Centre for Digital Citizens

Northumbria University NEWCASTLE

Co-designing platform governance policies

Tackling malicious flagging and de-platforming with impacted social media users

Table of Contents

Report summary

This report shares findings and policy recommendations from three Tech Policy Design Labs (TPDLs) led by Dr Carolina Are, Innovation Fellow at Northumbria University's Centre for Digital Citizens (CDC), in collaboration with The World Wide Web Foundation and Superbloom.

This report provides social media platforms with user-centred and research-informed recommendations to improve the design and effectiveness of their flagging and appeals tools. At present, research has shown these affordances to be inadequate at tackling online abuse, and to provide malicious actors with opportunities to exploit strict platform governance to de-platform users with whom they disagree. This has disproportionately affected marginalised users like sex workers, LGBTQIA+ and BIPOC users, nude and body-positive content creators, pole dancers, but also journalists and activists[1].

The idea underpinning this report is that content moderation often fails to take the human experience into account to prioritise speed and platform interests, lacking in the necessary empathy for users who are experiencing abuse, censorship, loss of livelihood and network as well as emotional distress. Indeed, being de-platformed from social media often leaves users unable to access work opportunities, information, education as well as their communities and networks – and research has found this has adverse mental health and wellbeing impacts[2]. In the words of a participant:

> "*If a post gets removed, instantly I'm angry about it, I'm upset, I'm shaking I'm stressed, but if my account gets removed I don't know whether I'm going to have another breakdown or whether I'm going to have to be institutionalised just because that control, that power has been taken away from me. And that's my whole career and livelihood.*" - Interview participant 12.

[1] See: Are (2023; forthcoming1; forthcoming2).
[2] See Are & Briggs (2023).

Report summary

The report's main contribution consists in a set of governance policies to improve transparency, fairness, education and design in platform governance. These policies were co-designed with marginalised, de-platformed end-users, who are too often ignored when drafting the rules governing the spaces they depend on for their social and work lives. 45 of these users participated in a series of round-table workshops to tackle the creation of policies to improve platform governance of malicious flagging and de-platforming.

The workshops' design and structure were informed by the experiences of and barriers faced by participants who took part in my previous CDC research.

Through this report, I hope to provide platforms, regulators and civil society with insights from participants' own stories and recommendations to tackle the problems that affect them in order to facilitate change towards a fairer, more equal and user-centred platform governance.

What is de-platforming and who does it affect?

> ***"You can tell that these platforms have been made and are run by cisgender white people, because that's what influences community guidelines. And while TikTok is a Chinese platform it's influenced by the tech world, which is predominantly mostly cisgender and usually white men. I just wish that there was more of a level playing field: I want big social media platforms that are made by queer people."* – Interview participant 9.**

De-platforming is a form of content moderation, or the practice of deleting and/or censoring online content, and a crucial aspect of platform governance without which social media would be unusable[3].

Through content moderation, social media and internet platforms make curatorial decisions over the visibility of what is posted on their spaces, enforcing rules established via 'community guidelines' or 'standards' based upon which a blend of human moderators and algorithms are trained to make decisions[4].

Content moderation – and particularly automated moderation – has so far disproportionately targeted marginalised users, over-focusing on nudity and sexuality instead of on violence particularly after the approval of the 2018 United States law known as FOSTA/SESTA – the Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) and the Stop Enabling Sex Traffickers Act (SESTA) respectively.

[3] Diaz & Hecht-Felella (2021).
[4] E.g., see Gillespie et al. (2020).

What is de-platforming and who does it affect?

FOSTA/SESTA removed the Section 230 exception to the US Telecommunications Act which ruled social media companies were not liable for what was posted on them.

Although Section 230 kept – and at the time of writing still keeps - online services immune from civil liability for the actions of their users, in 2018 FOSTA/SESTA reversed this immunity for to content that may facilitate sex trafficking, de facto making platforms liable for at least a portion of what was posted on them[5]. While fighting online sex trafficking may appear like the best moderation choice, a closer look at the groups campaigning for the approval of FOSTA/SESTA shows the joint bill was pushed into the US Congress by right-wing pressure groups and religious extremists, who used sex trafficking as a cover to push an anti-porn, anti-sex agenda[6].

Therefore, FOSTA/SESTA instead resulted in platforms over-censoring posts by sex workers, LGBTQIA+, plus size and BIPOC users, athletes, lingerie, sexual health brands, sex educators, disabled content creators and activists worldwide to avoid being accused of facilitating trafficking, applying this US legislation to content posted around the world[7].

The law changed platform governance through increasingly sex-averse community guidelines focused on nudity, sexual activity, and solicitation. This conservative approach to censorship has been linked to platforms' wish to protect their own commercial interests through being overzealous in following recent legislation[8] and to platform governance's tendency to replicate offline inequalities by over-sexualising and pathologizing content by queer, plus size, disabled and BIPOC users[9] due to a largely male, heterosexual, white, able-bodied and cisgender workforce.

[5] E.g., see: Are (2021); Blunt & Wolf (2020).
[6] E.g., see: Cooper (2021); Nolan Brown (2022)
[7] See: Blunt & Wolf (2020); Are (2021); Paasonen et al. (2019).
[8] See: Are & Paasonen (2021).
[9] E.g., see: Are (2021); Haimson et al. (2021); Salty (2021); Tiidenberg and van der Nagel (2020)

What is malicious flagging and who does it affect?

"[Platforms] need to understand power - that if somebody's got as many followers as JK Rowling, and she quote tweets or harasses somebody, that is actually worse than if somebody's got 13 followers and does something, right? [...] I don't think they understand how incitement works, that you can get your minion accounts to do it without directly saying, 'Go harass this person.'" **– Interview participant 1**

Flagging is a mechanism for reporting content to social media platforms, such as Instagram and TikTok, allowing users to express their concerns about platform governance and the enforcement of platform policies. However, both platform governance research[10] and data gathered as part of my CDC work have shown flagging can also be weaponised against accounts other users disagree with, affecting particularly nuanced content such as posts depicting bodies, sexuality, sex work, activism and journalism.

