

Networks of Care

Introduction

One of the earliest memories of my childhood is my parents asking me to go off the internet so that they could make a phone call. My whole life so far has been connected through the internet, and a bit later, with social media. When I was 11 years old, my friends were already posting pictures online in a social network called *Windows Live Spaces*. That was the platform where, one day after school, I read a lengthy hate comment about me. My first experience with hurtful comments was unexpected, but over time, it would become obvious that hateful content is uploaded to social media every day.

Throughout this text, I will focus on online hate that harms others. The word *hate* can accommodate a lot of actions, and it's hard to identify them without context. The most common demeanours are harassment, bullying, stalking, racism, threats, intimidation. These problems are getting more attention as the repercussions of online behaviours leave the screens to persist in our bodies. Furthermore, research shows that marginalised groups are bigger targets of online hate. (Silva et al., 2016) In this way, it becomes urgent to address some pending issues. How can we deal with social platforms that amplify hate against the users?

There is no single solution to end hate but diverse ongoing approaches. A fair answer is to insist on responsibility either from the government, tech companies, or international organisations. Laws, such as the NetzDG Law in Germany, are admirable initiatives. The NetzDG law (or the *Netzwerkdurchsetzungsgesetz*) gives legal importance to flagging, complaining, and reporting inside platforms. This law puts trust in governments to manage hateful content and, at the same time, preserve free speech. Not every country can rely on a responsible regime. However, these laws can set an example and encourage necessary legal discussions. Next to changes in official structures, it's stimulating to look at bottom-up strategies initiated by users.

This essay focuses on community movements that moderate social platforms. It demonstrates how users create systems to enforce their views on what is acceptable or not inside their online networks. In the first part, I look into digital vigilantism through *cancel culture*, an approach to callout problematic users. In the second part, I dive into Codes of Conduct, another efficient way to manage behaviour. In the last part of the text, I explore design tools that can hold off hate, such as *blocklists*. In the forefront of the fight against hate, there are users committed to creating better social media experiences for them, and for others.

User movements follow informal sets of rules which are clear for a specific community but often scatter over different groups and platforms. It is also true that online traces are often lost, movements morphed into others. This text offers a better understanding of online behaviours that establish counter-hate communities. As a designer, a media student and a social media user, I understand that what my online community encourages or dismisses deeply shapes me. Is it possible to fight hate within the platforms battlefield?

Fighting hate with hate, the case of cancel culture

In this chapter, I will take *cancel culture* as a case study to discuss how, with digital vigilantism, users oversee their social spaces. Cancel culture is an ongoing movement that prosecutes hateful content outside the conventional approaches. Online vigilantism brings users together to supervise social networks, creating communities with shared mindsets, rules, goals. For the users of social platforms, cancel culture allows a collective moderation of online content, through particular contentious methods.

Cancel culture evolved from the need to raise awareness for problematic behaviour online. When a mediatic figure does something unacceptable in the eye of the public, the outrage begins. Users shame others for reasons such as using hate speech, writing racist comments, making misogynist remarks or any other behaviour users perceive as unreasonable. The number of people that participate in the callout affects how viral the reaction on social media is – the shamed may lose followers, sponsors, job opportunities, or suffer other kinds of punishment. In short, they become *cancelled*. As the researcher Lisa Nakamura explains, in the attention economy, when you find someone not worthy of your attention, you deny them their sustenance. (Bromwich, 2018) Cancel culture happens the most through Twitter, Instagram, Facebook. The platforms that reach the mainstream audience are the most massive platforms for public opinion.

When it started, cancel culture was supporting the voiceless. The users behind the movement wanted to establish a more caring society, to show concern for marginalised groups that are frequently silenced and harassed on social media. It was an activist attitude. Safer online networks can only exist if hateful behaviour is regulated, especially if the misconduct comes from prominent identities. The online profiles of renowned brands, businesses or celebrities are powerful channels in which ideas broadcast to a vast number of people. With cancel culture, users criticised the careless exposure of hateful content, mainly coming from high-profile members of social spaces. And for the first time, if the outrage against a powerful identity was loud enough, it would reach them. The movement challenged the status of the elites that can often avoid the consequences of their bad behaviours.

The popularity of cancel culture brought problematic situations to the attention of a lot of people, which put pressure on gatekeepers to vocalise their opinions too, deciding what is allowed or not inside their platforms. Cancel culture pushes social platforms to act politically towards users, something that these businesses have been avoiding. In the US, publishers such as traditional newspapers curate content, so they have responsibility for what is published. US laws declare that an "interactive computer service" is not a publisher. (Communications Decency Act 1996) This means computer services can't be held accountable for what their users publish. Facebook is a computer service. However, when it starts banning content and deciding what is appropriate content, it's making editorial decisions. There's still some confusion about which legislation social media businesses should comply with.

Faced with the uncertain role of platforms, cancel culture had a particular aim: pursue social justice. The act of shaming always existed, but it gained a lot of momentum with social media. Some authors believe it's a characteristic of the technologically empowered yet politically precarious digital citizen. (Ingraham and Reeves, 2016) Ineffective politics pushes users to react, transforming shaming culture in meaningful political participation. According to Ingraham and Reeves, publicly shaming others distracts us from a larger crisis we seem to have little control over. It also allows us to perform agency on an obtainable smaller digital scale. The person shamed becomes a scapegoat. People blame them for a more significant issue. The accusers feel relieved that they identified and removed who

was causing a problem when, in reality, one person can't be the cause of all the recurring issues with social media.

Peter Thiel was the first external investor of Facebook as a result of his interest in the ideas of the philosopher René Girard. (Shullenberger, 2016) According to Thiel, Facebook was destined for success because people have a “mimetic” basis of desire. This expression means we have an instinct to copy and compare, to mimic everyone's behaviour and so desire what the others have. Facebook follows human nature's desire of imitation: the whole platform revolves around the events your friends went to, where they were, with whom. We *like* it. Social media is also the perfect environment to become resentful for not having what others have. The outcome is anger, violence and, eventually, scapegoating. Cancel culture, and other movements of vigilantism, do point to one person to make it a case. Holding someone accountable can be done in private, but cancel culture turns it in a public example of moral standards.

