MY SOCIAL MEDIA IS A BATTLEFIELD

Introduction

One of the earliest memories of my childhood is my parents asking me to unplug the internet cable 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 how social media grows in hateful content 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. It often includes unwelcomed harassment, bullying, stalking, racism, threats, intimidation. These problems are getting more attention as the repercussions of online behaviours leave the screens to persist on our physical bodies. Moreover, 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 questions. What can we do against violent social platforms filled with hate that harm 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*) is controversial, but it aimed to give legal importance to flagging, complaining, and reporting inside platforms. Not every country can rely on a democratic government. However, these laws can set an example for so many social media companies that are US based, as well as European data centres. These legal discussions deserve more encouragement.

Alongside, it's stimulating to look at bottom-up strategies. In the forefront of the fight against hate, there are users moderating content, writing rules, building tools. In this essay, I will highlight community movements that are regulating online platforms. 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 emerging way to manage behaviour. In the last part of the text, I explore design tools that can hold off hate, such as *blocklists*. The three parts show how users create systems to enforce and prosecute their views on what is acceptable or not inside their social networks.

User movements follow informal sets of rules which are clear for a specific community but often scatter through different groups and platforms. It is also true that online traces are often lost, movements morphed into others. This text attempts to offer 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?

Chapter 1 — Fighting hate with hate, the case of cancel culture

In this chapter, I will take *cancel culture* as a case study to discuss how digital vigilantism becomes a way for users to assert their agency. Cancel culture is an ongoing movement that prosecutes hateful content outside the conventional approaches. Online vigilantism brings users together to oversee social spaces on their own terms, creating communities with shared mindsets, rules, goals. For the users of social media, 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. Cancel culture starts when a mediatic figure does something unacceptable in the eye of the public. Therefore, they become cancelled. Users shame immoral deviants for reasons such as hate speech, racism, misogynism, or any other behaviour that they perceive unacceptable. The number of users that participate in the callout affect how viral is the reaction on social media – the shamed may loose followers, sponsors, or suffer other ways of online punishment. In the attention economy, when you find someone not worthy of your attention, you deny them their sustenance. (Nakamura cited in Bromwich, 2018) Cancel culture happens the most through Twitter, Instagram, Facebook. Capitalistic platforms are currently the mainstream, so they are the most massive platforms for public opinion. (Partido Interdimensional Pirata, 2019)

The cancel movement wanted to establish a more caring society, to show concern for marginalised groups that are frequently silenced and harassed on social media. Safer online networks can only exist if hateful behaviour is regulated, especially if the misconduct comes from prominent identities. Social media accounts of renowned brands, politicians or celebrities are powerful channels in which ideas broadcast to a huge number of people. Instead of accepting indisputable platforms, users were criticising mindless exposure of hateful content, particularly from high-profile members of social spaces. Cancel culture started as a movement of compassion for the voiceless — an activist attitude. And for the first time, if the outrage against a powerful identity was loud enough, it would reach them.

Cancel culture also puts pressure on 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 on which legislation social media businesses comply.

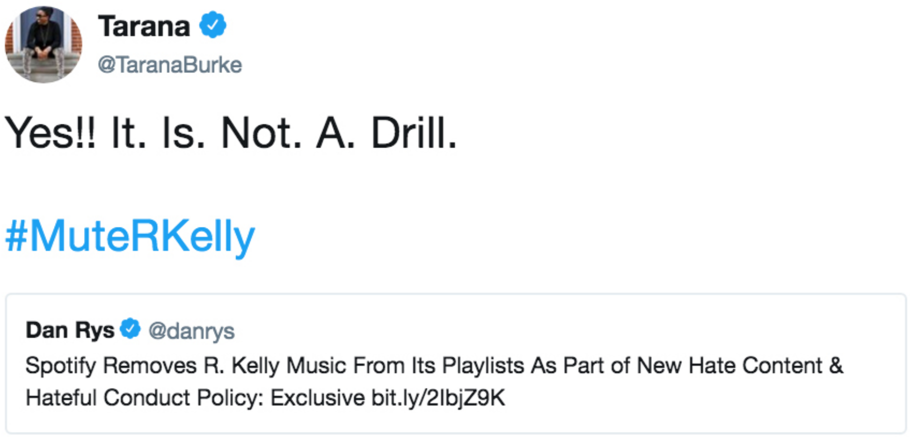
Faced with the uncertain role of platforms, cancel culture had a particular aim: pursue social justice. (Trottier, 2019) The act of shaming always existed, but it gained a lot of momentum with social media. Some authors believe it's a characteristic of the technologically empowered yet politically precarious digital citizen. (Anker, 2014) Ineffective politics pushes users to react, transforming shaming culture in meaningful political participation. (Ingraham and Reeves, 2016) 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. To blame one person as the cause of a more significant issue triggers a connection with scapegoating. The accusers feel relieved that they identified and removed who was causing a problem.