Over 130 de-platformed Instagram and TikTok participants who took part in my CDC project found these platforms' moderation to be discriminatory and inadequate[PB1] : e.g., sex workers were de-platformed after malicious flagging by as little as one user, while LGBTQIA+ users were targeted by cross-platform abuse in Telegram groups where members collaborated to flag them to de-platform them[11].

[10] E.g., see Tarleton Gillespie, Susanna Paasonen, Danielle Blunt & Ariel Wolf, Emily Coombes et al., Katriin Tiideberg etc.
[11] Are (forthcoming1:2).

The importance of empathy in content moderation

"It's a constant anxious battle. Instagram is always terrifying, and anxious and nerve-wracking and there's never any positive feelings around Instagram. It is always inadequacy, 'Oh my goodness, I'm not posting enough,' or 'Oh my God, if I don't post everyday then my followers are going to drop,' and of course, 'If I post everyday I'm more likely to get shadowbanned or blacklisted or deleted, and have posts removed.'" - Interview participant 7.

Empathy is a process through which a person centrally imagines the narratives, thoughts, feelings and emotions of another person[12]. These narratives generate specific needs, which become crucial when building and using technologies for people.

Empathy is crucial to Human-Centred Design when it is backed-up by tangible prosocial labour to improve the lives of others[13]. Human-Centred Design[14] allows designers to set aside their own assumptions about the world and users' needs to gain real insights from communities. However, empathy is often co-opted to drive sales rather than in pursuit of justice, and content moderation is no different.

As a result, the growth and hegemony of a handful of internet and social media companies has tricked even researchers into designing solutions to content moderation on Big Tech's terms. As Tarleton Gillespie wrote more eloquently[15]:

"*Solving the problems the industry created on the terms they offer can lead us to overlook the problems we are not being invited to solve, the communities the industry tends to ignore, the solutions that challenge the business models embraced by the industry, and those dilemmas that are in fact not solvable, but are actually meant to be perennially contested. We are kept from thinking about how else moderation might be, or how the very fact of content moderation configures public power. [...].*

But when we adopt a problem-solving approach on terms borrowed from social media companies, we risk accepting as a precept that content moderation exists, and must exist in the way that it does—to accept that social media exists in the way it does."

[12] Goldie (2000).
[13] E.g., see Bennett and Rosner (2019).
[14] E.g., see Giacomin (2014); Norman (2013).
[15] Gillespie, T. (2023, p.2).

The importance of empathy in content moderation

This project is instead influenced by design justice, a community-led movement that centres the needs of those most affected by technologies' harms: sex workers, BIPOC and LGBTQIA+ communities, disabled communities, etc.

Design justice focuses on the ways that design reproduces and/or challenges the matrix of domination (white supremacy, heteropatriarchy, capitalism, and settler colonialism)[16]. Centring the most marginalised users' needs, in this case, means working with communities instead of turning them into passive research subjects[17], meaning that they become co-designers and co-rule-makers.

With these communities, I wanted to reimagine content moderation policies outside of Big Tech, to challenge social media platforms to instil empathy into content moderation.

This approach is informed by my previous interactions with social media platforms, where I learnt that the time, resources and attention allocated to engagement with the stakeholders who are directly affected by technology is awarded sparingly. This way, I hope to create a free resource for both users to feel seen in a governance process that often erases them and, crucially, for platform workers to avoid escaping stakeholder engagement.

[16] Costanza-Chock, (2018).
[17] Floegel & Costello (2021, p.629).

Methodology

The workshops' design, structure and tasks were informed by a set of resources grounded in research, including:

My own personal experiences of malicious reporting and de-platforming gathered through autoethnographies[18];

A qualitative survey amongst 123 participants;

12 interviews with censored content creators.

The last two elements in particular originated from stage one and two of my CDC project, while my autoethnographies straddle my time at the CDC and my previous experiences as an independent researcher and censored pole dance instructor.

The workshops followed the Tech Policy Design Lab structure. TPDLs were created by the World Wide Web Foundation to help shape a safe and empowering web for everyone, as envisioned in the Contract for the Web. The TPDLs create space for policymakers, tech companies and researchers to learn directly from those affected by technology.

My workshops tested the TPDL Playbook, a set of best practices funded by The Bill and Melinda Gates Foundation and created by The Web Foundation.

For these workshops, the structure adopted was that of a Landscape TPDL, led by a sole researcher and lasting from two to three months. This format is suited to a problem that researchers are wishing to explore in more depth through preliminary participant interaction.

[18] E.g. see: Are (2021; 2022)

Workshop planning and timeline

Workshop planning

This research project (Submission reference 46139) was approved by Northumbria University and reviewed through the institution's Ethics Online system to safeguard participants' interests and rights to their data. Participants were sent information sheets and consent forms before any of their information was collected. Each participant was paid £50 via the Policy Support Fund for their time and expertise, provided they stayed for the whole length of the workshop.

This report follows three two-hour virtual workshops, each with 15 participants. Following Chatham House rules and through a World Café framework based on The World Wide Web Foundation's Tech Policy Design Lab (TPDL) Playbook, participants were invited to reimagine solutions to improve flagging and de-platforming policies drawing both from findings from my previous CDC project and from their lived experiences.

Workshops funding and timeline

The workshops were supported by Northumbria University's Policy Support Fund, which awarded me with £3620 to pay participants and facilitate dissemination of this project's findings.

Upon securing the funding in January 2023, planning stages for the workshops began, striking a partnership with The World Wide Web Foundation to test their TDPL playbook in collaboration with their non-profit design agency, Superbloom. This means that facilitation and design consulting happened free of charge thanks to this collaboration.