The R. Kelly case is an excellent example of how cancel culture evolves. R. Kelly is a very famous musician, recently arrested for multiple sex crimes. Over 20 years, the allegations were growing immensely but without any court conviction. His prominent presence on social platforms was seen as a systematic disregard for the well being of black women. The need for justice started a social media boycott under the name #MuteRKelly. Users felt he shouldn't be featuring in music streaming platforms, or continuing his career in general. Cancel culture supports the idea of first believing the victims, a concept supported by the #MeToo movement. The website muterkelly.org explains the reasons for the boycott.

By playing him on the radio, R Kelly stays in our collective consciousness. (...) That gets him a paycheck. That paycheck goes to lawyers to fight court cases and pay off victims. Without the money, he's not able to continue to hide from the justice that awaits him. It's not an innocent thing to listen to him on the car to work. That's what helps continue his serial sexual abuse against young black women. That makes us all an accomplice to his crimes. (#MuteRKelly, 2018)

People were encouraged to boycott him by sharing #MuteRKelly in all platforms. Report or perform similar actions on music streaming services, post about it as much as one could. At this time, Spotify removed R. Kelly from the auto-generated playlists and introduced the button *don't play this artist* across the platform. Some users were calling it the *R. Kelly button*, as the moment for the release of the feature seemed very connected with the boycott. Later, Spotify reversed all decisions. According to *Spotify Policy Update* of June 2018, “[At Spotify] we don't aim to play judge and jury.” The apprehension from Spotify to act adds to the discussion about the role of social media businesses. Does it fall onto the users or the platforms to fight the problematic topic of hate speech?



Yes!! It. Is. Not. A. Drill.

#MuteRKelly


Dan Rys  @danrys
Spotify Removes R. Kelly Music From Its Playlists As Part of New Hate Content & Hateful Conduct Policy: Exclusive bit.ly/2lbjZ9K

Fig – #MuteRKelly on the web

Unfortunately, hate draws attention. There is a term used in the art world for such a phenomenon. *Succès de scandale* is a French saying from the *Belle Époque* in Paris, meaning success from scandal. In 1911, Paul Chabas painted *Matinée de Septembre* portraying a nude woman in a lake. The nudity of the piece caused controversy. Several complaints culminated in a court case against the public exhibition of the painting. The conflict was dramatic: city council was making laws to prohibit nudes, meanwhile gallery owners were purposely placing copies of Chabas' work on their windows. This increased public's interest in the controversial painting.

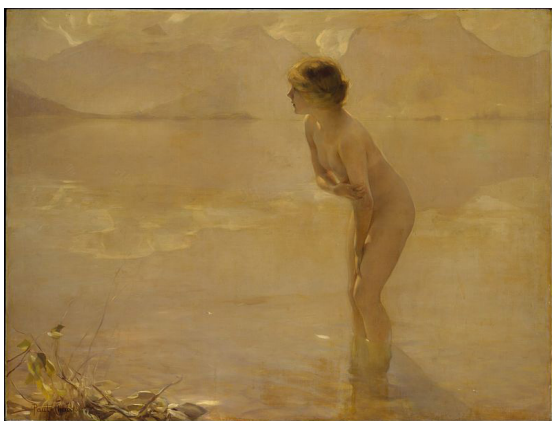


Fig – *Matinée de Septembre*, 1911

Likewise, on social media, scandals can be convenient. Engagement comes from negative or positive reviews, dislikes or likes. Social media rewards attention, even if this attention comes from absolute outrage. Views from haters on a youtube video will generate revenue for the creator. It is cold-blooded, but hate can bring the creator profit. Because of the way social media systems function, the virality of shaming benefits the social media business model. (Trottier, 2019) The success from scandal shows how cancel culture may fail to hold someone accountable through shaming. R. Kelly eventually went to prison, but many other celebrities enjoyed the status of the victim. It is possible that public shaming brings popularity to the offender, cancel culture excels in creating viral content.

Cancel culture uses techniques to spread quickly and gain visibility by finding its way to the trending topics, through hashtags, using popular location tags. The trends page of Twitter is a special place of interest. When an expression is used in abundance by the users, it gains a position of attention in the platform. On Twitter, the *trends* show by default. In the desktop version, they appear side by side with your profile, and on the mobile version they are on the search page. The *trends* are a pervasive feature of Twitter. This makes the words #MuteRKelly reach millions of people and spread the word to boycott this person. The *trends* created a stage to cover all kinds of messages, which is only possible because of the platforms design decisions.

Platforms are, without a doubt, essential moderators or promoters of user behaviours. Twitter *trends* are a prime example of weaponised design. Although they can be a news source, they also favour the mob mentality, typical in online trolling and harassment. Andrea Noel is a Mexican journalist. Through her investigation, the journalist obtained access to internal emails from a Mexican troll farm from 2012 to 2014. Troll farms are organisations that employ a vast amount of people to create conflict online, to distract or upset users. In the emails, Noel read how these people organise to divert online attention from important issues. One of the strategies was the fabrication of trending topics on Twitter. This falsification means that #FridayFeeling can be a topic tweeted every second by a company in Mexico to avoid #MuteRKelly to reach the trends. Publishing vast amounts of noise in social media prevents other conversations to happen.

Faced with Noel's research, I wanted to have a better understanding of the popularity of boycotting through what's trending on social media. For that reason, I created a bot that looks for trends in the United States related to cancel culture. The bot collects the trending topics methodically and saves them so I can look at them later. It listens for specific words I know are correlated with cancel culture, but I may be missing other specific hashtags of which I'm not aware yet. The bot isn't perfect, and it doesn't need to be. Throughout the time it's been running, it illustrated some of the activity of the users with digital vigilantism. In my research, in November 2019, *Halsey*, *Lizzo*, *Kpop stans*, *Uber*, *Amber Liu*, *John Bolton's book* and the cartoon *Booboo* all reached the trends to be boycotted. Lizzo made sexualised comments about a group of singers. Amber Liu spoke in favour of a racist arrest in the US. Both were actions viewed as morally condemning, which provoked a reaction on social media.



