment when providers continue to do minimal work when it comes to moderation, allowing hate and abuse to thrive. Hateware, then, exists as a consequence of this new kind of software platform emerging in the 21st century. Unwilling to claim full responsibility but also unwilling to completely step away from moderation, platforms like Facebook and Twitter straddle the (admittedly blurred) line between a media organization and an information tool. In order to understand the role these sites play in the current problems with online harassment, it is crucial that we see their decisions regarding moderation as not just a matter of policy but also a matter of design. Through their design decisions, these platforms have (knowingly or not) built a set of tools that bad actors have seized upon in order to troll, harass, and abuse others.

This portion of the U.S. Code is arguably the most important piece of our current digital infrastructure and has been described by Eric Goldman, a leading legal scholar on Section 230, as "the law that gave us the modern Internet," the "most important law in tech," and "the law that makes the Internet go" (Goldman, 2017). Section 230 has been at the center of a number of court cases since its passage, most recently in *Fields v. Twitter* (2017), a lawsuit attempting to make Twitter liable for the actions of terrorists using their platform. In January 2018, the Ninth Circuit Court of Appeals ruled in Twitter's favor, though it largely "sidestepped Section 230, leaving it [sic] applicability for another day and another court" (Goldman, 2018). Section 230 continues to be at the center of many legal and legislative battles, but its impact on our contemporary digital infrastructure is undeniable. Provided with legal coverage by Congress and the courts, not only have companies taken a hands-off approach to managing communities and policing harassment, they have actually shifted that responsibility to their users by crowdsourcing such activities.

Sowing Discord

Discord is a platform that has emerged as an important part of the conversation surrounding software and online harassment. At first blush, it seems to be little more than a chat application. However, our analysis demonstrates that Discord is much more than this and that its design provides those aiming to abuse and harass others with key tools for doing so. Designed to provide gamers with the ability to chat while gaming,

it is akin to IRC chat rooms, and it allows users to create their own servers to host like-minded conversations. Each server is divided into different channels, which are similar to chat rooms. So, each server has one or more channels, and each user can be a part of multiple servers. There is also a permissions system in place: the creator of a server can assign different roles to users with different permissions. Permissions dictate what users with particular roles can and cannot do, from what channels they can post in all the way down to *what* they can post in those channels. The permissions' system is most often used in hierarchy style: the server owner can have the most power and highest permissions, and they can create other roles, such as administrators and moderators.

Discord also gives users the option of remaining anonymous. There is not necessarily a real name attached to any Discord account: just a username, password, and an e-mail that realistically a user only ever needs access to once, to confirm the account. Users can also create a temporary free account that is automatically deleted if it is not confirmed via an e-mail address. Discord doesn't *require* anonymity, and users can link social media and gaming accounts to Discord accounts (such as Facebook, Twitter, YouTube, reddit, Battle.net, Twitch, and Steam). Even if users decide to connect these accounts, they have the option of not showing them on their profile. Much of what takes place on Discord takes place behind the scenes in a sense: user servers can be (and most often are) configured to be private and only accessible by desired parties (though, we will demonstrate below that infiltration of servers is indeed possible). This is a departure from the way platforms like Twitter and reddit work, because the latter platforms focus on providing the possibility of users to interact with those outside of their immediate social networks. In addition to attempting to silo conversations with its server model, Discord also allows a range of other features, including video calling and screen sharing, as well as an API that allows for the integration of Discord directly within specific games.

Discord is an interesting case study in hateware given that, at first glance, it is little more than a platform that makes communication among gamers easier. Players of online games such as *Overwatch* can establish a server and easily use Discord to chat while playing. Thus, one might initially balk at the idea that Discord enables abuse and harassment, and we would not fault such a response. If Discord is hateware, then what about e-mail platforms? What about text messaging? We have two responses to such questions. The first is that any platform or application could indeed be placed on the hateware spectrum, depending on design choices and features offered to users, and this is why we insist on understanding hateware as a sliding scale rather than a category. Our second response is that, upon closer analysis, Discord is much more than a mere chat application. Discord's design offers features that many have exploited to abuse, harass, and troll. We want to focus on Discord's

general approach to content moderation as well as techniques used by those abusing others on the platform. Discord outsources the labor of community management to users, a design decision that is core to nearly any social networking application or platform in use today. This establishes an environment that is rife with bad behavior. Those wanting to harass others have used a number of techniques, including the use of invitation links that allow users to invite others to join their server.