Peter Thiel was the first external investor of Facebook as a result of his interest in the ideas of the philosopher René Girard. 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."

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, whether it lies on the users or the platforms to fight the problematic topic of hate speech.



Tarana
@TaranaBurke

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* portraiting a nude woman in a lake. The nudity of the piece caused controversy. Several complains 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.

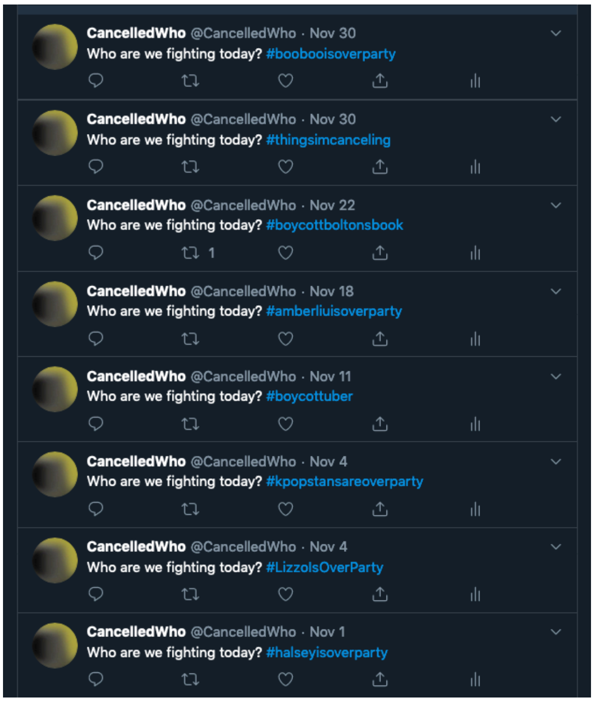
Fig – Matinée de Septembre

On social media, attention is quantified. 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. 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. (Trotier, 2019) The success from scandal shows how cancel culture may fail to hold someone accountable with shaming. R. Kelly eventually went to prison, but many other celebrities enjoyed the status of the victim. It is possible that shaming brings popularity to the offender. At the same time, it explains why cancel culture uses these strategies in the first place.

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. In the mobile version, in 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 precludes other conversations to happen..

I created a bot that uses the Twitter API to look for trends in the United States related to cancel culture. This way, I can have a better understanding of the popularity of boycotting through social media. However, the most mundane reasons affect the data. For example, the administrators shut off my computer every Sunday, so I'm not collecting any evidence of Sunday activism. Also important, the bot listens for specific words I know are correlated with the topic, but I may be missing other specific hashtags which context 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, I understood that the boycotts reaching the trending topics are grounded on social, political or cultural reasons. This supports the idea that users use cancel culture for social justice.



CancelledWho @CancelledWho · Nov 30
Who are we fighting today? #booboisoverparty

CancelledWho @CancelledWho · Nov 30
Who are we fighting today? #thingsimcanceling

CancelledWho @CancelledWho · Nov 22
Who are we fighting today? #boycottboltonsbook
1

CancelledWho @CancelledWho · Nov 18
Who are we fighting today? #amberliuisoverparty

CancelledWho @CancelledWho · Nov 11
Who are we fighting today? #boycottuber

CancelledWho @CancelledWho · Nov 4
Who are we fighting today? #kpopstansareoverparty

CancelledWho @CancelledWho · Nov 4
Who are we fighting today? #LizzoIsOverParty

CancelledWho @CancelledWho · Nov 1
Who are we fighting today? #halseyisoverparty

Fig – My Twitter bot

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. Makes sense that the popularity of sensationalism seen in magazines or the tv works as well in social media.