Fig – Bot to collect hashtag activism

The social platforms that are present in the daily lives of most people are focused on gathering attention. Attention comes from exaggerated actions, just like violence, harassment or SCREAMING. The friction benefits the social media business model, but not the well-being of the users. Since the #MuteRKelly movement, cancel culture gained other expressions. To become mainstream, cancel culture needs to be viral. Aggressive. Definitely scandalous. To spread awareness to the most social media users as possible, cancel culture started to ignore essential details of a story to escalate the situation to an unverified version that was more attractive. Just like tabloids and reality tv, people enjoy consuming reputations as entertainment. It makes sense that the popularity of sensationalism seen in magazines or the tv works as well on social media.

Although there are groups of people committed to using *cancel culture* as an instrument to call out hate, it's essential not to forget the ones who solely enjoy putting others down. The power to denounce others can be abused by who is already in a position of privilege. Boycotting also discards forgiveness, it turns away possible allies for the social issues it tries to bring attention to. Calling out bad behaviour, especially from marginalised groups, is a community-driven movement that follows bespoke rules. Users participate in social spaces in their conditions, forcing their visions of what should be unacceptable. Finding consensus on what constitutes hate speech, harassment or bullying is difficult, even

more on vast public spaces such as platforms with billions of people. At the end of 2019, Facebook reports showed 1.66 billion daily active users on the platform. (Facebook, 2020) This is a considerable fraction of the world population going on Facebook every day. It's impossible to aim for moral consensus amongst so many people, so some groups use their voices to push the values they would like to see applied over the platform.

The idea of pursuing justice collectively, allowing users participation in demanding accountability and change in ever-evolving society norms, fueled cancel culture. But is it possible to do it without following the same techniques as their opponents, where innocent people can become targets of a mob? Is it possible to build safe social networks in spaces that promote outrageous viral comments? It's impossible. Fighting hate with hate had controversial outcomes. Cancel culture today is an unforgiving movement, a massive confusion of harassment, shaming, fake morality, and a lot of pointing fingers. Nonetheless, controlling hate is still a very urgent issue.

New platforms, different rules

As seen in the previous chapter, users become digital vigilantes to denounce hateful content within social media platforms. Another strategy that has been receiving a lot of attention is the development of rigorous Codes of Conduct. The engagement from the community to create new guidelines supports the stronger regulation of online spaces.

Creating rules is essential. An explicit structure that is available and clear to every member makes space for participation and contribution. The lack of governance doesn't avoid the presence of informal rules. (Freeman, 1996) An unregulated group causes stronger or luckier users to establish their power and own rules, which prevents deliberated decisions and conscious distributions of power to be done at all. For this reason, users welcome the creation of Codes of Conduct within social media networks. A Code of Conduct is a document that sets expectations for users. It's an evidence of the values of a community, making clear which behaviours are allowed or discouraged, possibly decreasing unwanted hate. A Code of Conduct is very different from contractual Terms of Service or a Use Policy. Instead, it's a non-legal document, a community approach.

I followed the interesting public thread of discussions in CREATE mailing list, archived from 2014. This list shares information on free and open-source creative projects. The back-and-forth of emails discusses the need for a Code of Conduct in an upcoming international convention. One of the concerns is the proliferation of negative language in many Codes of Conduct. The group wishes to reinforce positive behaviours, instead of listing all the negative ones. A statement of what constitutes hate will indeed create a list of negative actions, but will that foreshadow a bad event? The discussion deepens. Is there a need for a Code at all? Some believe the convention is already friendly, while others feel that it is a privileged statement. A member compares the Code with an emergency exit, useful when you need it.

This CREATE thread is proof that what is obvious for us, may not be obvious for others. The mailing list was debating a physical event, but also online, where distance, anonymity and lack of repercussions dehumanise interactions, it's critical to be aware of the principles of our social networks. A useful Code of Conducts should make a clear distinction between anti-harassment policies and other general guidelines. (Geek Feminism Wiki, 2017) Is harassment publishing a personal address online? Is harassment a hurtful comment? It's essential to agree on the definition because it forces the group to make explicit decisions on what will be considered misconduct. A Code of Conduct doesn't only set rules but also includes people who are responsible for managing reports and possible malpractices. In this way, community rules are not only documents but labour intensive routines that imply human effort and involve the community.

An example of a massive platform that challenged their members to discuss misbehaviours was *League of Legends*. The online game *League of Legends* drives a powerful sense of sociality. The users create profiles, role-play different characters and form networks. The users have to work together in a team, and therefore the game provides chat tools for the players. The *League of Legends* has its formal documents – it specifies terms of use, privacy policies, support files. But the guidelines that govern the community are under the *Summoner's Code*. The *Summoner's Code* is a Code of Conducts that formulates the behaviours expected from the gamers. The *League of Legends* is an intriguing case to look at because it not only implemented community rules, but it also had a *Tribunal* where the community discussed the misconducts.

When users reported a gamer for frequently breaking the Code of Conduct, the case would go to Tribunal. For example, the reason for the report could be the explicit use of hate language. In the Tribunal, the system attributed the case at random to some users. It provided to each *judge* the statistics of the game where the incident happened, their chat log and the reported comments. The minimum of 20 users reviewed each case and then decided to *pardon* or *punish* the offender or *skip* the case as a whole. In the end, the most popular vote prevailed. The type of punishment, whether it was a warning, suspension or even banning, wasn't decided by the users, but by a member of the game administration team. In their profiles, the *judges* could see the cases, the outcomes of the decisions, and a personal ranking. This system was very popular, over the first year that the Tribunal was online, it collected more than 47 million votes.

