Robert Lewington (Twitch)
& The Fair Play Alliance Executive Steering Committee

April 2021

Being ‘Targeted’ about Content Moderation:

Strategies for consistent, scalable and effective response to Disruption & Harm

FAIR PLAY ALLIANCE

Content Moderation:

Best Practices for Targeted Reporting & reactive UGC Management At Scale

March 2021

Abstract

This document provides replicable best practice information on how to moderate User-Generated Content (UGC) in social applications or services (including digital media and video games). Its focus is on reactive moderation, a central component of the growing content moderation toolkit where a service provider responds to reports submitted by users of its service regarding UGC that may violate its Terms of Service. Specifically, the document explores and advocates for a ‘targeted’ approach to the creation of reporting mechanisms. This allows users to closely identify the specific infraction, utilise evidence of the infraction—access to which is facilitated as part of the design of the reporting process—enabling consistent content moderation at scale. Note, however, that we will also make passing-reference to pre, post and proactive (see Appendix A) moderation approaches.

Specifics of how best to tailor these best practices to a particular application or service will differ based on various parameters, including: type of service (social media, video game etc.); type of media (text, image, audio, video etc.); sharing mechanism (feed/gallery, avatar, communication etc.); persistence (ephemeral vs. static/umutable) and others, and therefore this document should be considered a set of high-level instructive principles rather than prescriptive guidelines.

Contents

Key sections in bold

Background

Reducing the prevalence of online disruption and harms[1] is a challenge with multiple facets. ‘Safety-by-design’ choices to product development and the field of Player Dynamics[2] in online gaming are fundamental in lowering or removing the possibility for harm before it ever manifests. Service providers can take critical decisions about the way in which content is surfaced; interaction are structured and exposure or reach is incentivised - all of which, in turn, impacts on matters of online safety   Not least,proactive technologies, such as ML detection models, are playing an important role in identifying and mitigating risks and incidents of online harms. To address the challenge comprehensively, a holistic approach and broad toolkit are increasingly indispensable.

However, it is also true that we cannot rely solely on these predominantly ‘ex ante’ mechanisms or choices. Per *The Verge*’s article, “Why AI Can’t Fix Content Moderation” from 2019[3], “...if you talk to actual industry insiders who will speak candidly and who are actually working directly in this area, they will tell you that there is no time that they can envision taking humans entirely out of this loop.”

The upstream work of safety-by-design is vital in reducing the potential for harms in online spaces, helping to create safer, more diverse spaces in which participants feel able to express themselves, while also contributing greatly to the ability of service providers to function positively and at scale. However, it would be naive to think we can ‘design out’ any and all possibility for harm to exist—at least, not without designing out any potential for fun, discourse, or debate—since, as the same Verge article goes on to contend, “...people are always going to try to defeat the algorithm. They’re going to try to defeat it. They’re going to try to game it. We can’t possibly imagine all the scenarios that will come online.” Similarly, some harm comes as a consequence of civic norms and laws, as well as the natural ambiguity that can arise at the intersection of culture and belief, meaning that we will never entirely ‘design out’ the need for content moderation.

Nor, it could be argued, should we necessarily want to. In the worlds of content moderation and Trust & Safety, professionals will often think in terms of the offline world to drive analogies for the application of rulesets and consequent enforcements—criminal justice, law enforcement, civil service etc. These mental models are instructive in framing how we think about both the problems we’re trying to solve and our intended solutions to them. At what point does the need for privacy intersect with or contradict the need for safety? Is a proactive technology intended to promote greater safety through the detection of online harms, such as offline surveillance cameras (CCTV), infringing on humanity’s desire for (and right to) privacy? In short: if we could detect or identify all crime by putting CCTV in every room in every building, would we still do it?

Since we live in an imperfect world, we can use this analogy to determine that, offline, there is no desirable example where humans have either created the conditions where laws are never broken nor where technological resources can always identify the breaking of laws prior to any harms taking place. And even if we borrow from Hollywood to imagine worlds where these systems of law enforcement are

[1] Disruption and Harms in Online Gaming Framework, FPA & ADL
[2] Player Dynamic 101, Riot Games
[3] Why AI Can’t Fix Content Moderation, The Verge

‘perfect’ (at least in terms of doing what they were designed to do), most of us would probably not want to live in such worlds:

i. In 1993’s *Demolition Man*, the future city of San Angeles has indeed created conditions in which virtually no crime exists; the one initial exception is automated graffiti, somewhat analogous to the type of ‘spam bot’ we might find online: popping up autonomously; designed primarily to create annoyance, chaos and disruption—and even this rule break is detected by proactive technology and removed before it can ‘harm’ any citizens. Not only is both speech and thought censored with the proactive monitoring of 1:1 conversations and the enforcement of even private ‘*sotto voce violations of the verbal morality statute*’, but it has the effect of pushing the subversive element of society underground (in this case literally), driving not only resentment but an entrenched feeling of disenfranchisement
ii. 2002’s *Minority Report* depicts a future in which all crime can be predicted and therefore prevented. Despite the sophistication of the technology in use, the system can still be ‘gamed’, as the protagonist is accused of a crime and where the fact he has not yet committed said crime is no longer irrelevant. His fate is entirely dependent on the algorithm proclaiming him guilty.

The purpose of these examples is to illustrate a point: that the purpose of safety-by-design is to minimise the potential for unwanted disruption and to reduce harms by being thoughtful and intentional about how social features and communities are constructed. It is not, however, a panacea or a 100% guarantee that no harm can exist in a product or service and successful safety-by-design will generally carry with it a requirement that safety does not come at the expense of freedoms, expression and enjoyment. Therefore, while safety-by-design and proactive detection are vital tools, they are a supplement to, not a replacement for, **optimised reactive content moderation**. Indeed, optimised safety-by-design *includes* the optimisation of content reporting and moderation mechanisms as part of that design, in the knowledge that these will always be required. As good as we get with preventative measures—and we should aim to be as good as we can while being mindful of the inherent tradeoffs—we will always need to ensure that users of online services can report harms they may experience and that these are properly actioned.

So what does **optimised content moderation** look like? Let’s again think about real-world examples of when a person has experienced harm, such as filing a police report, or is looking to improve the safety of a community, such as a petition for a crosswalk on a busy road. This paper takes the position that those with a genuine desire to combat harm or affect change will fill-out the relevant forms required, so long as the act of doing so aids in the application of meaningful action. While it may be counterproductive if forms are overly bureaucratic, designing them to facilitate the inclusion of supporting information and variable evidence is desirable for both the submitter—as it allows them to more fully state and support their grievance—and the processor, who can more easily identify and rectify the problem.

We therefore contend that two factors are vital when translating these examples into content moderation mechanisms:

1. That a user must always be able to report the harm that took place

2. That the mechanism used to report must be suitable in that it provides for action to be reliably and consistently taken

The juncture of these two factors is what we will refer to as '**targeted reporting**', that is, the system of UGC reporting that builds evidence capture into the mechanisms, facilitating scalability and actionability in content moderation.

The Content Moderation Flywheel

The following sections are a non-exhaustive list of some of the key areas on which to base a content moderation operational model, facilitating 'scalability-by-design' into UGC management and reporting mechanisms.