Although there are groups of people committed to use *cancel culture* as a tool 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. Boycott also discards forgiveness, which 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 bespoken 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 centrally-served platforms.

The idea of pursuing justice collectively, accommodating 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 where the platform is doomed to promote viral actions? It doesn't seem very easy. Cancel culture today is an unforgiving movement, a massive confusion of harassment, shaming, fake morality, and a lot of pointing fingers. Fighting hate with hate had controversial outcomes, but the urgency of controlling it continues very clear.

Chapter 2 — New platforms, different rules

As seen in the previous chapter, digital vigilantism allows users to denounce hateful content within the platform's structure. Another strategy that has been receiving a lot of attention is the development of rigorous codes of conduct. The emergence of significant commitment to creating new community guidelines favours healthy 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) In fact, 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.

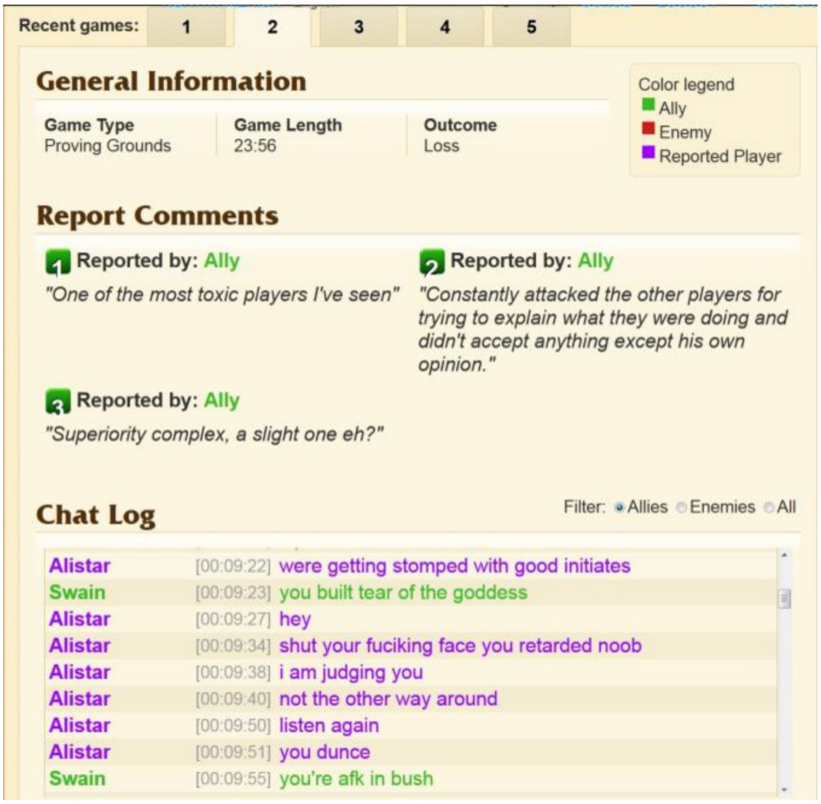
I followed the interesting public thread of discussions in Create mailing list, archived from 2014. This list shares information on free and open-source creative projects. The back-and-forth of emails discusses the need for a Code of Conduct in an upcoming international convention. One of the concerns 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 CoC 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 exist, 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. An effective code of conduct should make a clear distinction between anti-harassment policies and other general guidelines. (Geek Feminism Wiki, 2017) Is harassment publishing a personal address online, or is it a hurtful comment? This distinction is important 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 identities, have their profiles 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. Because it deals so much with user connections, it also has a *Summoner's Code.* The *Summoner's Code* is a code of conduct that formulates what good behaviours for the gamers are. 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 gamers reported repeated breaks of the Code of conduct, for example, because of explicit use of hate language, the case would go to Tribunal. The system attributes the case at random to some users with game statistics, the chat log and the reported comments. The minimum of 20 users would review the situation and then decide to *pardon*, *punish* or *skip.* In the end, the most popular vote for the case was enforced. The type of punishment, whether it was a warning, suspension or even banning, wasn't decided by the users. The judges could see their cases, the outcomes of the decisions, and a personal ranking, which calculated each decision agreed by the majority. Over the first year that the Tribunal system prevailed, 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 vigilante of others. In online forums where people share 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.