The League of Legends' Tribunal is, in essence, a court of public opinion. In a very similar way to the actions described in the first chapter, there is a community that enjoys being vigilant of others. In online forums where people reminisce about their time with the Tribunal, a lot of users seem to miss it. Some users reflect how proud they were for removing toxic players from the community. Others remember how the Tribunal made them entertained. However, one of the problems for the developers of the game was the time the Tribunal needed to achieve a decision, especially compared to automated systems. Nowadays, there are still platforms relying on community rules and human choices, but most of them deal with misconducts in private.

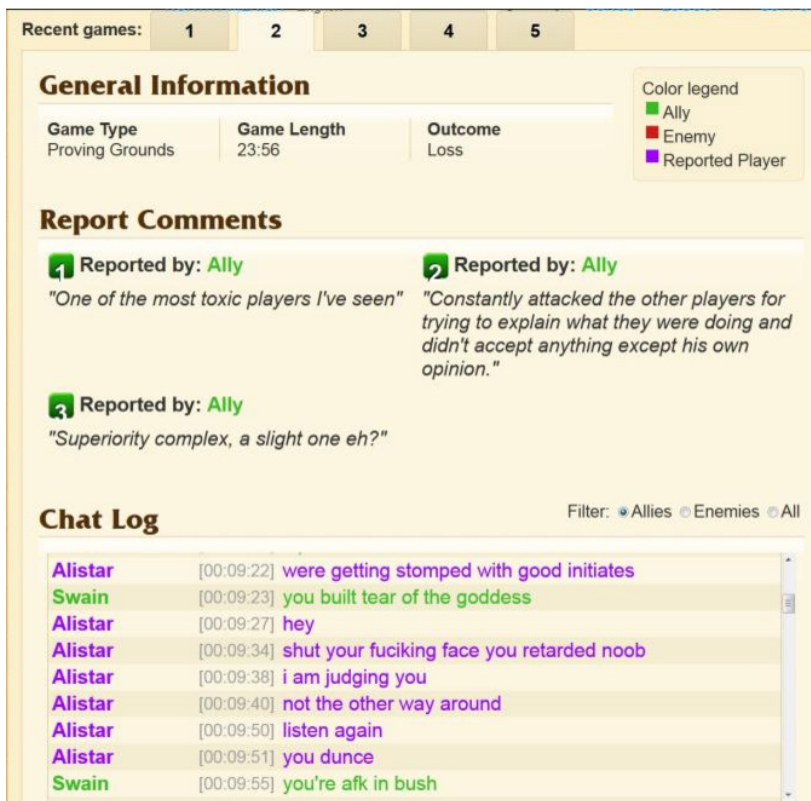


Fig – A Tribunal case

A social platform that promotes a diversity of guidelines within their community is Mastodon. Mastodon is a social media with microblogging features, similar to Twitter or Facebook. It is a community of communities, a federated and decentralised social media platform built on the open-source ActivityPub protocol. Decentralisation means the distribution of authority. Each server can implement their visions while sharing a common platform. Federation entails that users from different groups can socialise with each other, but everyone has their experience more tailored to their liking. Practically, while sharing

the same platform, a user can be part of a group which blocks a kind of content while another group allows it.

On the platform, the different community groups are called *instances*. Navigating through them, reveals the different rules sanctioned by the users. *Mastodon.social* is a prevalent instance created by Mastodon leading developers which has at the moment 459,189 users. As most Mastodon communities, there is a Code of Conducts that serves as guidelines for user-behaviours. These are informal rules moderated by the community, not legal documents. In mastodon.social one can understand there will be no tolerance for racism, sexism, casteism, violent nationalism, and many more. The writers of the Code of Conducts also took the time for clearly defining harassment.

I reached out to one of the moderators of mastodon.social. The most common situation in the other instances is that the moderators are voluntary members of the user base. However, this moderator explained how, in his instance, the Mastodon project pays for three moderators. On top of that, there are some volunteers. The CoC (Code of Conducts) of this group doesn't request anything farfetched. It operates at the essential level of human decency. However, I wanted to know how often they needed to remind someone of the rules and enforce the sanctions. The answer was that, although this instance has a lot of users, a lot are not active enough to disrespect the guidelines. The active ones follow the rules. A small community is indeed much easier to regulate, so the underlying structure of a federated project such as Mastodon already facilitates the moderators' jobs.

It's important to understand that user rules don't follow any particular view on morality. *CounterSocial* is another instance on the platform that blocks entire countries, such as Russia, China, Iran, Pakistan or Syria. The administrator of this cluster is *The Jester*, a well-known digital vigilante. The instance asserts that blocking countries aims to keep their community safe by not allowing nations known to use bots and trolls against the West. It can seem dubious behaviour, but this is entirely legitimate on Mastodon. The community is independent to create its guidelines. They choose who to invite and block from their network. The last question of CounterSocial frequent questions says it all: "Who defines these rules, anyways?" It's them!



Fig – banned countries

A Code of Conduct doesn't deter all misbehaviours, but in platforms that allow users to impose their rules, social media users can mitigate online hate in a much more direct way. Just like in cancel culture, community rules prosecute bad behaviours inside their system not having as a first defence the police or a governmental identity. However, in a very different approach from cancel culture, the repercussions of not following the conduct are predominantly dealt with in private. The people who manage the community have the role of moderating. The moderators make use of warnings, blocking, banning. While some groups have zero-tolerance policies, others employ more forgiving proposals – “If the warning is unheeded, the user will be temporarily banned for one day in order to cool off.” (Rust Programming Language, 2015)

An online conversation with the administrator of the Mastodon instance *witches.live* brought to light how the bottom-up initiative of moderating hate is a co-operative task. In this community, there are three moderators. The admin handles all tech work and daily maintenance, but two other persons bring additional perspective. The distribution of power is an excellent way of ensuring a diversity of viewpoints and avoid personal bias. Similarly to the feedback of other moderators, this group doesn't need to discipline very often. The written rules already form boundaries that keep people who don't agree with them away. For this moderator, the most critical part of making decisions is to understand where there is an “honest mistake” and a “trolling bigot”.