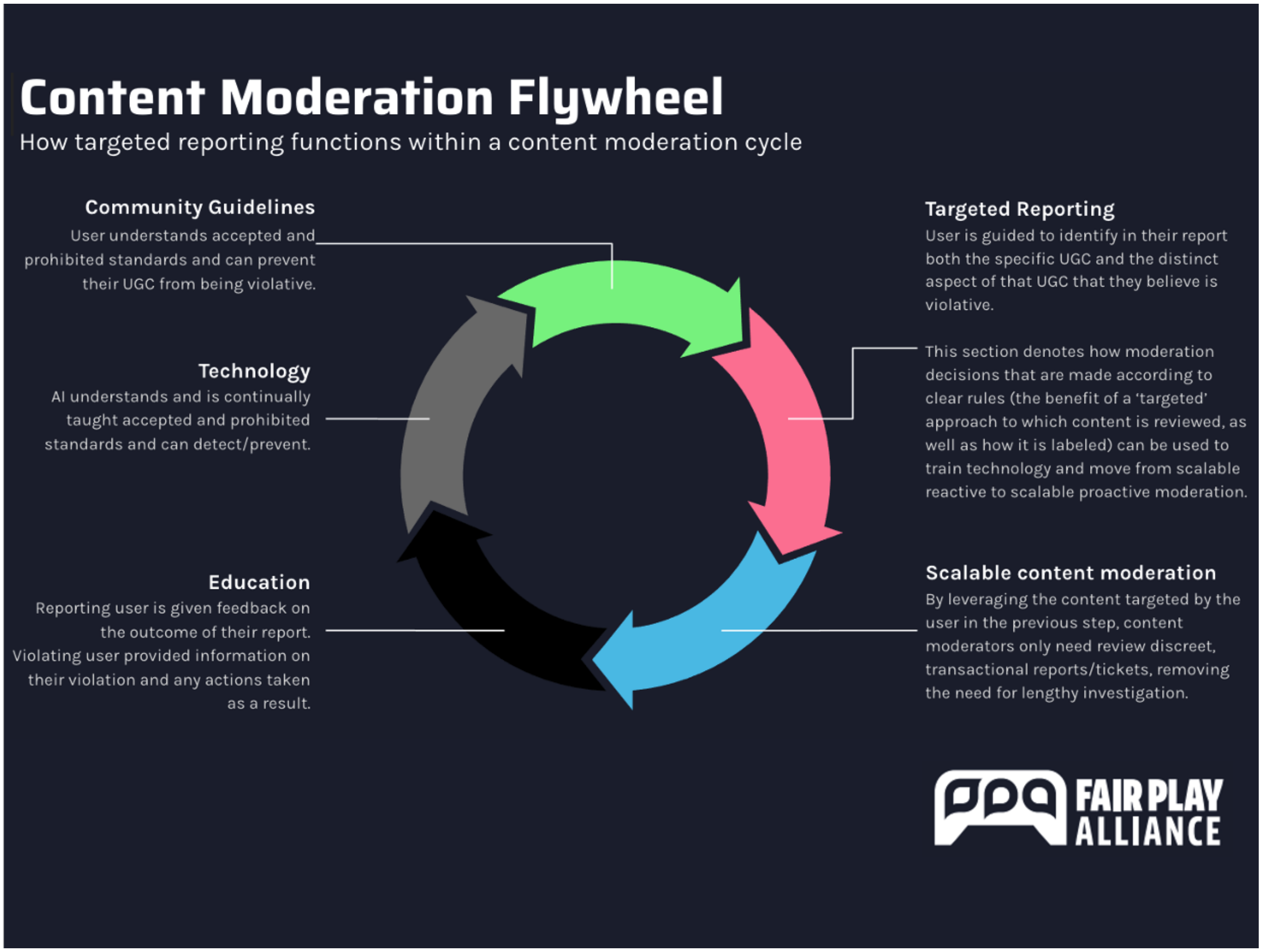
Note that there will likely be easier or quicker wins in the design of content moderation practices which are especially appealing while a service, and hence its associated content moderation needs, are relatively small. However, designing content moderation processes with scale in mind from the offset will support efficient growth, avoid major architectural changes as volumes increase, and therefore help to enable future success. Since greater use of an application or service is fundamentally good for its ongoing efficacy—and yet greater usage drives traffic drives moderation workload, which comes at a financial cost—scalability in content moderation helps to:

A. **Prevent degradation in quality of moderation with volume**: if reporting or moderation mechanisms are not designed to scale with your service, they will inevitably suffer as your service grows.
B. Reduce 'bottlenecks' in growth caused by an overwhelmed moderation team: the above degradation in service likely to negatively impact said growth as the UGC element can no longer be adequately managed to comply with Terms of Service. This both reduces the quality of the service offering—since content that violates the service's guidelines cannot be efficiently or quickly removed—and can contribute to mental wellness issues for moderators, with the potential for lawsuits in extreme cases[4] [5]. The FPA will produce future resources specific to the topic of moderator wellbeing.
C. Manage risk vectors (e.g. reputational, brand, legal etc.) associated with negative content: this degradation will not only impact users but will also reflect badly on the service and its reputation, with potential to increase legal risk and/or irrevocable brand damage.

The below flywheel denotes the key elements in designing a reporting and content moderation system that *will* scale. This document will focus primarily on the '**Targeted Reporting**' and '**Scalable Content Moderation**' components of the flywheel and so these sections of the flywheel will be emphasised.

[4] Facebook will pay $52 million in settlement with moderators who developed PTSD on the job, The Verge
[5] Former YouTube content moderator describes horrors of the job in new lawsuit, CNBC

However, in order to provide context for how this contributes to a virtuous cycle of improvement, the other components of the flywheel are also briefly outlined in sequence.



Content Moderation Flywheel

How targeted reporting functions within a content moderation cycle

Community Guidelines
User understands accepted and prohibited standards and can prevent their UGC from being violative.

Technology
AI understands and is continually taught accepted and prohibited standards and can detect/prevent.

Education
Reporting user is given feedback on the outcome of their report. Violating user provided information on their violation and any actions taken as a result.

Targeted Reporting
User is guided to identify in their report both the specific UGC and the distinct aspect of that UGC that they believe is violative.

This section denotes how moderation decisions that are made according to clear rules (the benefit of a 'targeted' approach to which content is reviewed, as well as how it is labeled) can be used to train technology and move from scalable reactive to scalable proactive moderation.

Scalable content moderation
By leveraging the content targeted by the user in the previous step, content moderators only need review discreet, transactional reports/tickets, removing the need for lengthy investigation.

Fig 1 *- the '<u>Content Moderation flywheel</u>' provides context for how targ*
reporting functions within a content moderation cycle. The pink se
denotes how moderation decisions that are made according to clear
(the benefit of a 'targeted' approach to which content is reviewed, as w
how it is labeled) can be used to train technology and move from sca
reactive to scalable proactive moderation. Note that the necessity and
for investment in such technologies will relate heavily to the scale o
UGC service in ques

While all applications or services will have some kind of Terms of Service or User Agreement, these will often not be the most easily-read or thoroughly consumed documents. Therefore the first component in a highly-scalable content moderation process is communicating acceptable usage of your UGC features in a way that is low effort and easy for users to understand; the intention is that the better users understand the "Do's and Don'ts" of content creation on a social applications or service, the less likely they are to create content that is in violation of those standards.

As outlined in the 'Creating and Maintaining Community Guidelines for Online Games and Platforms' from Fair Play Alliance's Disruption and Harms in Online Gaming Framework, "Going through this activity [to understand what values you want to see in your game/platform and what is important to your organization and players] can help ensure that the systems you build to support these values are in lockstep and create a mechanism through which you can build alignment and accountability"[6].

While this concept may seem intuitive—and even obvious —a study by the National Academy of Sciences[7] has shown that the application of this component is pivotal to success in this field; the research showed that highlighting a community's 'rules' in a persistent and highly-visible way was associated with a higher likelihood of users obeying said rules. Therefore how you display these guidelines within your application or service can have tangible impact on reduction of violative behavior and hence scalability; again, the Fair Play Alliance provides more practical advice on how to do this in the Community Guidelines resource.