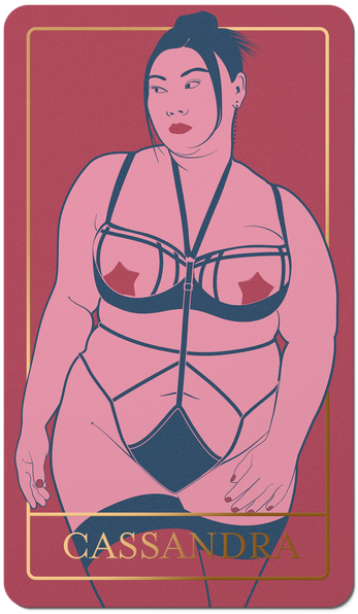
In January and February, Superbloom and I agreed on the workshops' structure. At the same time, I began working with CDC researcher and product designer Dr Henry Collingham to create the resources participants would need to read in advance to take part in the workshops.

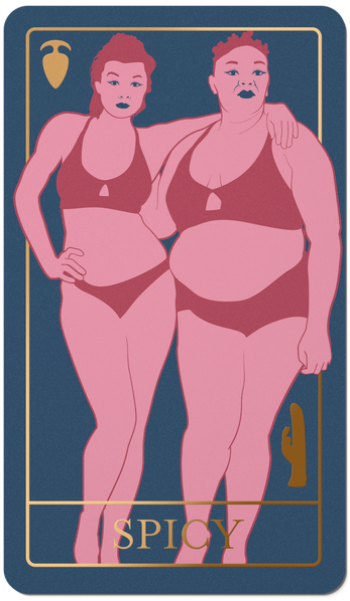
[18] E.g. see: Are (2021; 2022)

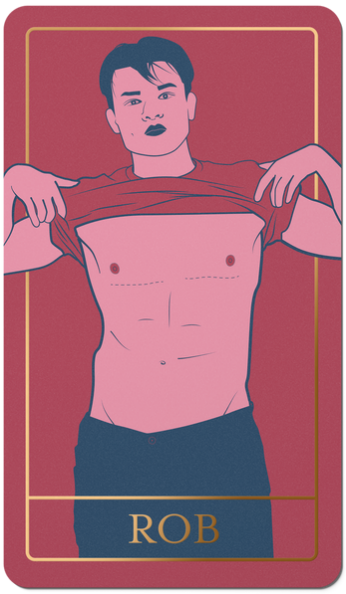
Workshop structure

Dr Collingham and I worked together to design a set of provocations for participants to read in advance of the workshops.

The provocations were presented to participants in the form of illustrations inspired by tarot cards, accompanied by content moderation experiences informed by the data gathered through the aforementioned two stages of my CDC project. The tarots' design was influenced by the stories of those directly targeted and erased from platforms. A purposefully queer, body- and sex-positive set of provocations, the cards helped participants empathise with the users they designed policy recommendations for. Some examples.







Workshop structure

The workshops took place in the following dates, with the following participants.

Workshop date	Workshop focus	Participants
01/03/2023	LGBTQIA+ content	Undergraduate and graduate participants from Berlin's Hertie School of Digital Governance's Pride Network
09/03/2023	Nude and sexual content	Sex workers, activists, artists, educators
19/04/2023	Activism and journalism	Activists and journalists

Workshop structure

Each workshop included an icebreaker, where participants introduced themselves by sharing their experiences of censorship to then answer the following two questions:

1) How might we protect users who think they are being censored and/or de-platformed because of mass reports?

2) How might we help de-platformed users after they experience censorship and/or de-platforming?

The questions were tailored to the specific aim and target audience of each workshop. To answer them, participants were split in four 20-minute long breakout groups of roughly four people each, where they were asked to propose three or more policy solutions to each question with the stories attached to the tarot provocations in mind. Each group appointed a presenter and note-taker to feed these solutions back into the main session.

The workshops were recorded via Zoom and resulted in eight sets of shared documents each: four containing the proposed solutions to flagging (four documents, one for each group) and four containing the proposed solutions to de-platforming (four documents, one for each group). Each document featured the workshop-specific tarot stories participants had to design for, and the 'how might we' questions they had to answer.

Workshop participants

Participants were recruited through a blend of direct outreach, word of mouth and social media recruitment. Those who participated in this study had to read and sign a consent form, providing information about why they wished to join and about their contributions to make sure they engaged with the tasks at hand.

This study featured a total of 45 participants – 15 participants per workshop. Participants were all over 18 years old. Similarly to previously documented experiences of censorship, women were over-represented among participants. Those who participated included both cis and transgender women and men, as well as non-binary people. Participants were both heterosexual and from the LGBTQIA+ community.

Those who took part in the TPDLs were located in and/or from a diverse set of locations: Europe (i.e., the United Kingdom, France, Germany); the United States; South America (i.e., Chile and Brazil); Asia (India, China, Hong Kong); and Africa (Ghana, Kenya, Nigeria, South Africa, Uganda).

While participants had personal experiences and/or knowledge of malicious flagging and de-platforming, the tarot provocations they were introduced to helped them focus on the specific areas of policy that were central to the TPDL they took part in.

Sometimes – and particularly in the case of the TPDL gathering recommendations from nude, sex working and sexual content creators – participants took the provocations further, building on them with their own experiences and generating more data in the forms of censorship and harassment examples. These informed the solutions they created, and they will be part of a rolling provocation project to raise awareness of participants' experiences.

Analysis

The workshops generated rich, detailed and nuanced data in the form of proposed solutions to tackle inequalities in platforms' approach to malicious flagging and de-platforming, as well as further examples of censorship from users' experience. The data was analysed by the lead researcher and the workshop facilitators through a stakeholder mapping and solutions analysis session, which happened remotely via a board on the visual collaboration software Miro.

Proposed solutions were sorted into four main stakeholder groups based on who would be responsible for each proposed solution:

Users

Platforms

Governments

Outside organisations (e.g., non-governmental organisations, volunteers, research centres etc.).

Subsequently, we drew our attention to platforms as the main stakeholder group, given that, too often, platforms, law enforcement and civil society place the burden of tackling issues in platform governance (e.g., online abuse, censorship etc.) on users themselves, replicating offline instances of victim-blaming and lacking support for women and marginalised communities facing hardship and aggression[19].