Mastodon offers an overview of some groups on the platform through its Application Programming Interface (API). Out of the top more populated servers listed, there is a home for all people: artists, developers, activists, gay men, and a whole big portion dedicated to adult content. It's not only marginalised communities that are enjoying more controlled networks. The idea of building safe spaces where users can be active participants and moderators of their social networks is proactive and resonates to a lot of people. However, safe spaces open the doors for fascists to make their protected networks as well. *Gab* is a social platform that advocates for free speech with no restrictions. Its terms of use don't ban bullying, hate, racism, torture or harassment. The only point that briefly mentions any liability is when to engage with actions that may perceive physical harm or offline harassment. Before 2019, its brand was the face of *Pepe the frog*, an alt-right symbol. As expected, *Gab* is known for hosting a lot of hateful content.

In 2019, *Gab* forked from Mastodon their custom platform. The migration was an attempt to dodge the boycott it was facing. Apple Store and Google Play had removed *Gab*'s mobile app from their services earlier. Although a lot of Mastodon communities have already their rules against racism and can block others that don't, *Gab* still benefits from the platform system as a whole. There was a lot of controversy on whether Mastodon should ban *Gab*'s instance as a general platform policy. In this case, the platform as a company felt pressure to intervene beyond community-driven rules. For the founder of Mastodon, the only possible outcome was to allow *Gab* in the fediverse. This situation upset some users. The perceived inadequate response to moderation of the alt-right from Mastodon was one of the reasons for the creation of *Parastat*.

Parastat is a new social media under development that aims to contribute to a more humane society. Their general Code of Conducts, for all users, is much stronger than Mastodon's. *Parastat* promises immediate ban for hate speech, threats or harassment. Beyond the norm of other platforms, it also doesn't allow flirting, conspiracy theories, homoeopathy, healing crystals and many others. *Parastat* is very serious in their moderations policies. In the present online environment where hate proliferates, there are enough reasons to build safe spaces. Creating online networks where people come together, can express themselves and feel protected from outside abuses. A CoC applied in the context of social media takes the stand that platforms will not welcome everyone. In this way, the rules challenge the idea of having social networks open for everyone. Strict moderation policies, such as the ones in *Parastat*, will always polarise social media users due to different ideals of freedom of expression.

Conspiracy theories, pseudoscience and scams

It's important that our communities are places for constructive ideas, and not ones that are used as weapons, tools of disinformation, or spreading information that can cause harm.

Some kinds of conspiracy talk and discussions of the occult and woo are fine, but these are examples of the kinds of things that are unacceptable here:

- Anti-vaxxer
- Climate change denial
- Flat earth
- NWO, Illuminati, Chemtrails, Pizzagate, etc.
- Hate group conspiracy theories like Holocaust Denial.
- Pseudoscience like Homeopathy, 'healing crystals', etc.
- 'Creation Science' and other religious forms of pseudoscience.

Promoting MLMs (Multi-Level Marketing schemes) or Pyramid

Fig – policies of Parastat

Codes of Conducts make the intentions of a social space known. As explained at the beginning of this chapter, groups with no rules don't exist. At best, there are groups with no rules announced. In such way, if the members acknowledge the goals of a community, this action can support users that understand each other better. The emergence of Codes of Conduct on social media provides more agency to the users. Users choose how they want to interact with their networks. On a big scale, is it possible to manage billions of different-minded people with one set of rules? Small communities seem more capable to regulate online hate, as it's easier to share similar ideals. What can mainstream platforms with massive amounts of people do? Big platforms still have a long way to go in the way they manage hate, but one crucial step is to work on their policies – to be straightforward on what constitutes hateful actions and how they won't tolerated. Isn't it the time to accept one social network can't cater to all?

Designing change

Throughout this text, I analysed the popularity of vigilantism and the development of Codes of Conduct. Still, it is compelling to mention the potential of software tools. Users build tools outside the formal development of social media businesses to moderate content on their terms. Together, the community shares notions of morality and customises their platforms, gaining more control over the way they participate in their networks. The interface is a crucial component of social media to deal with online behaviours. The design shows the actions we can do, what and how we see content on the platform. Add-ons, plugins, and other tools can be very efficient in avoiding hate when they tweak, remove or add to the design of the interface. In this way, to begin this chapter, it's necessary to understand the importance of interface design.

In 1990, Don Norman wrote that “the computer of the future should be invisible” (Norman, 1990), meaning that the user would focus on the task they want to do instead of focusing on the machine. Much like a door, you go through it to go somewhere else. But the designer and researcher Brenda Laurel reminds us that closed or opened doors allow different degrees of agency. A door that opens for you, a small door for children, a blocked door: the interface defines the user role and establishes who is in control. What the platform allows the user to do, the possibilities for a person on social media to write, post, and reach others, are *affordances* of the platform. The term *affordance*, as Norman has interpreted it, is now a buzzword in the field of design.

If platforms have intrinsic characteristics that guide user behaviours, social platforms become responsible for the way users share hate, especially if they facilitate or perform abusive actions. To understand how platforms can accommodate hate is valuable to look at *Yik Yak*, a former social media app targeted at college students. The platform allowed users to post messages to a message board, in anonymity. The privacy policy of *Yik Yak* did not approve the identification of the users without specific legal action. The app bounded a small community as the user would only see the posts of people around them. *Yik Yak* was anonymous and local. It was also community-monitored. Users upvoted or downvoted the posts of the message board, and as a result, the upvoted messages would be more visible on the interface. The app launched in 2013, and at one point in 2014, *Yik Yak*'s value reached 400 million dollars. Only three years later, the developers published a farewell note, and the app shut down.