General Information

Game Type: Proving Grounds

Game Length: 23:56

Outcome: Loss

Color legend
🟩 Ally
🟥 Enemy
🟪 Reported Player

Report Comments

1 Reported by: Ally
"One of the most toxic players I've seen"

2 Reported by: Ally
"Constantly attacked the other players for trying to explain what they were doing and didn't accept anything except his own opinion."

3 Reported by: Ally
"Superiority complex, a slight one eh?"

Chat Log

Filter: ● Allies ○ Enemies ○ All

```
Alistar   [00:09:22] were getting stomped with good initiates
Swain     [00:09:23] you built tear of the goddess
Alistar   [00:09:27] hey
Alistar   [00:09:34] shut your fuciking face you retarded noob
Alistar   [00:09:38] i am judging you
Alistar   [00:09:40] not the other way around
Alistar   [00:09:50] listen again
Alistar   [00:09:51] you dunce
Swain     [00:09:55] you're afk in bush
```

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 advertising 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 Conduct 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 conduct also toke the time for clearly defining harassment.

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 last question of CounterSocial frequent questions says it all: "Who defines these rules, anyways?" They are.

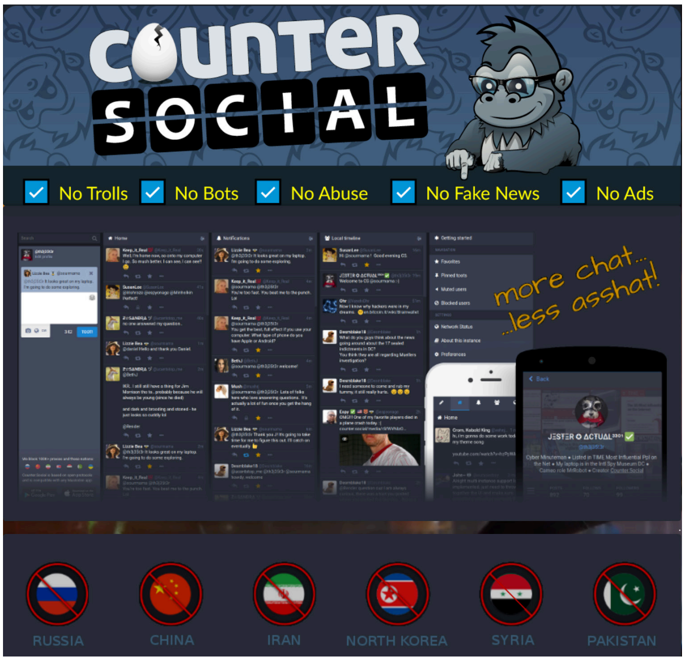


Fig – banned countries

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 deviants inside their own system not involving law enforcement or governmental identities. However, in a very different approach, the repercussions of not following the conduct are almost always dealt in private. The people who manage the community rules 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 subreddit, 2015)