Discord could not exist without the extensive labor carried out by its users. The platform offers very little in the way of community management, and it relies on users to report bad behavior. As we have noted, this is the norm with contemporary websites and services, and it is largely due to Section 230 establishing an environment that disincentivizes any attempt to actively filter content or police behavior. It also allows companies like Discord to avoid paying a labor force of content moderators and community managers. Even after raising massive amounts of venture capital in 2017, Discord had hired only five customer experience personnel and zero moderators (Menegus, 2017). Discord provides tools to those administering servers; however, they do not necessarily make life very easy for server administrators. For instance, the default setting for any Discord server is "no verification," meaning that if an administrator makes no changes to the settings, they are inviting anyone to join the conversation without having to verify their identity (Ravenscraft, 2016). This is, of course, an open invitation to those who want to disrupt a server.

A number of journalistic accounts have pointed out that Discord has shown little concern for harassment. Menegus of *Gizmodo* has offered a frank evaluation of Discord's approach to this problem: "From what I've seen, users who wish to engage in harassment, raid servers, or bombard chats and users with child pornography suffer no lasting repercussions for doing so" (Menegus, 2017). While this approach changed after the Unite the Right controversy (which we address later), the overall approach from Discord appears to be one of unconcern. One Discord user quoted by Menegus even describes a number of de facto moderation techniques that administrators have had to develop on their own. For instance, many administrators have had to create entire channels for new users, a kind of holding pen for the abusers that arrive to disrupt the conversation. This strategy results in a kind of "manual verification process, where new users can only talk in two rooms which the majority of the server have muted" (Menegus, 2017). Creating these containment rooms (one can imagine a chatroom in which harassers scream at one another without an audience) has now become a normal operating procedure for administrators attempting to deal with harassment.

Those using Discord to harass others have developed a range of techniques, and we will focus on the most common: server raiding. Discord uses invitation links to allow users to send invitations to particular

servers, and these invitations can be set to expire after a period of time determined by the user. This is an easy way to invite people to a server, but it is also an easy way to infiltrate a discussion and cause chaos. This kind of action is known as a raid on a Discord server, and it happens fairly regularly. Invitation links can be generated by those posing as members of the community. These users essentially embed themselves in order to later invite other trolls and abusers. One can sit in the weeds, participate in the discussion, and after a certain amount of time gain the permission to generate invitation links. They can then invite others to raid the server, meaning that a swarm of users arrives to post offensive material and to overrun the conversation with thousands of simultaneous posts. However, these invitation links are not only generated by harassers—they are often used by members of a server to draw in new members. They are a convenient way to gain community members, and therefore many server administrators continue to use them even if they do invite raids. One server administrator described the problem to Menegus:

> Some time ago, the owner of the server put out a few invite links on Reddit, and as such, the search term "furry discord" brings up the Reddit post containing said link as one of the top results.... The simplest way to stop these raids is to revoke said link, however doing so cuts off the main source of legitimate people wishing to join.
>
> (Menegus, 2017)

Thus, server administrators are caught between creating the possibility of growing their community and inviting raids on their server.

Discord claims to be updating the software to address raids, which the company says is a violation of their terms of service and is against their values. However, certain updates seem to have actually exacerbated the problem. For instance, one 2017 change log explains that the default expiration time for links was changed from 30 minutes to 24 hours, "so less people run into dead ends and more people get to hang with their fun friends" (Nelly, 2017). Furthermore, the same change log provides a claim that Discord has upgraded systems to address raiding, but we are given no details. Instead, we get cute Internet lingo: "A ton of internal systems have been added to combat spam and raiding. THEY ATTAC. WE PROTEC" (Nelly, 2017). Discord has also attempted to address this problem with verification levels, which establish "a basic level of security a user must meet before they're allowed to send text messages in a channel" ("What are verification levels?," n.d.). Implementing such changes does attempt to address raids and the creation of invitation links by nefarious actors. However, it still places the onus on users (those administering servers) to address the problem. This design decision is framed as the one that gives users freedom, but

the result is that administrators are tasked with dealing with raids and attacks on a regular basis. Apparently, with freedom comes overwhelming responsibility.