[19] E.g. see: Are (2021; 2022)

Analysis

Policy solutions to malicious flagging and de-platforming expected from platforms were then ranked using Diamond Ranking[20], a visual method in which data can be sorted according to a given criteria, to decide which solutions would be easier and quicker to adopt for platforms.

Diamond ranking has been used across several cultures and countries e.g., by researchers and teachers interested in students' views of learning environments and practices to inform the redesign of space. Given that we were interested in redesigning platform governance through user-centred and informed policies, this method was appropriate to reimagine the running of these spaces. During coding, quick wins or short-term solutions for platforms to action were ranked at the top of the diamond, while longer-term, more time and resource-intensive solutions were assigned to the bottom.

[20] Newcastle University. (n.d.). https://www.ncl.ac.uk/cored/tools/diamond-ranking/

Analysis

At this juncture, we should therefore emphasise that the longer-term solutions – e.g., better and quicker customer service, investment in human moderation and more personalised notifications of de-platforming – are by no means less important than the quick wins. Quick fixes cannot be seen as the only way to address the more systemic harms of platform governance and its inequalities[21]. They just rely on information and tools that platforms may have already, and that they can harness while working towards long-term change.

Finally, this report also features users' recommendations for support from outside organisations, to supplement the slow pace of change in user-centred content moderation.

[21] Are & Gerrard (2023).

Co-designed policies to improve platform governance

"The issue is that there's no real people that seem to run this side of the platform, while I think the platform should be run more by people than by algorithms created by these tech lizards." – Interview participant 4.

Social media are important spaces for work, promotional and civic life. Malicious flagging and de-platforming can therefore affect the livelihoods, lives, and visibility of a set of users who become powerless in trying to understand or appeal de-platforming decisions and in trying to recover their profiles, content and social media history. It is therefore important to re-design these specific moderation functionalities centring users' safety, work, and freedom of expression.

The following sections feature users' recommendations to improve responses to malicious flagging and de-platforming.

01

Malicious flagging

Policies to improve platforms' responses to malicious flagging

02

De-platforming

Policies to improve issues surrounding de-platforming

03

Help from outside organisations

How can charities, NGOs, think-tanks, unions etc. help improve user experience?

04

Considerations towards systemic change

Ideas beyond single policy changes.

Policies to improve responses to malicious flagging

"Users should have more of a say in their experience. I think that is the most important fix: being able to curate your own content helps fix a lot these issues in general, you don't have to see what you don't want to see online. There should be different communities that have a say in how the moderation works, like different teams e.g. for sex workers who are on Instagram, for artists, for comics, for activists etc. I think that there should be teams of people that can help with their own communities online in those areas they're experts in to have a say, instead of leaving it up to these algorithms." – Interview participant 5

Users recommended that platforms should overhaul their public-facing flagging uptake policies, foregrounding transparency. Proposed solutions in transparency could lead platforms to better understand the challenges and effectiveness of their tools, which could in turn lead to better education towards users' understanding of platforms' policies, and towards data-informed design solutions. Users therefore recommended the following improvements.

Policies to improve responses to malicious flagging

Transparency policies:

1. Publishing regular reports about the efficacy, frequency of use, and uptake of flagging tools, to understand whether specific groups are being targeted with it.

2. Letting flagged users know the number of accounts and/or of reports that affect them, with information about what they were reported for.

3. Detecting and notifying users of reporting by (suspected) bots, AI, and humans.

Safety and equality policies:

1. Allowing users to report themselves as a protected category to expedite reviews and notify companies of malicious flagging against specific groups.

2. Querying flags by and against specific demographics/interest groups (particularly if repeated) to recognise patterns in the flagging tools' use.

3. Detecting if specific accounts continually report others to prevent harassment.

Education policies:

1. Publish more detailed Terms of Use to provide further clarity around criteria for take-down after flags.

2. Provide rights and responsibilities training for users upon sign-up.

3. Notifying users about country-specific contexts in which content is not allowed, similarly to copyright notices.

Design recommendations:

1. Slow reporting: a lengthier process to flag content, asking users to provide written reasons why they are reporting, with the potential to analyse speech in case it is malicious.

2. Granular content controls: periodic requests to confirm if users wish to see specific content to prevent flagging due to disagreement.

3. A 'Close Friends' version of posts, for these to be shown only to the right audience.

Long-term recommendations (malicious flagging)

While greater transparency and education, and their subsequent design recommendations, rely on information that platforms already have, longer-term solutions towards systemic change may require more investment in time, resources, technologies or, sometimes, the widening of certain privileges to more user groups. To this end, users also provided three crucial recommendations for long-term commitment to fairness in flagging:

1. Hiring more human intermediaries to intercept flags, in order to directly contact flagged users and to give them the chance for direct recourse.

2. Similarly to the process already in place for celebrity and high-profile users known as Cross Check (Oversight Board, 2022), there should be a period of hiatus (e.g., 24/48 hours) for moderators to review flagged content ahead of removing it instead of outright banning it through automation.

3. When the above fails, platforms should provide specific forms for recourse to be sent to human moderators, e.g., in cases when users believe they are being targeted by flagging hate campaigns.

"A lot of the people that moderate are gig workers in the Global South, in countries where they're not being paid very much, English isn't their native language, and they're looking at content without any cultural or linguistic context on what this even means, and they are just seeing horrific stuff all day. They're seeing this content and their job is to take it down and it's traumatising for them." – Interview participant 3.

The workshops highlighted users' lack of contact with platforms after de-platforming of content and profiles, resulting not just in community and identity loss, but also in a lack of understanding of the application of platform policies in practice. An over-reliance on automation and a lacking appeals infrastructure also resulted in lack of fairness in content moderation. Users therefore recommended the following improvements to social media de-platforming.

Policies to improve issues surrounding de-platforming

Fairness and due process policies:

1. Give users access to their case history or violations record when their account is taken down to expedite appeals and identify repeat glitches and/or repeated malicious reporting.

2. Enable off-platform tools or specific communication about which posts triggered de-platforming and why to allow users to learn from their mistakes and expedite appeals.