One day at college, the student Jordan Seman saw a horrible message about her and her body on *Yik Yak*. The hyper-localisation of the app meant that whoever *yaked* the insults, was very very close to her. She then would write an open letter to her school and peers, where I found her story. The letter was published for the Middlebury College community, but it definitely resonated to other groups using the app. *Yik Yak* is a significant example of weaponised design. The features of the platform could allow a close self-regulated community. Instead, the same characteristics tolerated the spread of hate on college campuses without any accountability. The message board was a *burn book*, a place to vent, to make jokes about others, to bully. In the case of *Yik Yak*, the platform design facilitated the shaming of Jordan. She asks in her open letter – “Is this what we want our social media use to be capable of?” (Seman, 2014)

Yik Yak's structure is very similar to Reddit. *Yik Yak* also maintained message boards, allowed pseudonyms and kept a karma system. Identical design choices on Reddit, its algorithm and platforms politics, have been analysed and implied to support anti-feminist and misogynistic activity. (Massanari, 2017) It's clear that platforms affordances deeply shape user behaviours. In this way, it's not surprising that while *Yik Yak* developers were

dealing with hate on their platform, the same was happening at Reddit. In August 2014, a controversy around the gaming industry culture instigated coordinated attacks, mainly targeted at women. The movement spread and escalated with the usage of the hashtag Gamergate on Twitter. The repercussions of such actions were hateful. The #gamergate harassment included doxing, intimidations, SWAT interventions, life threats, bomb alerts, and shooting warnings.

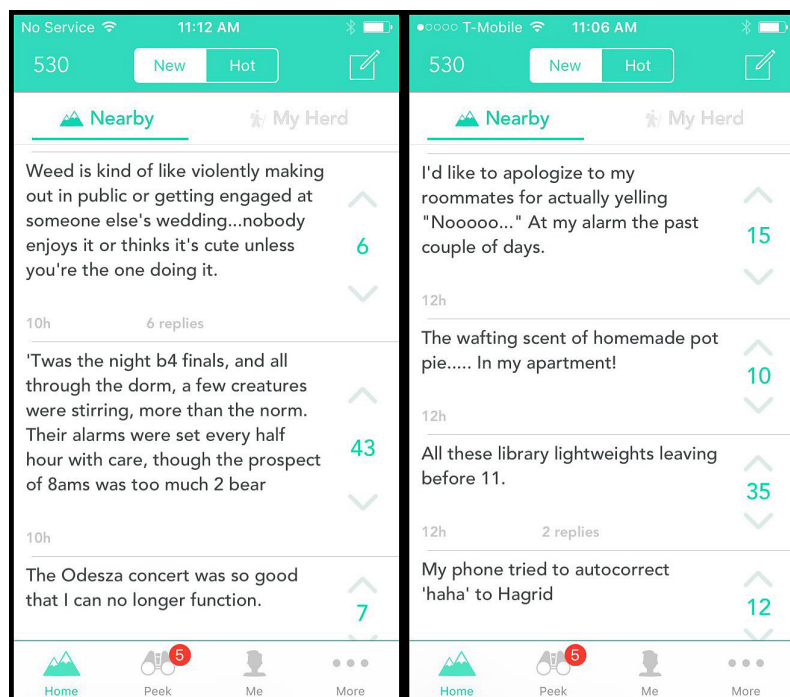


Fig – Yik Yak

The interface can act as an agitator. For this reason, technical tools to reduce hate through the interface become meaningful and required. A feature that allows shutting down harassment is to stop listening to the source by blocking the user. However, there are some situations where individual blocking is not enough. As a result of the Gamergate controversy, the developer Charles Hutchins created *BlockAllTwerps*, his block list on Twitter. *BlockAllTwerps* programmatically collects and blocks users that are harassing, following or retweeting harassment. (Hutchins, 2016) When a user subscribes to a blacklist, their feed will ignore the presence of the people added to the list – no tweets, notifications, messages. In a broad sense, if a user subscribes to *BlockAllTwerps*, they will stop seeing content from potential harassers. The idea of who should be blocked derives from Hutchins' ideals. The mass blocking may reproduce discriminating views of the developer, and the creator of *BlockAllTwerps* is well aware of it.

Feminists have used mass blocking strategies before Gamergate. The first shared block list was *TheBlockBot*. It maintained a list with three levels of strictness, level 1 for users who posted hateful content until level 3 for microaggressions. Shared blocklists are developed and supported by the community. They are bottom-up strategies to individually and collectively moderate Twitter experiences. (Geiger, 2016) A community co-operates a list, deciding on who is listened to or silenced. Blocklists follow shared views of morality, ruling themselves by what each member feels is harassment, hate speech, or any target the list has. The practice of preserving a blocklist creates an informal structure, a network of affection. Some of the tasks of the members of the group include adding more people to the list, removing some, explaining the reasons for the block, providing tech support, dealing with complaints.

Blocklists use a different approach to cancel culture to reduce hate. Blocklists don't aim to remove problematic users from online spaces but choose instead to not engage with them. Users who use block bots are not escalating a discussion but trying to stay away from it. The action is generally quiet, as a person may not even detect it was blocked. However, if they do, it may raise some questions about the reasons why it happened. Different people manage the list, so who is blocked or not doesn't reflect strict guidelines. In the process of adding someone to a blocklist, it is common to add the reason for such blocking. In one hand, the explanation adds disclosure for users. On the other hand, it shames the dreadful users and their behaviours. Bots like *TheBlockBot* give email addresses to forward the complaints. Although there's a word of advice – "...make peace with the possibility that some people on twitter may not wish to talk to you and that's okay." (TheBlockBot, 2016)