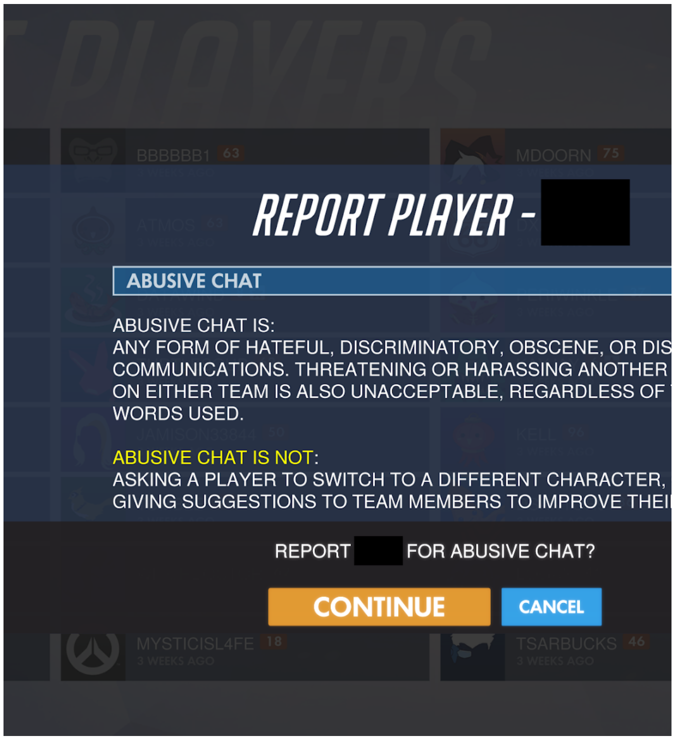
Targeted reporting

In a model that utilizes reactive moderation, the quality of the reports you receive from users is vitally important to scalability. Going back to the offline mental model discussed in the introduction, we can think of this as akin to the offline world of police work, where 'chasing down leads' is often only as effective as the quality of the leads provided; the advantage we have in the online world is that we can tailor the method of submitting a 'police report' such that the evidence required to judge the offence is built in, removing the 'chase' entirely.

Continuing from the previous section on Community Guidelines/Code of Conduct, reporting mechanisms should first play in reinforcing and communicating the standards you have already set. The categories available for the reporting user to select should reflect the things that you have set out as inappropriate as in your guidelines and be as specific as is needed to clearly communicate the transgression without creating the paradox of choice for users.

[6] Creating and Maintaining Community Guidelines for Online Games and Platforms, Disruption and Harms in Online Gaming Framework
[7] Proceedings of the National Academy of Sciences, Nathan Matias, Ph.D.

Example 1: Blizzard Entertainment's *Overwatch* features a reporting mechanism that utilises the report categories to clearly signpost which one the user should select for a given infraction:



REPORT PLAYER - [redacted]

ABUSIVE CHAT

ABUSIVE CHAT IS:
ANY FORM OF HATEFUL, DISCRIMINATORY, OBSCENE, OR DIS
COMMUNICATIONS. THREATENING OR HARASSING ANOTHER
ON EITHER TEAM IS ALSO UNACCEPTABLE, REGARDLESS OF
WORDS USED.

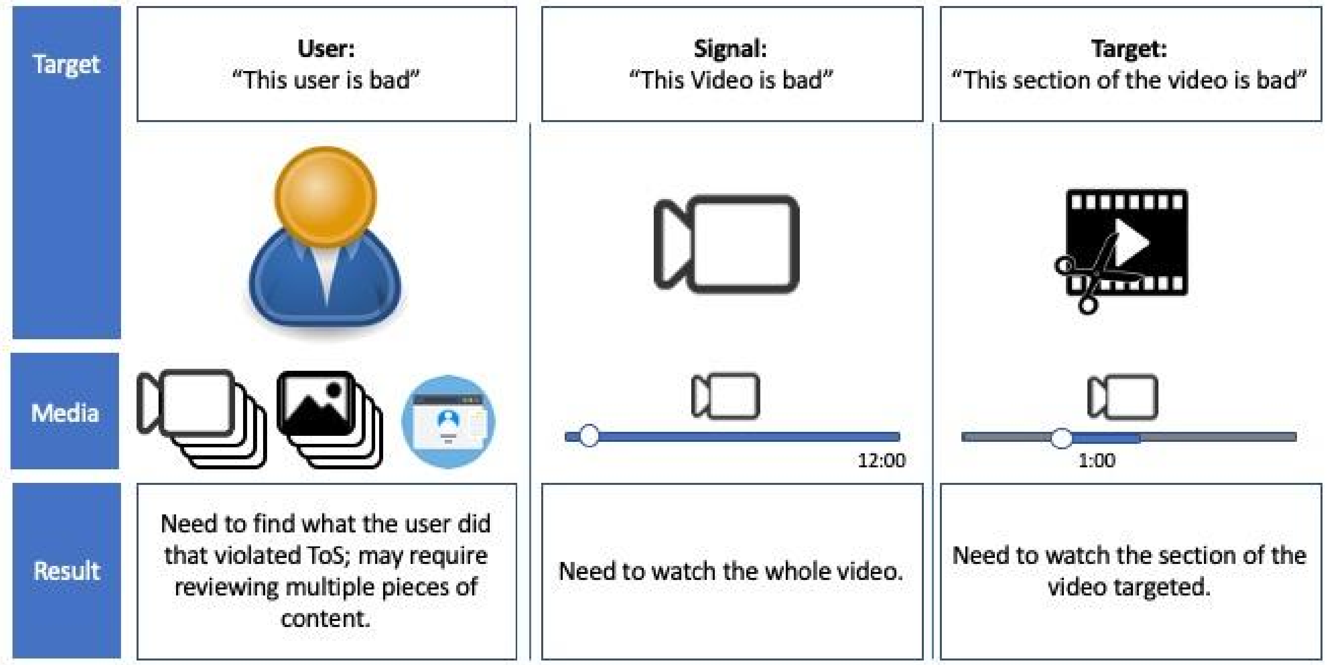
ABUSIVE CHAT IS NOT:
ASKING A PLAYER TO SWITCH TO A DIFFERENT CHARACTER,
GIVING SUGGESTIONS TO TEAM MEMBERS TO IMPROVE THEI

REPORT [redacted] FOR ABUSIVE CHAT?

Overwatch's reporting options are particularly instructive in that it not only describes what constitutes each violation, but also what does not. This helps reinforce the educational elements set out in Blizzard Entertainment's Code of Conduct and sets expectations with the reporting user on what the outcome of a given report is likely to be.

Circling back to the offline analogy of police work, this helps to drive a higher quality of 'leads', increasing the proportion received that will be genuine violations of the rules.

The next aspect of the 'targeted reporting' approach is detailed in Fig 2, where three potential options for a reporting mechanics are shown, each narrowing down the specifics of the complaint to an increasingly focused 'target':



Target	User: "This user is bad"	Signal: "This Vīdeo is bad"	Target: "This section of the video is bad"
Media		12:00	1:00
Result	Need to find what the user did that violated ToS; may require reviewing multiple pieces of content.	Need to watch the whole video.	Need to watch the section of the video targeted.