To fully understand how Discord has responded to problematic users, it's best to turn to the event that put Discord in the spotlight—the Unite the Right rally and subsequent protest in Charlottesville, Virginia. Originally designed for gaming communities, Discord does play host to other conversations, including a collection of alt-right organizations. These organizations communicated (privately and out of sight) with each other on Discord to organize the rally, plan raids on other servers, and target individuals for abuse. The ultimate impact of their digital actions (a real-life hate rally) was immense and demonstrated that online harassment does not necessarily remain online. The actions of these Discord users exploded into the real world, and that seemed to be where Discord began to draw the line. The Unite the Right rally was not organized overnight and the hatred exhibited was not spontaneous: Discord servers that helped organize the event (such as National Socialist Army, Führer's Gas Chamber, Blood and Soil, /pol/, and Centipede Central) had been hosting forms of abuse since before the creation of this messaging client. But in addition to raiding other servers, groups were also organizing internally. Their outwardly facing harassment doesn't always reveal what is happening within the walls of the servers, where "they posted swastikas and praised Hitler" (Roose, 2018). Hatred was the norm even inside those walls, with defamation of certain groups of people and blatant hate speech running rampant. According to *New York Times* reporter Kevin Roose, even infighting between these various servers was common, meaning no one was safe.

The biggest offense of the alt-right on Discord, however, *was* outwardly facing (and extremely so): the Unite the Right rally in Charlottesville, Virginia, which led to the death of Heather Heyer. This form of harassment was not innovative. In fact, it was depressingly banal. Simply by exploiting the platform and its features, users of the alt-right servers were able to organize themselves and coordinate a massive meet-up. There was no special tactic used here, like the ones we discussed earlier on in this chapter, no server raids, no invitation link sharing. They merely took advantage of Discord's hands-off approach to community management in order to hide in plain sight.

After years of being aware that their platform had been utilized by the alt-right, white supremacists, and neo-Nazis, Discord finally began taking action by eliminating servers and issuing warnings to others (Newton, 2017). After the Unite the Right rally, Discord said that it was taking a proactive, long-term stance to deal with those violating terms of service. Interestingly, they policed this behavior by, once again, relying on users to report bad actors: "Though we do not read people's private messages, we do investigate and take immediate appropriate action

against any reported ToS violation by a server or user" (Alexander, 2018). However, Discord also claims to track certain patterns that are associated with raids, such as the use of bots to swarm a server, meaning that their trust and safety team is not only relying on user reports. The details of these measures are closely guarded by the company: "We do not disclose the exact measures we take as we don't want to give people clues for how to work around those measures" (Alexander, 2018). Still, Discord insists that they never listen in and that they are not policing the content of a server.

This insistence of remaining at a remove from content and staying out of the nitty gritty of actual language is in keeping with the hands-off nature of the libertarian bias of many digital spaces, a bias that we can see at least as early as the harassment studied by Herring. Language is not policed, because this would violate the free speech ethos that still dominates many digital platforms. Discord will go to great lengths to not police actual speech (or to convince users that they are not doing so) for fear of alienating users who would see such policing as off-putting. But Discord can also just as easily claim that it is following current U.S. law, which does not require them to engage in this way. Furthermore, the very draw of Discord for alt-righters (and perhaps others) is the anonymity and privacy of Discord servers. This is key to their platform, and a violation of it would actually transform their service. Discord remains adamant that the company will only track conversation on a metalevel, and that they don't and will not track content. They have removed people for abuses of terms of service in their messages, but only after receiving tips via other users. All of these ways of dealing with abuse deal in technicalities. One way to understand this is that Discord avoids the question of ethics altogether. Another way to understand it—the way we choose to read the situation—is that the company's ethics are embedded in the platform, its policies, and its community. However you understand it, it should be clear that Discord's approach leaves a lot of wiggle room for offenders. Alt-right servers such as The Right Goys and Centipede Central are still active even after Discord banned a number of servers in the wake of Charlottesville (Liao, 2018).