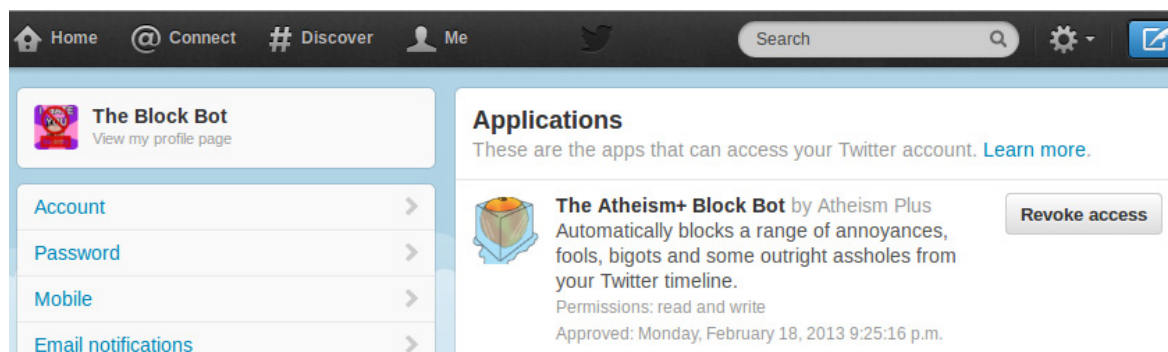


Fig – The Block Bot on Twitter

Software approaches reshape the way users interact with social platforms. Voluntary developers create blocklists because of the lack of a comparable feature on the platform. Even before block bots, Twitter users helped each other identify people to block by posting the hostile user id on the public timeline. In 2015, Twitter CEO Dick Costolo would write in a leaked internal memo "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years." (Independent, 2015) On that year, Twitter added the feature to share block lists into their source code. Today, sharing who is blocked is not available anymore, so blocklists continue as parallel activities. Nonetheless, since 2015 a lot more attention has been given to moderation on social media. Not only the platforms have been researching and trying new approaches, but also more plugins, extensions and bots are created every day.

Nowadays, there is a variety of tools useful for social media users. *Tumblr Savior* is a web browser extension for Tumblr that allows the user to choose which words they want to blacklist or whitelist. *Social fixer* or *F.B. Purity* filters Facebook's news feed. For Youtube, *Channel Blocker* blocks separate videos, users or whole channels. An open directory online is full of suggestions from the community to add features that could decrease hate on social media. A user suggests blocklists, such as *BlockAllTwerps*, others want the ability to disable replies. Some would like to improve *muting*. It's not uncommon for grassroots tools to turn into real features on social platforms. For example, on Twitter, *flagging* started as a petition from 120,000 users that wanted more report mechanisms to deal with online abuse. (Crawford and Gillespie, 2014)

Flagging takes the expression of the nautical red flag, meaning danger, a warning, and on social media, a report of something improper. It is a method for users to show discontent towards something or someone. In some platforms, the action of flagging is binary – the user is either against the content or not. In others, flagging is more thorough. For example, Youtube asks for the user to choose from nine options why the video violates community guidelines. Flagging can be capable of removing hate, mainly when used as a collective tool. As the outcomes of individual flagging are often undisclosed, it is frequent that a community organises and demands change by using the tool in cooperation with others. A

call to action is posted online for users to use the report button against some post, or user. This amount of feedback will put pressure on the platforms to act — to remove someone from the network, for example.

Flagging is a feature on a lot of social platforms, and a tool to moderate content. However, users can use it to report all kinds of things, including genuine valid material. Different users of social media can use flagging in varied ways, which explains how tools are just a means to do something. They don't obey single handling but rely heavily on the user. An unfortunate example of flagging is the report of female biographies on community platforms such as Wikipedia. Last year, the flagging and subsequent removal of women pages generated a lot of commotion and media coverage. Wikipedia members used the flagging system to ask for the removal of pages of several women, in a platform that already lacks female contribution and exposure. As of February 2020, only 18.3% of biographies in the English Wikipedia are about women. (Denelezh, 2020)

Alongside the flagging, Wikipedia is interesting to analyse for their other software tools. Without moderators, the task to edit content on Wikipedia articles is the result of the public collaborative discussion between users. As anti-hate measures, the editors get help from tools such as ClueBot NG, ORES or the AbuseFilter extension. These software tools detect and remove hateful content. The tools are always evolving to more sophisticated forms, for example, with the implementation of machine learning. The automatism of moderation is becoming a common practice on social media. But so far, the intricate nature of hate and its context, still require a lot of human action. Until someone comes up with better social solutions, technical tools can help users to deflect the hateful content.



Fig – One of the biographies deleted on Wikipedia that generated most controversy.

In this chapter, I discussed the possibility of creating tools on the margins, as complements or plugins, as is the case of blocklists. Also relevant is the manipulation of some already integrated features, like flagging. The openness of forums is also a great place to conceptualise what tools are needed. The technical tools referred to in this text are used within coordinated strategies to help shape social spaces. They are generous approaches to filter out hate from the users' networks. The tools, when used collectively, share software knowledge, design skills, and media know-how. This cooperation is especially helpful for users without the resources to make adjustments that can make a difference in their experiences with social media. The community that shares their knowledge, and is active in removing hate for themselves and others, creates important support systems — networks of care.

Conclusion

Online hate exists since people could share messages on computers. In 1984, a bulletin board system called Aryan Nations Liberty Net was carrying racist material, years before the internet was widespread. Two decades later, the participatory web 2.0 foreshadowed a cultural revolution. The potential for social media to connect people grew, as well as the ability to spread mean comments, harass someone, make threats. To say that online spaces are filled with hate is not a novelty, but a commonplace at this point. However, the ways to deal with hate continue to increase and improve, always trying to stay as progressive as possible, aiming to catch up with the most recent hurdles. Discussions about moderating social platforms are challenging issues making the headlines right now.

Throughout this text, I pinpointed the multiplicity of efforts to reduce hateful content from the users perspective. Users, fed up with encountering harmful behaviour online, started coming up with their ways of protecting and maintaining their networks. Valuable clusters of people organise on the margins to make social media spaces more enjoyable. The communities that grow within these actions build networks of care. I suggest that these bottom-up strategies are essential to the existence of social networks. Official responses from the platforms are necessary, but I propose that informal community movements are crucial to managing social platforms and that they deserve more attention, debate and recognition.