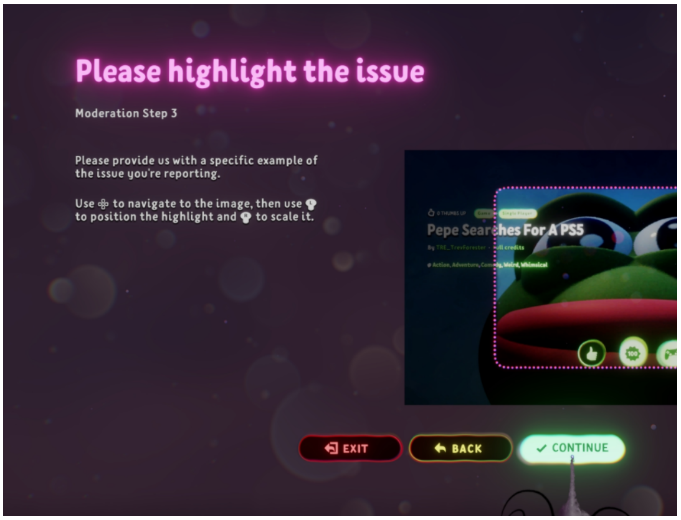
Fig 2 *- note: this is a high-level illustration of the premise of targeted reporting; in practice there are multiple wrinkles to be considered such as context (see below)*

For the purposes of illustration, we have assumed an average handle time of 30 mins per 'user' report (that is, a report where a user, rather than a specific piece of that user's content, is reported). While clearly this will vary greatly depending on the nature of the report and/or social application or service in question, we can take this as a broadly conservative illustrative estimate if it is assumed that the user may have created multiple pieces of content, including videos of 12 minutes in length (as denoted by the 'media' and 'signal report' sections).

The examples in Fig 2 are also intentionally simplistic and this type of content moderation can take many and varied forms. Some real-world applications of the targeted reporting principle:

Example 2: Media Molecule's *Dreams* video game includes an in-game reporting mechanism[8] for flagging user-generated creations or 'Dreams'. Each creation takes the form of a fully-fledged interactive experience comprising a vast amount of individually created objects and/or media; these creations are essentially games within the main game. Therefore investigating 'Dreams' flagged as inappropriate could be an extremely labor intensive process, but Media Molecule have utilized a targeted reporting approach in order to mitigate this:

[8] Dreams, Moderation Guidelines



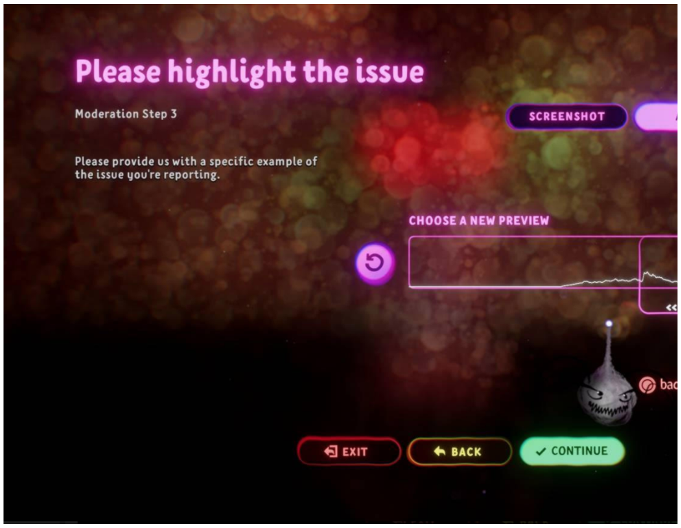
Please highlight the issue

Moderation Step 3

Please provide us with a specific example of the issue you're reporting.

Use ✥ to navigate to the image, then use L to position the highlight and R to scale it.

When the user initiates the reporting process, a screenshot is generated capturing what the player was looking at that point; the player can then crop or navigate the image to focus on the specifically offending UGC or, indeed, cancel the reporting process entirely to ensure the screenshot correctly captures the issue in question.

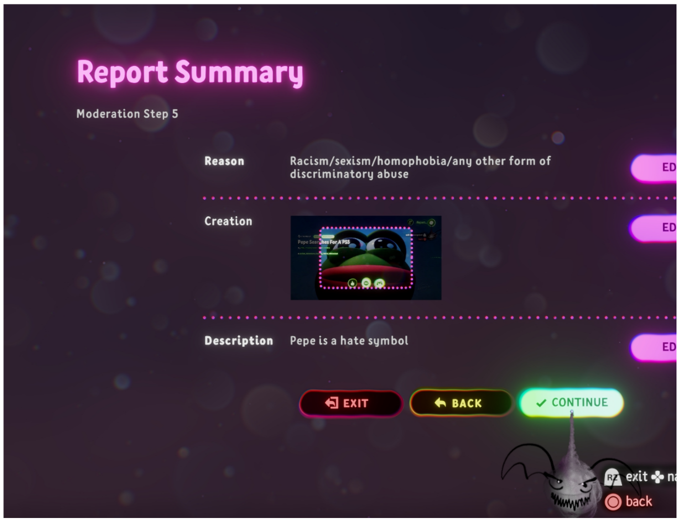


Please highlight the issue

Moderation Step 3

Please provide us with a specific example of the issue you're reporting.

Similarly, if the subject of the report is audio content, the user can toggle from a screenshot to the audio clip they were just listening to; they can then select the section of the clip that they wish to report.



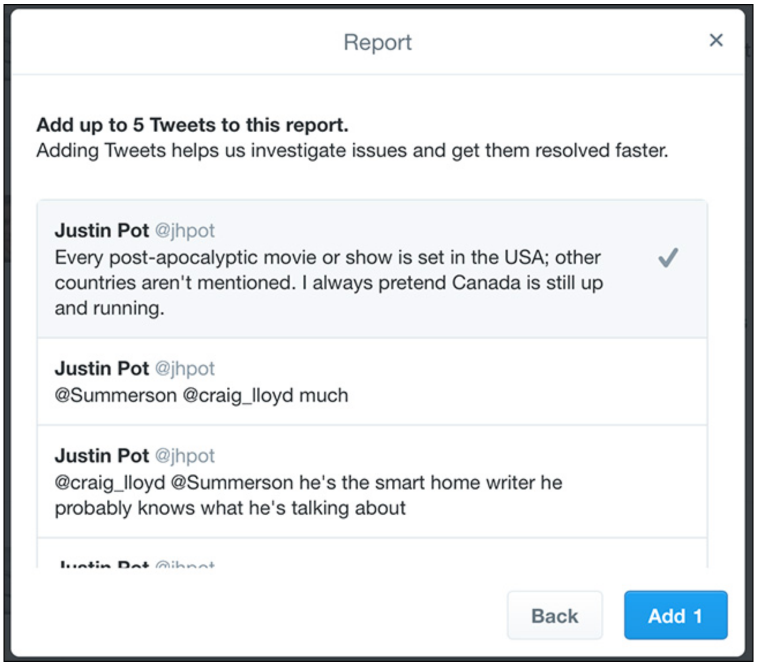
Report Summary

Moderation Step 5

Reason	Racism/sexism/homophobia/any other form of discriminatory abuse	ED
Creation		ED
Description	Pepe is a hate symbol	ED

In both cases a summary of the content (screenshot or audio clip) the user is about to report is shown to them to confirm prior to submission. While it’s possible there are also additional investigative tools available for moderators to deep-dive edge cases, this type of mechanism empowers the user to make an actionable report and still facilitates a discrete, scalable workflow for resolution.

Example 3: Twitter's reporting mechanism[9] which allows the reporter to specific up to 5 additional tweets from the reported user[10].



Add up to 5 Tweets to this report.
Adding Tweets helps us investigate issues and get them resolved faster.

Justin Pot @jhpot
Every post-apocalyptic movie or show is set in the USA; other countries aren't mentioned. I always pretend Canada is still up and running. ✓

Justin Pot @jhpot
@Summerson @craig_lloyd much

Justin Pot @jhpot
@craig_lloyd @Summerson he's the smart home writer he probably knows what he's talking about

Justin Pot @jhpot

This will facilitate an operational workflow/design of tooling where reviewers can review a finite number of tweets—rather than needing to review the user's general activity—aiding in multiple aspects of AHT (average handle time) measurement, planning and scalability.

As we will also see in the Education section, this 'targeting' of specific tweets also facilitates the inclusion of the specific UGC that led to enforcement within suspension notifications.