Design Justice and Insourcing Responsibility

In a 2017 *BuzzFeed* story, Discord CEO Jason Citron laid out the company's design philosophy in clear terms: "We're very focused on making an amazing communication product for gamers.... I had a hunch that it would be used outside of gaming, but it wasn't anything we thought specifically about" (as quoted in Bernstein, 2017). When asked about the popularity of his platform among the alt-right, Citron essentially shrugged his shoulders: "It's inevitable that there will be actors using the product for things that are not completely wholesome"

(Bernstein, 2017). Seven months later, it became clear just how unwholesome some of Discord's users were. Those organizing the Unite the Right rally in Charlottesville had exploited many of the platform's features to gather together white supremacists at an event that resulted in the death of Heather Heyer and two law enforcement officers. The most prominent of these organizers was Jason Kessler, who proclaimed on Twitter that Heyer was a "a fat, disgusting Communist" and said her death was "payback" for deaths caused by Communists (Pearce, 2017). Discord had been a key tool for Kessler and others, something that Citron knew about in January. However, he seems to have seen such uses of his platform as collateral damage, something that would inevitably happen in a space that took a hands-off approach to moderation. Even after Charlottesville, Citron has said that the company has no plans to change its model of relying on users to report bad actors (Crecente, 2017).

But what would Discord look like if Citron and the rest of the design team "thought specifically about" the other potential uses of Discord? What if software designers began to think more deeply about how their platforms might be enabling bad behavior and designed these platforms with such potentialities in mind? How might the designers of these platforms *insource*, rather than outsource, responsibility? This is the kind of approach many scholars are beginning to advocate for. For instance, Costanza-Chock has argued for a "design justice" framework that sees design as directly tied to questions of racism, sexism, and domination:

> For example, at the personal level, we might explore how interface design affirms or denies aspects of a person's identity through features such as, say, a binary gender dropdown during account profile creation. More broadly, we might consider how design decisions play out in the impacts they have on different individual's biographies or life-chances. At the the community level, we might explore how platform design fosters certain kinds of communities while suppressing others, through setting and implementing community guidelines, rules, and speech norms, instantiated through different kinds of content moderation systems. At the institutional level, design justice asks us to consider the ways that various design institutions reproduce and/or challenge the matrix of domination in their practices.
>
> (Costanza-Chock, 2018, p. 5)

We believe that the term "hateware" can be a key part of a design justice approach since it presents critics and designers with a way of analyzing and evaluating platforms that are enabling the abuse of marginalized people. Understanding the rhetoric and ethics of harassment will require

that we do more than just analyze the language and actions of those carrying out this abuse. As many scholars have already noted, the analysis of bad behavior will have to be coupled with an understanding of how our digital infrastructure is a crucial part of this story.

Discord's name suggests from the very start that disharmony and disagreement are core to what users might expect. As Milner (2014) has argued, disagreement and agonism are important for healthy, thriving communities. Drawing on the work of Chantal Mouffe, Milner argues that we should notice the difference between agonism and antagonism in online communities. Antagonism describes argument between "enemies," while agonism describes disagreement between "adversaries." Discord's problem with raiding demonstrates that its siloed model provides for agonism on particular servers but also too easily allows for *antagonism*:

> Antagonism… [pushes] voices out of the public sphere. For the logic of lulz [to] afford vibrant, agonistic public discourse, multiple perspectives and counter perspectives should be evident. Voice should be evident over exclusion, even if that voice is not monolithic in content or tone.
>
> (Milner, 2014)

Antagonism is built on exclusion and silencing others, and certain design decisions can encourage that kind of behavior. Discord is but one example of a site that has made such design decisions without fully recognizing the far-reaching implications of those decisions. Section 230 and a broader notion of free speech drive a great deal of design on the Internet, meaning that nearly every site and service on the Internet outsources responsibility.

The concept of hateware is our attempt to understand how design can help address problems of abuse, harassment, and antagonism by thinking more carefully about the design of software and community standards. Rather than addressing these problems after the fact, design justice can attempt to address them when building platforms and communities. Rather than outsourcing responsibility, design justice can insource it. To prevent the problems of hateware, we will need to identify and diagnose the portions of software that are easily gamed toward nefarious ends and then learn from those lessons as we attempt to build software that avoids landing on the hateware spectrum.

References

Alexander, J. (2018, February 28). Discord is purging alt-right, white nationalist and hateful servers. *Polygon*. Retrieved from https://www.polygon.com/2018/2/28/17061774/discord-alt-right-atomwaffen-ban-centipede-central-nordic-resistance-movement

Arola, K. L. (2010). The design of web 2.0: The rise of the template, the fall of design. *Computers and Composition*, *27*(1), 4–14.