3. Layer de-platforming into a chain of decisions akin to a tribunal process with majority votes amongst moderators before an account is suspended entirely.

Context policies:

1. Create less broad and more granular content categories to avoid grouping all posts under "sensitive content". This way, users could choose (upon sign-up and with periodical checks after) whether they wish to see violence, nudity, sexual activity separately, based on their preference.

2. Ensure moderators have contextual knowledge by hiring content moderation teams and/or volunteers with local/language/personal expertise to avoid false positives and pick up false negatives.

3. Fast-track reviews for important content and accounts – i.e., journalistic information, relief organising, activism information.

Policies to improve issues surrounding de-platforming

Community support policies:

1. Officialise the already existing help from their communities that they harness after de-platforming by adding an 'ask your followers for help' element to appealing.

2. Allow users to select friends or followers to provide additional reports or vouch for mass reported and/or de-platformed users.

3. Include a form of Reddit-style, community service moderators to complement agency and platform moderators for further context.

Sign-posting policies:

1. Add further layers of content signposting to prevent or mitigate de-platforming, e.g., by tagging content instead of removing it (i.e., tagging something as nudity instead of de-platforming it, allowing users to decide if they want to see it).

2. Limiting age-restriction to content instead of to whole profiles by giving users the option to age-restrict certain posts (e.g. to allow teens to see sex education).

3. Allow specific account categories, such as media or sex workers, to go through a kind of verification process – i.e., for journalists to be verified with their news organisation so that companies can act on their behalf upon de-platforming, or for sex workers to express consent on their account being used for sex work.

Post de-platforming rights policies:

1. Allow users to download and access their content post-de-platforming, to hold onto their intellectual property, data and memories.

2. Give the option to pay a small, affordable fee to platforms to expedite profile review, particularly for content creators and non-profit and/or media organisations needing speedy profile reviews, or needing quick human review.

3. Give more options to debate deletions beyond appeals: i.e. through policies allowing users to either make a case for themselves personally or to appoint case representatives to do so for them, potentially by allowing them to juxtapose de-platformed content with existing content on platforms, highlighting that what they have shared is already allowed on other users' social media profiles.

Long-term recommendations (de-platforming)

"I think a lot of people that complain about too much moderation and push for more free speech are often more right-wing free speech activists, they want to be allowed to say anything. I don't really think that's the solution, because I do believe that moderation is important: people need to be safe on the Internet. But I do believe that there's this overall stinginess of Meta, that they really only care about making money and also about having as little blood on their hands as possible, so they over-correct and over-moderate and ineffectively moderate their flawed algorithmic design. And so I think that there needs to be better design in general: there needs to be a larger team of people." – Interview participant 4.

Similarly to policy solutions to flagging, users' policy solutions to de-platforming rely on information that platforms already have. However, systemic change to improve platforms' approach to content moderation requires more investment in time, resources and technologies – and overall, ensuring better due process. To this end, users recommended the following long-term changes to improve platforms' content moderation.

1. Platforms should create more opportunities (i.e., forums or inboxes) for more direct and personalised customer service allowing users to write a response to de-platforming, to ask questions and to request feedback on specific content and profiles.

2. More workers from diverse backgrounds and areas of expertise should be employed in the moderation fields towards better diversity, contextual knowledge and sheer scaling of responses, even with pre-emptive communications ahead of de-platforming.

3. Users agreed platforms should be developing better, more accurate AI to prevent false positives.

Assistance from outside organisations

Aside from platforms-led solutions, users expressed the wish to receive help from outside, third-party organisations when dealing with malicious flagging and de-platforming. Users recommended the following ways in which these organisations, which span from NGOs to independent oversight bodies, could step in to aid users with their platform governance issues.

- Agencies and/or apps could provide third-party, specific support for content creators in a reliable and affordable manner as and when their profiles were deleted.

- Stakeholder advisory councils at companies, such as the Oversight Board, should be extended to a cross-platform system, allowing public comments, and pushing for visible due process.

- Different ombudsman style associations and/or trade unions could step in as mediators and representatives of de-platformed users across different fields to defend cases and work towards community-specific solutions;

- Directories of volunteers with contacts at platforms could be created to help share the load when trying to recover lost accounts.;

- NGOs could offer back-up services (e.g., data back-ups or transfers so that users do not lose posts, and can easily migrate or access content across platforms).;

- Participants and facilitators alike were shocked to realise that there is no formal watchdog program (e.g., from organisations such as the Pew Research Centre or Nieman Lab) to report de-platforming of journalists and to track which platforms are doing it. As a result, they recommended formalising this type of support and investigations.

- ·The mainstream media should engage in continuous awareness raising of content moderation issues – particularly when they affect marginalised communities.

Considerations towards systemic change

The goal of this project was not just to design potential quick fixes to platform policies – we aimed to be unruly, to reimagine a more just overhaul to the system that is platform governance.

Towards real, systemic change therefore, users pushed for radical transparency and, crucially, for a duty of care by platform conglomerates, demanding information, rights, and compensation when platforms fall short of protecting their users. For this, they wished to redesign platform governance to:

- Ensure politicians and regulator take mandatory social media awareness training to be aware of the space they are regulating.
- Push governments to invest in digital literacy for users to know about their social media rights and responsibilities.
- Provide civil society with more information on who trains algorithms for accountability.
- Recognise content creation and promotion as work both on the platform and civil society side.
- Allow users to ask platforms for financial compensation and mental health support upon de-platforming.
- Create pots of legal funding for de-platformed accounts.
- Improve privacy laws to account for surveillance of marginalised users by platforms, which triggers de-platforming.
- Improve human rights law to make sites responsible for the harm of removing content and censoring profiles, moving on from only making them liable for allowing the publication of harmful content.

Conclusion

This report generated policy solutions and long-term recommendations to tackle the adverse impacts platform governance can have on internet users.

Through co-design workshops with 45 affected participants, we proposed a series of solutions to tackle malicious flagging and de-platforming, with a focus on transparency, safety and equality, education, design, fairness, due process, contextual knowledge, community support and sign-positing.