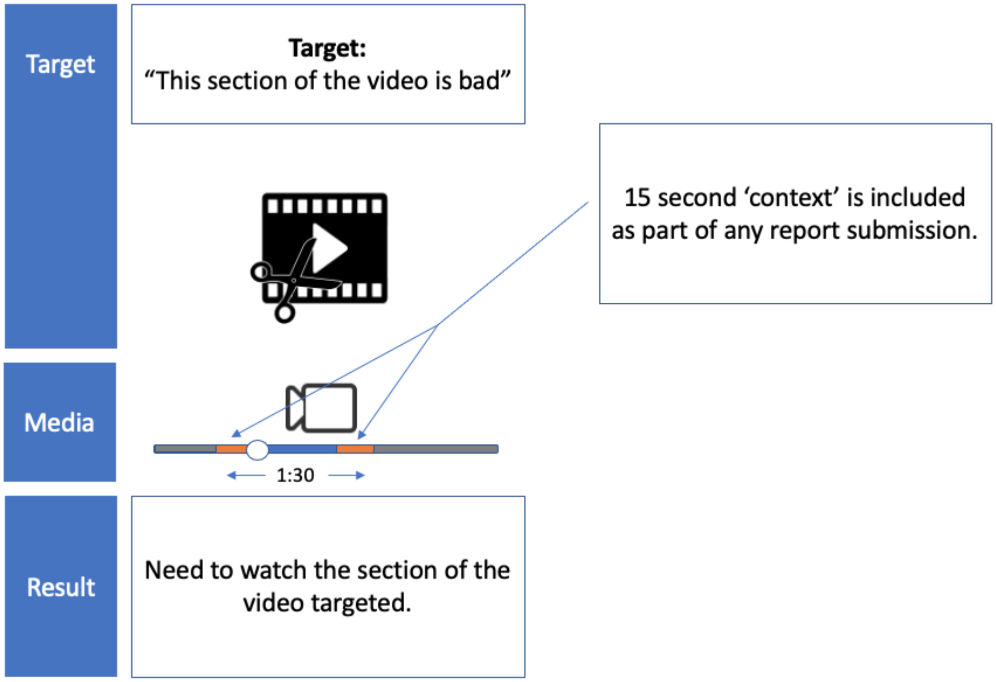
Context

The counterpoint to targeted moderation is context; we can think of this like narrowing down our field of vision to a specific object. This has benefits for clarity, focus and the removal of 'visual noise', but can be restrictive in terms of removing the context in which that object exists. An umbrella can mean two different things, for example, depending on whether the surrounding climate is hot (for shade) or raining (for shelter).

Similarly, if reporting mechanisms are too targeted, content which is offensive or inappropriate when viewed in isolation may be more understandable when surrounded by proper context (and potentially no longer violative, depending on the service or social applications' terms of use). Perhaps the user who has been reported was goaded or provoked (or indeed was the initial victim of violative content themselves); perhaps the reported content exists within a sub-community or conversation in which its tone is in-keeping with 'banter' or 'trash talk'.

Accounting for context while maintaining the advantages of a targeted moderation approach can be managed as illustrated in Fig 3:

[9] Report Abusive Behavior on Twitter
[10] How to Report a Tweet on Twitter, How-to-Geek



Target:
"This section of the video is bad"

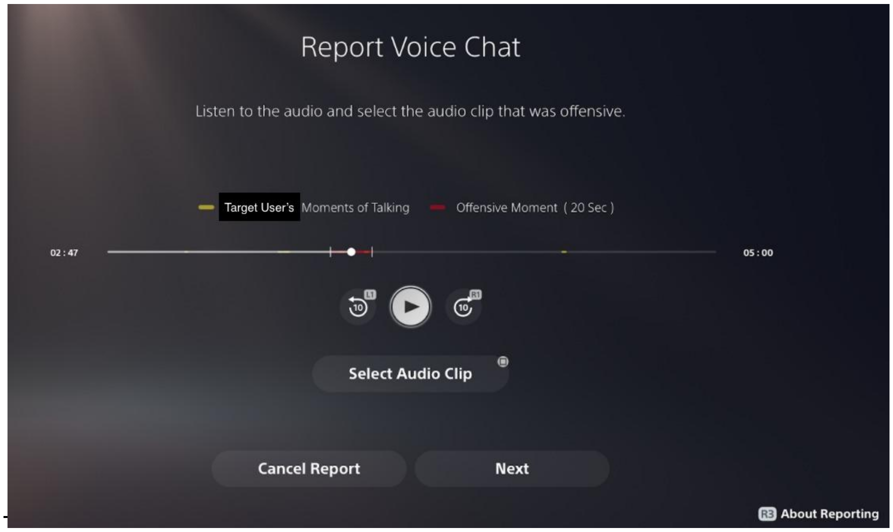
15 second 'context' is included as part of any report submission.

Need to watch the section of the video targeted.

Fig 3 *- The requirement for [context](#) is designed into the reporting mechanic by always including a small amount of surrounding content*

While the content that requires review has now grown by 50% versus the narrower 'target' approach—therefore reducing some of the operational gain described in Fig 2—it is still 8x shorter than the 'signal' approach (also shown in Fig 2). This therefore incorporates many of the gains from 'targeting' while also benefiting from additional contextual information driving a higher likelihood of accuracy in the content moderation decision.

Again, the example shown in Fig 3 is both broad and intended purely as an illustration of the concept, but a very similar real-world application of the concept can be found in the PlayStation 5's voice communications moderation system[11]:



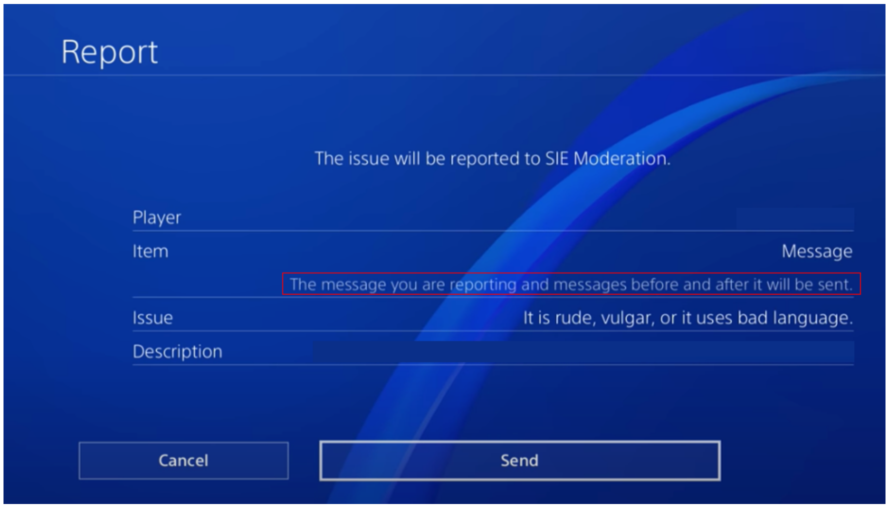
Report Voice Chat

Listen to the audio and select the audio clip that was offensive.

Per PlayStation's official blog:

"...if a PS5 player needs to file a harassment report, they will be able to include up to a 40 second-long Voice Chat clip in their report — 20 seconds of the main conversation with the other player, plus an additional 10 seconds before and after the conversation selection."

[11] https://blog.playstation.com/2020/10/16/details-on-new-voice-chat-functionality-coming-to-ps5/


Report

The issue will be reported to SIE Moderation.

Player

Item: Message

The message you are reporting and messages before and after it will be sent.

Issue: It is rude, vulgar, or it uses bad language.

Description

A similar system existed on PS4 for written messages. As circled in red, the report confirmation screen informs the reporting user that:

“The message you are reporting and messages before and after it will be sent [to SIE Moderation].”