Bernstein, J. (2017, January 23). A thriving chat startup braces for the alt-right. *BuzzFeed News*. Retrieved from https://www.buzzfeed.com/josephbernstein/discord-chat-startup-braces-for-the-alt-right

Buhr, S. (2017, May 24). R.I.P. Imzy. *TechCrunch*. Retrieved from http://social.techcrunch.com/2017/05/24/r-i-p-imzy/

Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. *Social Science Research Network*. Retrieved from https://papers.ssrn.com/abstract=3189696

Crecente, B. (2017). Discord: 87M users, nintendo switch wishes and dealing with alt-right. *Rolling Stone*. Retrieved from https://www.rollingstone.com/glixel/news/discord-87m-users-switch-dreams-dealing-with-alt-right-w513598

Goldman, E. (2017). The ten most important Section 230 rulings. *Social Science Research Network*. Retrieved from https://papers.ssrn.com/abstract=3025943

Goldman, E. (2018, January 31). Twitter didn't cause ISIS-inspired terrorism. *Fields v. Twitter. Technology & Marketing Law Blog*. Retrieved from https://blog.ericgoldman.org/archives/2018/01/twitter-didnt-cause-isis-inspired-terrorism-fields-v-twitter.htm

Herring, S. (1999). The rhetorical dynamics of gender harassment on-line. *The Information Society*, *15*, 151–167.

Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum*, *30*(3), 284–297.

Jeong, S. (2015). *The internet of garbage*. Jersey City, NJ: Forbes Media.

Kaufer, D. S., & Butler, B. S. (1996). *Rhetoric and the arts of design*. Mahwah, NJ: Lawrence Erlbaum.

Liao, S. (2018, February 28). Discord shuts down more neo-Nazi, alt-right servers. *The Verge*. Retrieved from https://www.theverge.com/2018/2/28/17062554/discord-alt-right-neo-nazi-white-supremacy-atomwaffen

Massanari, A. (2015). #Gamergate and The Fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

Menegus, B. (2017). How a video game chat client became the web's new cesspool of abuse. Retrieved from https://gizmodo.com/how-a-video-game-chat-client-became-the-web-s-new-cessp-1792039566

Milner, R. (2014). FCJ-156 Hacking the social: Internet memes, identity antagonism, and the logic of lulz. *The Fibreculture Journal, 22*. Retrieved from http://twentytwo.fibreculturejournal.org/fcj-156-hacking-the-social-internet-memes-identity-antagonism-and-the-logic-of-lulz/

Nelly. (2017, July 21). 7.21.17—Change Log. Retrieved from https://blog.discordapp.com/7-21-17-change-log-c9acad667d67

Newton, C. (2017, August 14). Discord bans servers that promote Nazi ideology. Retrieved from https://www.theverge.com/2017/8/14/16145432/discord-nazi-ban-white-supremacist-altright

Pearce, M. (2017). Tweet from the account of Charlottesville rally organizer insults slain protester Heather Heyer. Retrieved from http://www.latimes.com/nation/la-na-charlottesville-organizer-20170818-story.html

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture* (Reprint edition). Cambridge, MA: The MIT Press.

Protection for screening of offensive material, 47 U.S. Code § 230 (1996). Retrieved from https://www.law.cornell.edu/uscode/text/47/230

Ravenscraft, E. (2016, August 17). Discord is the voice Chat App I've Always Wanted. *Lifehacker*. Retrieved from https://lifehacker.com/discord-is-the-voice-chat-app-i-ve-always-wanted-1785403197

Roose, K. (2018, January 20). This was the alt-right's favorite chat app. Then came Charlottesville. *The New York Times*. Retrieved from https://www.nytimes.com/2017/08/15/technology/discord-chat-app-alt-right.html

Selfe, C. L., & Selfe, R. J. (1994). The politics of the interface: Power and its exercise in electronic contact zones. *College Composition and Communication*, *45*(4), 480–504.

Sparby, E. M. (2017). Digital social media and aggression: Memetic rhetoric in 4chan's collective identity. *Computers and Composition*, *45*, 85–97.

Tarsa, B., & Brown Jr, J. J. (2018). Complicit interfaces. In W. S. Hesford, A. C. Licona, & C. Teston (Eds.), *Precarious rhetorics* (pp. 255–275). Columbus: Ohio State University Press.

What are verification levels? (n.d.). *Discord Server Setup*. Retrieved from http://support.discordapp.com/hc/en-us/articles/216679607-What-are-Verification-Levels-