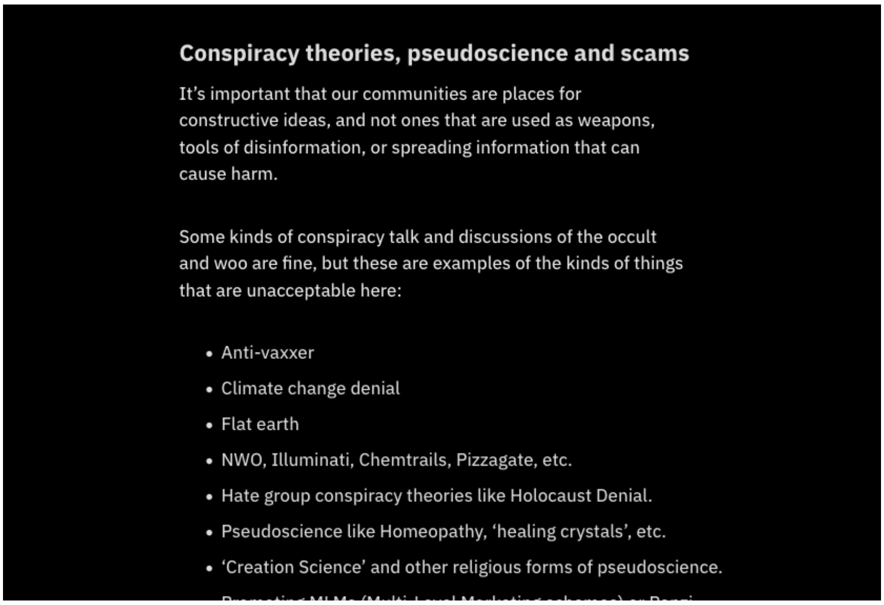
I reached out to one of the moderators of a very populated instance on Mastodon. The most common situation in the other instances is that the moderators are voluntary members of the user base. However, this moderator explained how, in his instance, the Mastodon project pays for three moderators. On top of that, there are some volunteers. The CoC 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 break them. The rest seems to respect 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.

A different online conversation with an administrator of a smaller instance brought to light how the bottom-up initiative of moderating hate is a co-operative task. In this community, there are also three moderators. The admin handles all tech work and daily maintenance, but two other persons bring additional perspective. Similarly to the other group, the moderators don't need to discipline people 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”.

The idea of building safe spaces where users can be active participants and moderators of their social networks is proactive. Mastodon API offers an overview of some groups on the platform. Out of the top more populated servers listed, there is a home for all people, artists, developers, activists, gay men, and a whole building for NSFW content. It’s not only marginalised communities that are enjoying more controlled networks. However, this opens the doors for fascists to make their protected spaces 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, 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 conduct, 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 Ponzi

Fig – policies of Parastat

Codes of conduct make the intentions of a social space known. As explained by Freeman at the beginning of this chapter, groups with no rules don't exist. At best, there are groups with no rules announced. In this 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 within their networks. For this reason, small communities seem more capable to regulate online hate, as it's easier to share similar ideals. On another scale, is it possible to manage billions of different-minded people with one set of rules? 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 be tolerated.

Chapter 3 — Designing change

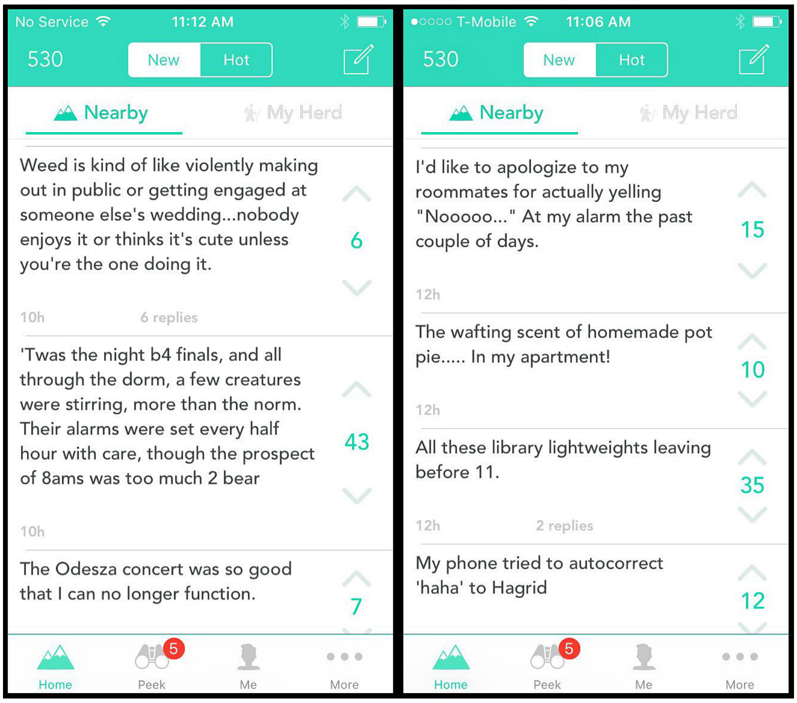
Throughout this text, I touched on the popularity of vigilantism and the increase of stronger user rules. Still, it is compelling to mention the potential of software tools. Users build tools outside the formal development of social media businesses to moderate content on their terms. Together, the community shares notions of morality and customises their platforms, gaining more control over their social media experiences.