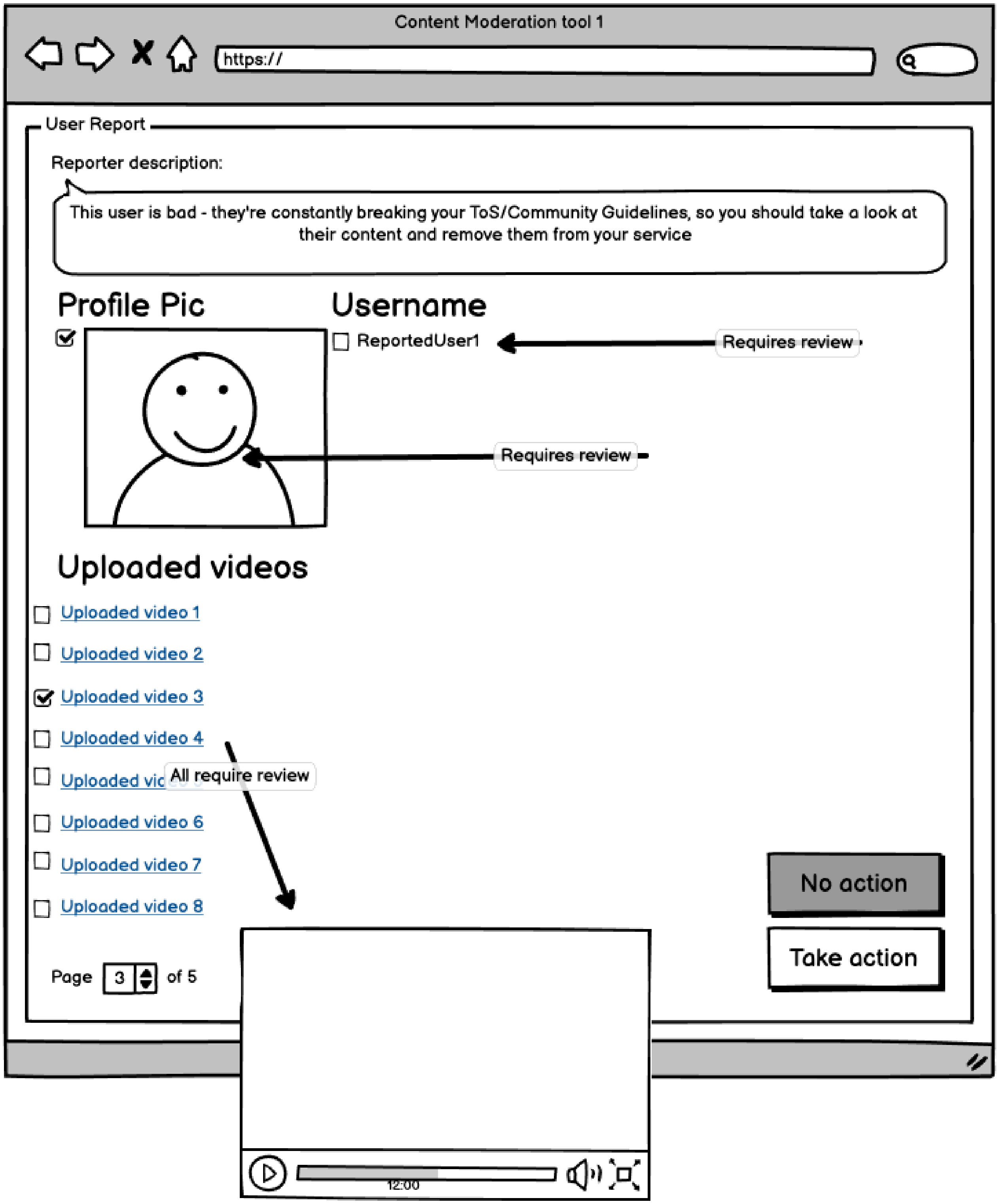
Scalable Content Moderation

Using the examples of types of UGC reporting described in Fig 2 and Fig 3, we can see how targeted reporting can translate to scaled content moderation by reducing the length of time taken on each assessment. In essence, we have started to **design our content moderation tool when designing the report mechanism**, as we are constructing the way in which we ‘package’ the content to be moderated into an easily transactable form.

This means the content moderation tool can simply ‘queue’ the specific pieces of content identified by the reporting users (plus any relevant context) and present it to the moderator in a single tool and within a single view, without the need to click into different pages. It also means that the moderator can perform consistent and replicable processes on the content they are reviewing, as opposed to having to perform a deep-dive investigation for every single user, which will naturally vary from case-to-case. It is often best practice to include mechanisms of escalation so that this type of investigation can still be performed if initial review of ‘targeted’ content indicates but does not fully capture potential violations; however, these will almost certainly be the exceptions and apply to only a very small subsection of reports, presuming that the reporting mechanism has been properly designed (in addition to other principles of safety-by-design and Player Dynamics) . Designers of reporting UIs can then continue to iterate on those mechanisms to ensure the reporting user is able to capture the relevant information in their report submission, in turn reducing the ratio of cases that require investigation.

Fig 4 shows an example of how a content moderation tool might look if the reporting mechanism utilizes ‘User Reporting’ (as described in Fig 2). This requires a significant amount of content to be reviewed for each individual report.

User Reporting



User Report

Reporter description:

This user is bad - they're constantly breaking your ToS/Community Guidelines, so you should take a look at their content and remove them from your service