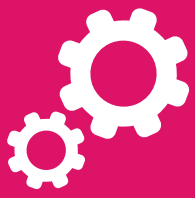
Our report's main highlights include the following.

Communication

Communicating detailed information about content moderation educates and empowers users, improving due process.

Recognising vulnerability

Allowing users to report themselves as protected categories can mitigate current platform governance inequalities.



Granular controls

Making ALL content visible to everyone causes more governance issues than it solves. Empowering users to choose what they see can help mitigate this.

Acknowledgements

This report was supported by Northumbria University at Newcastle Upon Tyne for supporting this project through the internal Policy Support Fund.

This work was also supported by the Engineering and Physical Sciences Research Council (grant number EP/P011454/1).

We would also like to thank the World Wide Web Foundation and Superbloom for supporting this report as part of the Tech Policy Design Lab Playbook project.

More than anything, we are incredibly grateful to our participants who took the time to co-design policies and solutions with us.

We hope this report can improve the experiences of participants affected by online abuse and censorship.

Contact

Dr Carolina Are
Centre for Digital Citizens
Northumbria University
College Ln, Newcastle upon Tyne NE1 8SG

https://digitalcitizens.uk/
carolina.are@northumbria.ac.uk
@bloggeronpole

Research team

Dr Carolina Are, Principal Investigator (she/her)

Dr Carolina Are is an online moderation researcher with a PhD in online abuse and conspiracy theories, currently working as Innovation Fellow at Northumbria University's Centre for Digital Citizens. Following her experiences of online censorship, she has been researching on algorithmic bias against nudity and sexuality on social media, and has published the first study on the shadowbanning of pole dancing in Feminist Media Studies.

Her work has been published in Social Media + Society, Media, Culture & Society, First Monday and Porn Studies. It has appeared in The New York Times, The Atlantic, The Conversation, the BBC, Wired, The Guardian and the MIT Technology Review. She is also a content creator and blogger, as well as an activist and a pole dance instructor, on social media at @bloggeronpole.

Dr Henry Collingham (he/him)

Dr Henry Collingham is a designer and researcher, using participatory methods to foreground the voices of traditionally marginalised populations in sensitive design contexts. Henry is currently working as an Innovation Fellow for Design and Creative Technology at the Centre for Digital Citizens, based in Northumbria University's School of Design. Henry principally works in the context of Health and Wellbeing.

Research team

Ann Marie Carrothers (add bio)

Eriol Fox (they/them)

Eriol Fox works at Superbloom design on research, open-source and technology project and is a part-time funded PhD researcher at Newcastle University's Open Lab, looking at how designers participate in humanitarian and human rights focussed open-source software projects.

They have been working as a designer for 10+ years, in for-profits, NGOs and open-source software organisations. Their work addresses complex problems like sustainable food systems, peace-building and crisis response technology. Eriol is also part of the core teams at Open Source Design (Dr Carolina Are, Principal Investigator (she/her)

Studies from this project underpinning this report

Are, C. (forthcoming 2023) 'Flagging as a silencing tool: exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok.' New Media + Society, in production.

Are, C. & Briggs, P. (2023) The emotional and financial impact of de-platforming on creators at the margins. Social Media + Society, 9(1). https://doi.org/10.1177/20563051231155103

Are, C. (2022) An autoethnography of automated powerlessness: lacking platform affordances in Instagram and TikTok account deletions. Media, Culture & Society, 45(4), 822–840. https://journals.sagepub.com/doi/10.1177/01634437221140531.

Are, C. (Forthcoming1) '"Dysfunctional" appeals and failures of algorithmic justice in Instagram and TikTok content moderation.' Under review.

Are, C. (Forthcoming2) 'Investigating the assemblages of flagging and de-platforming against content creators at the margins.' Under review.

Bibliography

Alacovska, A. and Karreman, D. (2022). 'Tormented Selves: The social imaginary of the tortured artist and the identity work of creative workers.' Organization Studies, pp. 1-25.

Allan, B. A., Autin, K. L., & Wilkins-Yel, K. G. (2021). 'Precarious work in the 21st century: A psychological perspective.' Journal of Vocational Behavior, 126, pp. 103491.

Apouey, B., Roulet, A., Solal, I., & Stabile, M. (2020). 'Gig workers during the COVID-19 crisis in France: financial precarity and mental well-being.' Journal of Urban Health, 97(6), pp. 776-795.

Are, C. (2021a). 'Petition update: sharing community experiences of moderation with big tech.' Blogger On Pole. https://bloggeronpole.com/2021/03/petition-update-moderation/

Are, C. (2021b). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram.' Feminist Media Studies, 22: 2002-2019

Are, C. (2021c). 'Linking to other platforms on IG? You may be at risk of deletion.' Instagram. https://www.instagram.com/p/CRMfgJNIFEJ/.

Are, C. (2020). How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. Feminist Media Studies, 20 (5), pp. 741-744

Are, C. (2020a). A corpo-civic space: A notion To address social media's corporate/civic hybridity. First Monday, 25(6), https://firstmonday.org/ojs/index.php/fm/article/view/10603/9549..

Are, C. (2020b). 'Does Instagram's verification discriminate against nudity?' Blogger On Pole. https://bloggeronpole.com/2020/08/does-instagrams-verification-process-discriminate-against-nudity/

Are, C. (2020c). How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. Feminist Media Studies, 20 (5), pp. 741-744.

Are, C. and Paasonen, S. (2021). 'Sex in the Shadows of Celebrity.' Porn Studies, 8(4), pp. 411-419.

Artefact Group. (n.d.) 'The Tarot Cards Of Tech.' The Tarot Cards Of Tech. http://tarotcardsoftech.artefactgroup.com/.

Banet-Weiser, S. (2018). Empowered: Popular feminism and popular misogyny. Durham, NC: Duke University Press.

Bibliography

Beebe, B. (2022). '"Shut Up and Take My Money!": Revenue Chokepoints, Platform Governance, and Sex Workers' Financial Exclusion.' International Journal of Gender, Sexuality and Law, 2, pp.140-170.