The point where it gets complicated is where to embrace bottom-up strategies such as Codes of Conduct, and where these rules limit freedom of expression online. Another problem with some approaches, such as cancel culture, is that they can assume moral righteousness, where someone's morality becomes superior to others, and therefore more important to spread through media. Karl Popper's Paradox of Tolerance clarifies the impossibility of allowing everything and being completely tolerant. The philosopher has explained how it's essential to set boundaries to create a truly tolerant society. I believe the same applies to social media platforms. In Popper's words, "We should therefore claim, in the name of tolerance, the right not to tolerate the intolerant." (Popper, 1945) Going forward, it's evident that a big, open for all, social media doesn't work. Platforms need to implement clear values, set goals, social rules, and ways to enforce them.

In my artistic practice, I'm giving attention to Codes of Conducts by recording, annotating, commenting, drawing or transcribing with people that are working with these community rules. I understood that it was crucial to give space for a variety of stances on the topic and create a mesh of informative experiences. On this text, I recognised broader approaches to reduce hate content online, such as digital vigilantism and technical tools. And still, a lot of paths weren't mentioned — manifestos, protests, low tech devices, memes — these are all appropriate strategies. It's noticeable for me that these actions aim for a future where it's possible to achieve an ideal social media experience for every user.

While writing this thesis, I was, at times, feeling defeated. At the beginning of my research, I was trying to put the blame for online hate on the interface, on platforms' gatekeepers, on capitalism if it seemed feasible. In the end, this work doesn't focus on hate towards others on social media but sheds light at the intricate communities that work against it with no formal responsibilities. These networks of care share ideas and mindsets to what should be acceptable, and work as collectives to cut down hateful behaviours from their social spaces. Even if the outcomes are dubious at times, these are very generous approaches to moderate social media. There's now a clear answer to the question I contemplate on the introduction — is it possible to fight online hate? Absolutely.

Bibliography

- Bromwich, J.E. (2018) Everyone Is Canceled. *The New York Times*, 28 June.
Available at: <https://www.nytimes.com/2018/06/28/style/is-it-canceled.html> (Accessed: 9 February 2020).
- Communications Decency Act 1996*, section 230.
Available at: <https://www.law.cornell.edu/uscode/text/47/230> (Accessed: 5 January 2020).
- Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18 (3): 410–428.
- Denezh (2020) *Gender Gap in Wikimedia projects*.
Available at: <https://www.denezh.org/> (Accessed: 28 February 2020).
- Facebook, Inc. (2020) *Facebook Reports Fourth Quarter and Full Year 2019 Results*.
Available at: <https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-Fourth-Quarter-and-Full-Year-2019-Results/default.aspx> (Accessed: 2 March 2020).
- Freeman, J. (2013) The Tyranny of Structurelessness. *WSQ: Women's Studies Quarterly*, 41 (3–4): 231–246.
- Geek Feminism Wiki (2014) *Community anti-harassment/Policy*.
Available at: https://geekfeminism.wikia.org/wiki/Community_anti-harassment/Policy (Accessed: 9 March 2020).
- Geiger, R.S. (2016) Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19 (6): 787–803.
- Hutchins, C. (2016) @BlockAllTwerps. *Art Meets Radical Openness (#AMRO16)*.
Available at: https://publications.servus.at/2016-AMRO16/videos-lectures/day2_charles640lq.mp4 (Accessed: 8 March 2020).
- Ingraham, C. and Reeves, J. (2016) New media, new panics. *Critical Studies in Media Communication*, 33 (5): 455–467.
- Rust programming language (2015) *Our Code of Conduct (please read)*.
Available at: https://www.reddit.com/r/rust/comments/2rvrx/our_code_of_conduct_please_read/ (Accessed: 9 February 2020).
- Massanari, A. (2017) #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19 (3): 329–346.
- MuteRKelly (2018) *Why Mute R. Kelly?*
Available at: <https://www.muterkelly.org/about> (Accessed: 8 March 2020).
- Norman, D. (1990) Why interfaces don't work. *The art of human-computer interface design*.
Available at: https://www.academia.edu/2849717/Why_interfaces_don_t_work (Accessed: 9 February 2020).
- Popper, K.R. (1945) *The open society and its enemies*. London: Routledge.
- Seman, J. (2014) A Letter on Yik Yak Harassment. *The Middlebury Campus*.
Available at: <https://middleburycampus.com/27709/opinion/a-letter-on-yik-yak-harassment/> (Accessed: 9 February 2020).
- Sherwin, A. (2015) Twitter CEO: "We suck at dealing with abuse and trolls". *The Independent*, 5 February.
Available at: <http://www.independent.co.uk/news/people/news/twitter-ceo-dick-costolo-we-suck-at-dealing-with-abuse-and-trolls-10026395.html> (Accessed: 4 February 2020).
- Shullenberger, G. (2016) The Scapegoating Machine. *The New Inquiry*.
Available at: <https://thenewinquiry.com/the-scapegoating-machine/> (Accessed: 2 March 2020).
- Silva, Mondal, Correa, et al. (2016) Analyzing the Targets of Hate in Online Social Media. *International AAAI Conference on Web and Social Media*.
Available at: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13147/12829> (Accessed: 3 February 2020).
- Spotify (2018) *Spotify Policy Update*.
Available at: <https://newsroom.spotify.com/2018-06-01/spotify-policy-update/> (Accessed: 8 March 2020).
- The Block Bot (2016) Introduction to @TheBlockBot.
Available at: <http://archive.is/WJ19U> (Accessed: 8 March 2020).
- Trottier, D. (2019) Denunciation and doxing: towards a conceptual model of digital vigilantism. *Global Crime*, pp. 1–17.