Profile Pic

☑

Username

□ ReportedUser1

Uploaded videos

□ Uploaded video 1

□ Uploaded video 2

☑ Uploaded video 3

□ Uploaded video 4

□ Uploaded vic

□ Uploaded video 6

□ Uploaded video 7

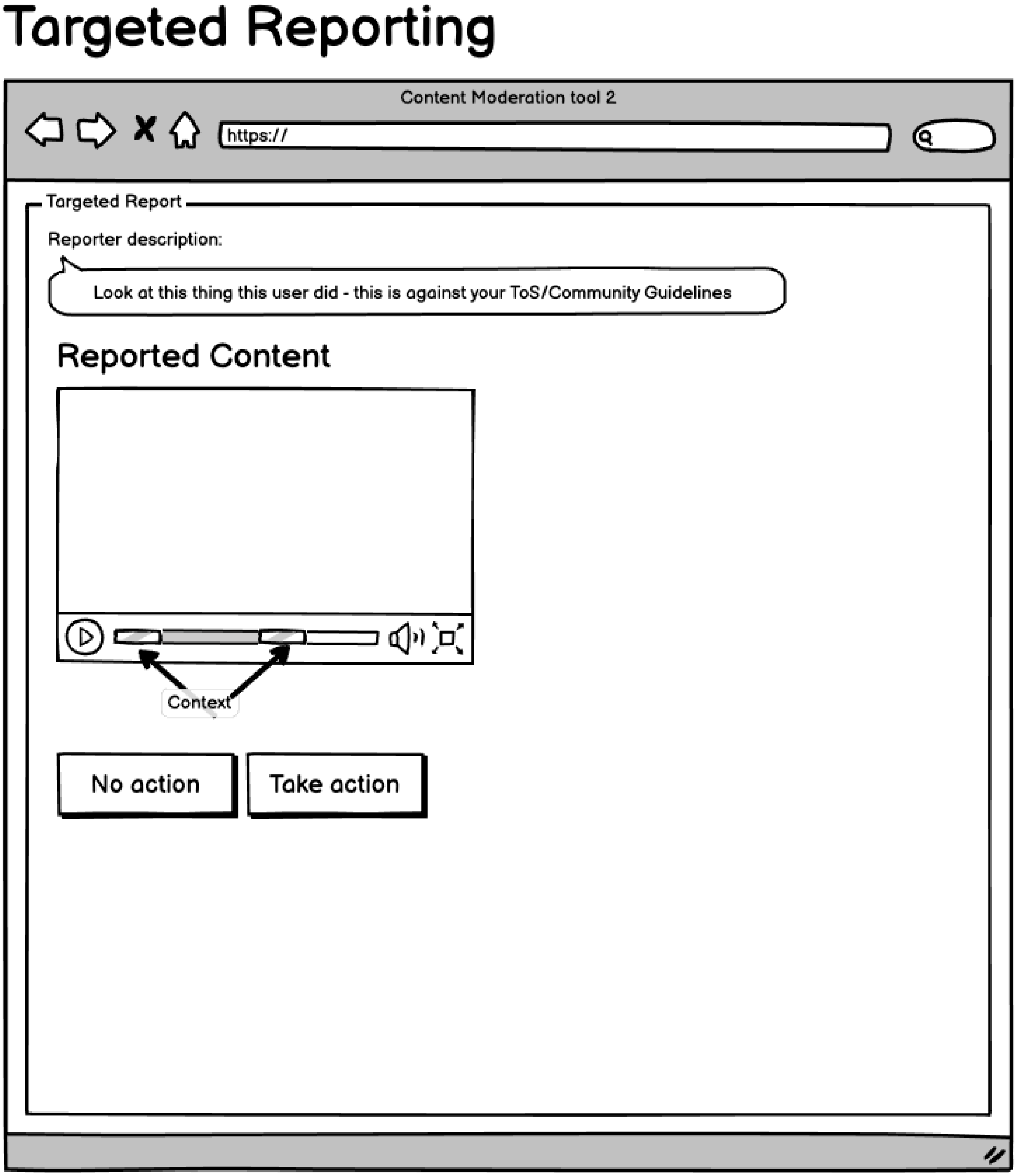
□ Uploaded video 8

Page 3 of 5

Fig

Fig 4 - User reporting content moderation tool

Fig 5, on the other hand, shows an example of how a content moderation tool might look if the reporting mechanism utilizes ‘Targeted Reporting’ (as described in Fig 2). Due to the nature of how the report was submitted, only the UGC identified by the reporter, plus a small amount of context (as described in Fig 3), is required to be reviewed.



Look at this thing this user did - this is against your ToS/Community Guidelines

Fig 5 - Targeted reporting content moderation tool

While these are again intentionally simplistic examples, they demonstrate the downstream effects of decisions made during the design of reporting mechanisms. Moreover, while the effects of these decisions when reviewing individual reports may not seem significant, they multiply exponentially with scale.

In the below **example** we demonstrate how this would translate to total workforce hours required across one thousand reports received by a service or game:

	User report	**Signal report**	**Targeted report**
Requirements	Investigate the reported user's activity	Review an entire piece of User-Generated Content	Review the section of UGC that has been identified as violative by the reporting user
Volume of reports	1000	1000	1000
Average Handle Time (mins)	30	12	1
Total (mins)	30000	12000	1000
Total (Hrs)	500	200	17

While the above metrics are framed around resourcing hours—and this risks the perception of being overly-focused on cost-efficiencies vs. user benefit—in reality these two factors are inextricably linked. Unless cost-efficient, scalable mechanisms are utilized in content moderation, systems designed to safeguard users become cost-prohibitive and ultimately less effective, especially as user bases scale. Ensuring transactional processes remain predictable in terms of handle times, while empowering users with mechanisms that allow them to accurately identify violative content, will drive a more efficient and hence ultimately more effective reactive moderation solution.

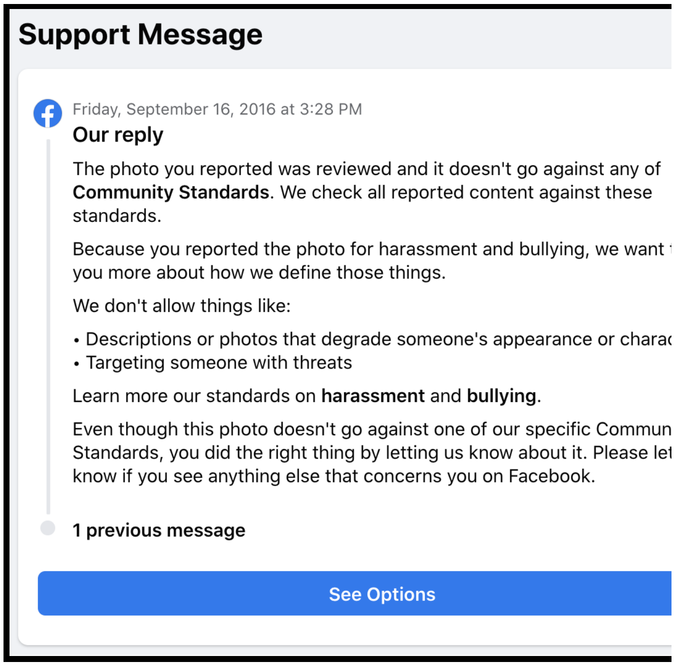
Education

While the preemptive communication of standards described in section Community Guidelines/Code of Conduct is the first step in preventative measures against violative content, the component of the flywheel we are calling 'Education' is the reactive equivalent. When violative content is identified (whether via the reactive content moderation model primarily described here or, indeed, via any of the other approaches outlined in Appendix A) communicating the results of content moderation review to relevant parties is similarly important in preventing future violation and/or misuse of reporting functions.

Some of this approach already in use include:

Example 1 - Facebook

Education for reporters - Facebook allows reporters to monitor the status and outcome of their report via a Support Inbox.



Support Message

Friday, September 16, 2016 at 3:28 PM

Our reply

The photo you reported was reviewed and it doesn't go against any of **Community Standards**. We check all reported content against these standards.

Because you reported the photo for harassment and bullying, we want t you more about how we define those things.

We don't allow things like:

• Descriptions or photos that degrade someone's appearance or charac
• Targeting someone with threats

Learn more our standards on **harassment** and **bullying**.

Even though this photo doesn't go against one of our specific Commun Standards, you did the right thing by letting us know about it. Please let know if you see anything else that concerns you on Facebook.