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. Designers shouldn't wish for their interfaces to appear invisible, as they are the main translator between the user and the system. The tendency to build attractive, easy to use platforms, can overlook and oversimplify some problems. Especially when dealing with online hate, designers have to increase the attention to the lack of disclosure and choice in what they are creating.

Design becomes dangerous when it facilitates or performs abusive actions. The designer Cade Diehm has been focusing his work on weaponised design. He exemplifies the pertinence of his research with Facebook's celebration of memories. Facebook at the end of a year, reminds the user of their old posts. The suggestion can trigger bad memories, like someone's death, by assuming users only share happy events on the platform. Creators can integrate the weaponised design on purpose, but it's commonly accidental. It happens when designers don't take into consideration all possible outcomes of a design, or it's assumed a perfect user in an ideal situation. Weaponised design is often hard to recognise and understand.

An example of weaponised design would be *Yik Yak*, a social media platform targetted at college students. Before closing, the app allowed users to post messages to a message board, in anonymity. The privacy policy of Yik Yak did not allow 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.

One day at college, the student Jordan Seman saw a horrible message about her and her body on the board. 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. 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)



Weed is kind of like violently making out in public or getting engaged at someone else's wedding...nobody enjoys it or thinks it's cute unless you're the one doing it.
6
10h 6 replies

'Twas the night b4 finals, and all through the dorm, a few creatures were stirring, more than the norm. Their alarms were set every half hour with care, though the prospect of 8ams was too much 2 bear
43
10h

The Odesza concert was so good that I can no longer function.
7

I'd like to apologize to my roommates for actually yelling "Nooooo..." At my alarm the past couple of days.
15
12h

The wafting scent of homemade pot pie..... In my apartment!
10
12h

All these library lightweights leaving before 11.
35
12h 2 replies

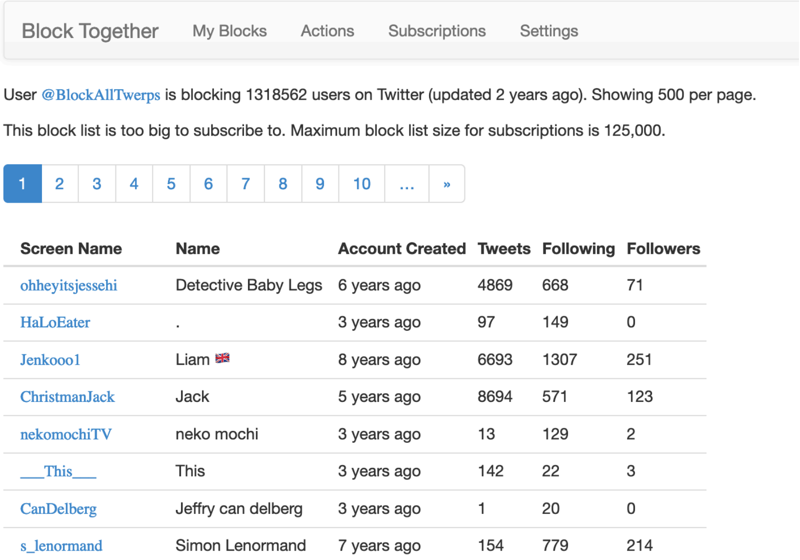
My phone tried to autocorrect 'haha' to Hagrid
12

Fig – Yik Yak

Yik Yak's structure was comparable to Reddit. Yik Yak also maintained message boards, allowed pseudonyms and kept a karma system. Similar 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.