Beech, N., Gilmore, C., Hibbert, P., & Ybema, S. (2016). 'Identity-in-the-work and musicians' struggles: The production of self-questioning identity work.' Work, Employment and Society, 30, pp. 506–522.

Benach, J., Vives, A., Amable, M., Vanroelen, C., Tarafa, G., Muntaner, C. (2014). 'Precarious employment: Understanding an emerging social determinant of health.' Annual Review of Public Health, 35, pp. 229-253.

Bhandarker, A. and Rai, S. (2019) 'Toxic leadership: emotional distress and coping strategy.' International Journal of Organization Theory & Behavior, 22 (1), pp. 65-78.

Bennett, C. L., & Rosner, D. K. (2019). 'The Promise of Empathy.' CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–13. Alacovska, A. and Karreman, D. (2022). 'Tormented Selves: The social imaginary of the tortured artist and the identity work of creative workers.' Organization Studies, pp. 1-25.

Blunt, D. and Wolf, A. (2020). 'Erased: The impact of FOSTA-SESTA and the removal of Backpage on sex workers.' Antitraffickingreview.org, Special Issue – Technology, Anti-Trafficking, and Speculative Futures, 14: 117-121.

Blunt, D. & Stardust, Z. (2021). (2021) 'Automating Whorephobia: Sex, Technology and the Violence of Deplatforming'. Porn Studies, 8 (4), pp.350-366.

Bronstein, C. (2021). 'Deplatforming sexual speech in the age of FOSTA/ SESTA.0 Porn Studies, 8(4), pp. 367-380, DOI: 10.1080/23268743.2021.1993972.

Bucher T. and Helmond A. (2018). 'The affordances of social media platforms.' In: Burgess J, Poell T and Marwick A (eds) The SAGE Handbook of Social Media. London and New York: Sage, pp. 233–253.

Cooper, D. (2021) 'OnlyFans' policy switch is the latest victory in Big Banking's war on sex', Engadget. https://www.engadget.com/onlyfans-big-banks-war-adult-content-174041161.html.

Bibliography

Costanza-Chock, S. (2018).'Design justice: Towards an intersectional feminist framework for design theory and practice.' Design Research Society Conference 2018. University of Limerick. https://doi.org/10.21606/drs.2018.679.

Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. MIT Press.

Cotter, K. (2021). '"Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms.' Information, Communication & Society, DOI: 10.1080/1369118X.2021.1994624.

De Veirman, M., Cauberghe, V. & Hudders, L. (2017). 'Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude.' International Journal of Advertising, 36:5, pp. 798-828, DOI: 10.1080/02650487.2017.1348035.

Diaz, A. and Hecht-Felella, L. (2021). 'Double Standards in Social Media Content Moderation.' Brennan Centre For Justice, pp. 1-39.

Dijkstra-Kersten, S.M., Biesheuvel-Leliefeld, K.E., van der Wouden J.C., Penninx, B.W., van Marwijk H..W. (2015). 'Associations of financial strain and income with depressive and anxiety disorders.' Journal of Epidemiology and Community Health. 69(7), pp:660-665.

Duffy, B. E. (2020). 'Algorithmic precarity in cultural work.' Communication and the Public, 5 issue, 3-4, pp.103-107.

Duffy, B.E. & Meisner, C. (2022). 'Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility.' Media, Culture & Society. https://doi.org/10.1177/2057047320959855.

Easterbrook-Smith, G. (2022). 'OnlyFans as gig-economy work: a nexus of precarity and stigma.' Porn Studies, https://doi.org/10.1080/23268743.2022.2096682.

Eslami, M.; Karahalios, K.; Sandvig, C.; Vaccaro, K.; Rickman, A.; Hamilton, K-; Kirlik, A. (2016). 'First I "like" it, then I hide it: Folk Theories of Social Feeds.' CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 2371–2382.

Bibliography

Fiksenbaum, L.; Marjanovic, Z; Greenglass, E. (2017). 'Financial threat and individuals' willingness to change financial behavior.' Review of Behavioral Finance, 9 (2), pp.128-147.

Floegel, D. and Costello, K.L. (2021) 'Methods for a feminist technoscience of information practice: Design justice and speculative futurities.' Journal of the Association for Information Science and Technology, 73 (4) pp: 625-634.

García-Rapp, F. (2017). 'Popularity markers on YouTube's attention economy: the case of Bubzbeauty.' Celebrity Studies, 8:2, pp. 228-245, DOI: 10.1080/19392397.2016.1242430.

Glatt, Z. (2022). '"We're all told not to put our eggs in one basket": Uncertainty, precarity and cross-platform labor in the online video influencer industry'. International Journal of Communication, Special Issue on Media and Uncertainty, 16, pp.1-19. https://ijoc.org/index.php/ijoc/article/view/15761.

Giacomin, J. 2014. 'What Is Human Centred Design?' The Design Journal, 17 (4), pp: 606–623.

Gillespie, T. (2023) 'The Fact of Content Moderation; Or, Let's Not Solve the Platforms' Problems for Them.' Media and Communication 11(2), pp.1-4. https://doi.org/10.17645/mac.v11i2.6610.

Goanta, C. and Ortolani, P. (2021). 'Unpacking Content Moderation: The Rise of Social Media Platforms as Online Civil Courts.' Forthcoming, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3969360.

Goldie, P. (2000). The Emotions: A Philosophical Exploration. Oxford: Oxford University Press.

Haimson, O.; Delmonaco, D.; Nie, P. and Wegner, A. (2021). 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas.' Proceedings of the ACM on Human-Computer Interaction,5, (CSCW2), No.: 466. pp 1–35, https://doi.org/10.1145/3479610.

Hill, K. (2021). 'Algorithmic Insecurity, Schedule Nonstandardness, and Gig Worker Wellbeing.' Population Association of America Poster Presentation. Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3773164.

Bibliography

Instagram. (n.d.). https://help.instagram.com/477434105621119.

Kaye, D. (2019). Speech Police - The Global Struggle To Govern The Internet. New York: Columbia Global Reports.