1 previous message

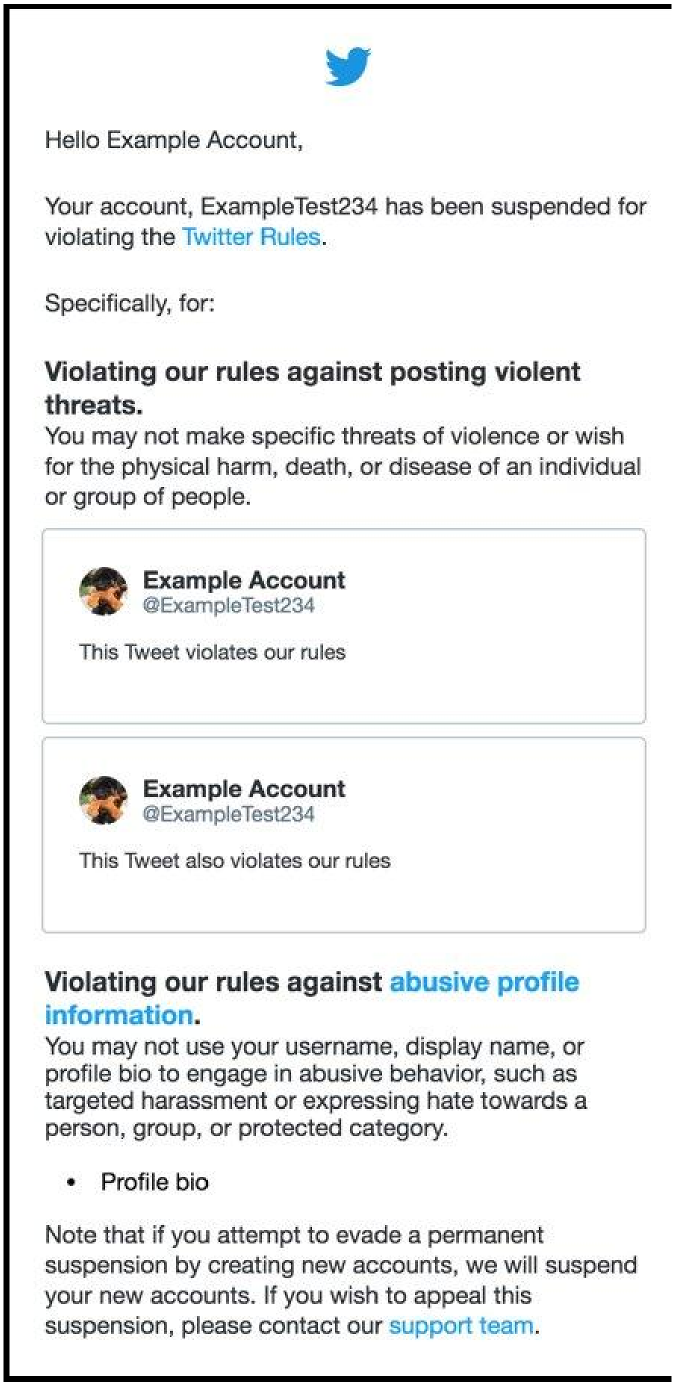
Even if no moderation action is taken following review, the response a reporter receives still includes additional information about the specific aspect of the Community Standards the user selected in their original report.

This kind of information helps to:

- Continue to reinforce the Community Guidelines/Code of Conduct
- Improve the community's understanding of what is and isn't allowed on the service
- Improve the quality of future reports so they are more closely aligned with the rules of the community

- Example 2 - Twitter
  - *Education for offenders* - If a Twitter user is suspended, they receive a notification[12] informing of the action that has been taken and why.

[12] Twitter Safety



Hello Example Account,

Your account, ExampleTest234 has been suspended for violating the Twitter Rules.

Specifically, for:

Violating our rules against posting violent threats.
You may not make specific threats of violence or wish for the physical harm, death, or disease of an individual or group of people.

Example Account
@ExampleTest234

This Tweet violates our rules

Example Account
@ExampleTest234

This Tweet also violates our rules

Violating our rules against abusive profile information.
You may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

- Profile bio

Note that if you attempt to evade a permanent suspension by creating new accounts, we will suspend your new accounts. If you wish to appeal this suspension, please contact our support team.

The notification includes all of the following information:

- The account suspended
- The rules that have been violated
- The specific content (in this case the tweet) which led to the enforcement action.
  - This is a key point that is not often present in enforcement notifications but is vital in helping the offender understand what it was they did that violated the rules.
  - In many cases users will create a large amount of content and so helping them to understand which specific piece of UGC was violative will help them avoid repetition.
  - The fact that the reporting user can specify the offending tweet (and up to 5 additional tweets) during the reporting process will likely aid in building mechanisms that allow it to also be included in the enforcement notification. This demonstrates how the **content moderation flywheel** can be a virtuous cycle, as one aspect of it informs another.

Not only does the available data support this approach, it’s also far from new information for the industry. As far back as 2013, Riot Games published[13] that by simply providing the information relevant to an account action via ‘Reform Cards’[14] helped to reduce future reports by as much as 13.2%.

[13] Using science to reform toxic player behavior in League of Legends, Arstechnica
[14] In-Client Reform Cards FAQ, League of Legends Support

And subsequent analyses have continued to reinforce this finding. According to a study on Reddit Moderation undertaken by the Georgia Institute of Technology and the University of Michigan[15]: “Our regression analyses... show that when moderated users are provided explanations, their subsequent post removal rate decreases”, going on to calculate that ...the odds of future post removals would reduce by 20.8% if explanations were required to be provided for all removals”.

Technology

Numerous standalone papers can be written about the use of technology to detect and prevent violating content on online services; this is therefore well beyond the scope of this paper. However, it is important to note the role that being targeted about reporting and content moderation can play in improving the efficacy of these systems.

By enabling reporting mechanisms to target distinct and finite elements or moments of violating behaviour, we facilitate labelling via content moderators’ actions of more discrete data sets. These can be used to feed machine learning and therefore improve our ability for proactive moderation. This completes the virtuous cycle of learning and improvement set out in the Content Moderation Flywheel by ‘narrowing the funnel’ and continually reducing the volume of violative content users need to report.

[15] Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit, Jhavar, Shagun et al

Conclusion

The purpose of this whitepaper is to cater and optimise for a specific use case, which is as follows:

i. I experienced a bad thing
ii. I want to tell someone about it
iii. I want the people I tell about it to do something meaningful

Providing for a ‘perfect’ system that eliminates (i) and ensures that disruption and harm can never occur online is well beyond our scope. If such a thing is achievable, we will likely look upstream towards the systems of design in our products, education of digital citizens and the mechanisms we use to incentivise types of behaviour as larger societies.

However, by building our systems to optimise for (ii) and by aiding the user who experienced the ‘bad thing’ to identify and specify what it was–allowing them to harness the verifiable evidence of the event captured by service provider’s systems–we create the conditions in which (iii) is consistently possible, even at extreme scale.

Appendix A - Glossary of terms

Pre-moderation: This is the process of reviewing all UGC before it goes live or is visible in your application or service. Only 'accepted' (i.e. does not violate ToS/guidelines) will be accessible within the app. This minimizes risk of ToS-violating content being present in your application (only ever appearing due to moderation error) and is often used for applications intended for younger or vulnerable audiences and/or content that is in some way viewed as 'endorsed' (e.g. recommendations).

Post-moderation: This is the process of reviewing all UGC but without requiring an 'accept' flag before content is visible in your application or service. All content will be accessible within the app as soon as it is generated but all will be given an 'accepted' or 'rejected' flag at some point. The benefits of this approach is that it provides a complete view of the nature of UGC present in your application or service, but is highly-labor-intensive and does not scale efficiently.

Reactive-moderation: This is the process of only reviewing UGC when it is flagged to you by a user who thinks it may violate your ToS/guidelines. This has the benefits of being more scalable and is likely to result in a higher 'Validity' of work handled than either pre or post moderation, but drawbacks of providing a much lower sample size in terms of the nature of UGC present in your application, as well increasing the risk that users of your application or service are exposed to ToS-violative content. This is often the first line of defence used by large social services or platforms but is increasingly supplemented by technology-driven proactive detection/moderation[16].

Proactive-moderation: This is the process whereby potentially-violative UGC that is live and accessible in your application or service is sought out and flagged for review and/or action taken. Traditionally, moderators might use simple search mechanisms to seek out this kind of content - e.g. keywords searches in online forums or in-game admin functionality in online games[17]—increasingly Artificial Intelligence (AI) is used to create proactive flagging (or reporting) for human-review, the results of which are fed back into the AI to further improve its efficiency (i.e. Machine Learning). This may include action that is immediately taken by the proactive technology (e.g. hiding the content while it is reviewed and/or removing it entirely with review for high-confidence cases) or a human-in-the-loop process that functions similarly to reactive moderation, only that the source of the initial 'report' comes from a system and/or agent whose job it is to identify violative content (as opposed to a report from a user of the service).

[16] For Q4 2019, YouTube's transparency report shows that ~91% of content removals were a result of flags (or reports) created by technology, with only ~5% a result of human-generated flags/reports (the more traditional approach to 'reactive moderation').

[17] Admin tools used in the PlayStation 3 online social service 'PlayStation Home'