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 his block list on Twitter called *BlockAllTwerps*. *BlockAllTwerps* programmatically collects and blocks users that are harassing, following or retweeting harassment. (Hutchins, 2016) When a user subscribes to a blocklist, 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 is a harasser 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 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, adding the reasons for the block, provide tech support, dealing with complaints.



User @BlockAllTwerps is blocking 1318562 users on Twitter (updated 2 years ago). Showing 500 per page.

This block list is too big to subscribe to. Maximum block list size for subscriptions is 125,000.

Screen Name	Name	Account Created	Tweets	Following	Followers
ohheyitsjessehi	Detective Baby Legs	6 years ago	4869	668	71
HaLoEater	.	3 years ago	97	149	0
Jenkooo1	Liam 🇬🇧	8 years ago	6693	1307	251
ChristmanJack	Jack	5 years ago	8694	571	123
nekomochiTV	neko mochi	3 years ago	13	129	2
___This___	This	3 years ago	142	22	3
CanDelberg	Jeffry can delberg	3 years ago	1	20	0
s_lenormand	Simon Lenormand	7 years ago	154	779	214

Fig – Blocklists

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 more a less quiet, as a person may not even detect that was blocked. However, if they do, it may raise some questions about the reasons why it happened. As the list works with different users managing the people who are blocked, it doesn't follow a scrupulous control. 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. For the other hand, it shames the deviants 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."

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 again, so blocklists continue as parallel activities. However, it's not uncommon that grassroots tools turn into real features on social platforms.

On Twitter, *flagging* also 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. As the outcomes of individual flagging are often undisclosed, it is frequent that a community organises and demands change by using the tool collectively. 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.

Paragraph
— A good example of flagging: take down hate speech, racism.

Paragraph
— An unfortunate example of flagging: mass-report of feminist pages. Which is what happens in community platforms such as Wikipedia.
— Finish with the increased automation of tools.
Without moderators, the task to edit hateful content on Wikipedia articles is the result of the public collaborative discussion between users. The editors get help from tools such as ClueBot NG, ORES or the AbuseFilter extension. These software tools automate the detection and removal of hateful content or perform preventive actions. This 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.

Paragraph
Conclusion of the chapter. Power of coordinated strategies using tools.
— Whether is on the margins, as complements, plug-ins. (blocklists), or
— Manipulating usage of some already integrated features (flags)
— Or if it for asking/debating more features (forums)

References

47 U.S. Code § 230 - Protection for private blocking and screening of offensive material (1986). Available at: https://www.law.cornell.edu/uscode/text/47/230 (Accessed: 9 February 2020).

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).

Crawford, K., Gillespie, T. (2016) *What is a flag for? Social media reporting tools and the vocabulary of complaint.* Available at: https://journals.sagepub.com/doi/abs/10.1177/1461444814543163 (Accessed: 9 February 2020).

Freeman, J. (2013) The Tyranny of Structurelessness. *WSQ: Women's Studies Quarterly*, 41 (3–4): 231–246.

Geek Feminism Wiki (2017) *Community anti-harassment*. Available at: https://geekfeminism.wikia.org/wiki/Community_anti-harassment (Accessed: 9 February 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.

Ingraham, C. and Reeves, J. (2016) New media, new panics. *Critical Studies in Media Communication*, 33 (5): 455–467.

Massanari, A. (2017) #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19 (3): 329–346.

Norman, D. (1990) Why interfaces don't work. *The art of human-computer interface design*.

Pereira, L. (2019) *13 - Fed Up!* Available at: http://ilu.servus.at/category/13%20-%20Fed%20Up!.pdf (Accessed: 9 February 2020)

r/rust - Our Code of Conduct (please read) (n.d.). Available at: https://www.reddit.com/r/rust/comments/2rvrzx/our_code_of_conduct_please_read/ (Accessed: 9 February 2020).

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).

Trottier, D. (2019) Denunciation and doxing: towards a conceptual model of digital vigilantism. *Global Crime*, pp. 1–17.

Twitter CEO: "We suck at dealing with abuse and trolls" (2015). 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).