Lingel, J., Naaman, M., & Boyd, D. M. (2014). 'City, self, network: transnational migrants and online identity work.' Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1502-1510.

Lustig, C. and Rosner, D. K. (2022) 'From Explainability to Ineffability? ML Tarot and the Possibility of Inspiriting Design'. Proceedings of the Designing Interactive Systems Conference, 123-136.. https://dl.acm.org/doi/10.1145/3532106.3533552..

Madsen, I.E.H., Nyberg, S.T.; Magnusson Hanson, L.L. et al. (2017). 'Job strain as a risk factor for clinical depression: systematic review and meta-analysis with additional individual participant data.' Psychological Medicine, 47 (8), pp. 1342-1356.

Marks, A. (2022). 'How Are Trans Bodies Monitored on Instagram? Meta's Oversight Board Takes Up Its First Gender Identity Case.' Rolling Stone. https://www.rollingstone.com/culture/culture-news/meta-oversight-board-trans-body-gender-identity-case-1388117/

MIT Co-design Studio and Sasha Costanza-Chock. The Oracle for transfeminist technologies. The Oracle for transfeminist technologies. https://www.transfeministech.codingrights.org/.

Newcastle University. (n.d.). 'Diamond Ranking'. CoReD. https://www.ncl.ac.uk/cored/tools/diamond-ranking/

Nolan Brown, E. (2022). 'The New Campaign for a Sex-Free Internet.' Reason. Available at: https://reason.com/2022/04/09/the-new-campaign-for-a-sex-free-internet/.

Norman, D. (2013). The Design of Everyday Things, Revised and Expanded Edition. Cambridge: MIT Press.

Parkman, D. (2022) .'Sex Workers, Performers Invited to Participate in Platform Discrimination Study.' XBIZ. https://www.xbiz.com/news/266649/sex-workers-performers-invited-to-participate-in-platform-discrimination-study.

Bibliography

Pitcan, M., Marwick, A. E., & Boyd, D. (2018). 'Performing a vanilla self: Respectability politics, social class, and the digital world.' Journal of Computer-Mediated Communication, 23(3), pp. 163-179.

Pittman, M. and Abell, A. (2021). 'More Trust in Fewer Followers: Diverging Effects of Popularity Metrics and Green Orientation Social Media Influencers.' Journal of Interactive Marketing, 56, pp. 70-82. https://journals.sagepub.com/doi/full/10.1016/j.intmar.2021.05.002.

Salty. (2021). 'Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram.' Saltyworld.net. https://saltyworld.net/algorithmicbiasreport-2/

Salty. (2019). 'Exclusive: Victoria's Secret Influence on Instagram's Censorship Policies.' Saltyworld.net. https://saltyworld.net/exclusive-victorias-secret-influence-on-instagrams-censorship-policies/.

Savolainen, L. (2022). 'The shadow banning controversy: perceived governance and algorithmic folklore.' Media, Culture & Society, pp. 1–19. https://journals.sagepub.com/doi/10.1177/01634437221077174.

Schoenebeck, S. and Blackwell, L. (2021). 'Reimagining Social Media Governance: Harm, Accountability, and Repair.' Yale Journal of Law and Technology, 23, pp. 113-152.

Scully, M.Swords, L.; Nixon, E. (2020). 'Social comparisons on social media: Online appearance-related activity and body dissatisfaction in adolescent girls.' Irish journal of psychological medicine, pp.1-12.

Sherman, L..; Payton, A.; Hernandez, L.; Greenfield, M.; Dapretto, M. (2016). 'The Power of the Like in Adolescence: Effects of Peer Influence on Neural and Behavioral Responses to Social Media.' Psychological science, 27 (7), pp.1027-1035.

Stardust, Z., van der Nagel, E., Tiidenberg, K., Lee, J., Coombes, E. and Miller-Young, M. (2022). 'Manifesto for Sex-Positive Social Media.' ARC Centre of Excellence for Automated Decision- Making and Society. https://sexpositivesocialmedia.org/ DOI 10.25916/vnkc-zn96.

Bibliography

Stardust, Z., Garcia, G. and Egwatu, C. (2020). 'What can tech learn from sex workers? Sexual Ethics, Tech Design & Decoding Stigma.' Berkman Klein Center Collection. https://medium.com/berkman-klein-center/what-can-tech-learn-from-sex-workers-8e0100f0b4b9.

Stefanello, V. (2021). 'Instagram bans popular pole-dancing blogger and activist, who says now "no one is safe" with its sex policies.'The Daily Dot. https://www.dailydot.com/debug/instagram-ban-pole-dancer-bloggeronpole/.

Stokel-Walker, C. (2021a). 'TikTok censored a pole-dancing PhD who studies how social media silences women.' Input Mag.https://www.inverse.com/input/culture/tiktok-censored-banned-pole-dancer-phd-carolina-are.

Stokel-Walker, C. (2021b). 'OnlyFans is abandoning the sex workers who made the platform a success.' New Statesman. https://www.newstatesman.com/science-tech/2021/08/onlyfans-abandoning-sex-workers-who-made-platform-success.

Sverke, M., Hellgren, J., & N˙aswall, K. (2002). 'No security: A meta-analysis and review of job insecurity and its consequences.' Journal of Occupational Health Psychology, 7(3), pp. 242-264.

Tiideberg, K. and van der Nagel, E. (2020). Sex and Social Media. Melbourne: Emerald Publishing.
TikTok (n.d). 'Adult nudity and sexual activities.' Community Guidelines. https://www.tiktok.com/community-guidelines?lang=en#30.

Uttarapong, J., Bonifacio, R., Jereza, R., & Wohn, D. Y. (2022). 'Social Support in Digital Patronage: OnlyFans Adult Content Creators as an Online Community.' In: CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1-7.

Wood, A.J., Graham, M; Lehdonvirta, V. and Hjorth, I. (2019). 'Good gig, bad gig: autonomy and algorithmic control in the global gig economy.' Work, Employment and Society, 33, 1, pp. 56–75.