

QUICK REFERENCE

Being ‘Targeted’ about Content Moderation

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

By **Robert Lewington** (Senior Director of Safety Operations @Twitch) & the Fair Play Alliance Executive Steering Committee.
All examples are expanded in the [full paper](#).

About this document

This document is intended to be a quick reference for a number of key concepts in the “Being ‘Targeted’ about Content Moderation” white paper (see [full whitepaper](#)).

This purpose of the whitepaper is to provide 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 primarily reactive moderation, where a service provider responds to reports submitted by users of its service regarding UGC that may violate its Terms of Service.

Contents

Community Guidelines

Targeted Reporting

and the importance of context

Scalable Content Moderation

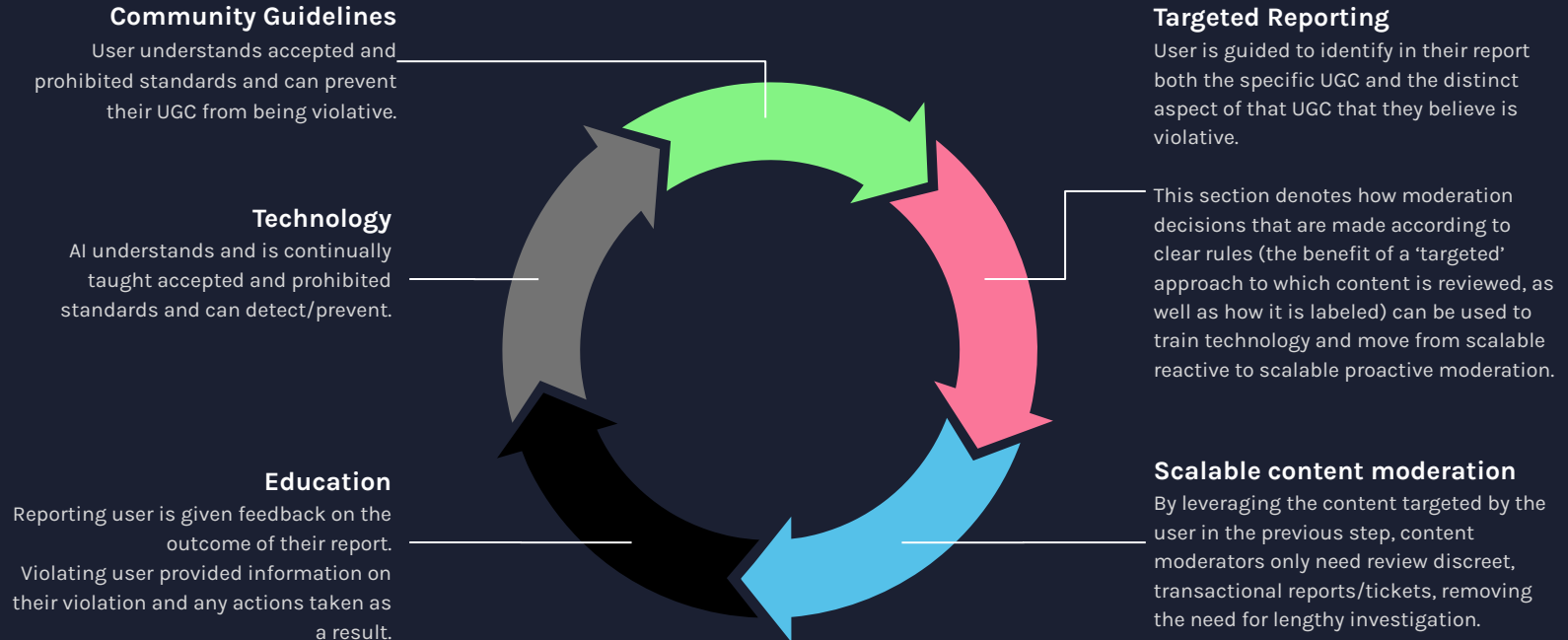
Design the tool upstream

Education

Technology

Content Moderation Flywheel

How targeted reporting functions within a content moderation cycle



Community Guidelines / Code of Conduct

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.

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”

Did you know?

A study by the [National Academy of Sciences](#) 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.

Targeted reporting

In a model that utilizes reactive moderation, the quality of the reports you receive from users is vitally important to scalability. Using a mental model of comparing 'online' content moderation to 'offline' forms of investigation and enforcement, we can think of this as akin to offline police work. '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.

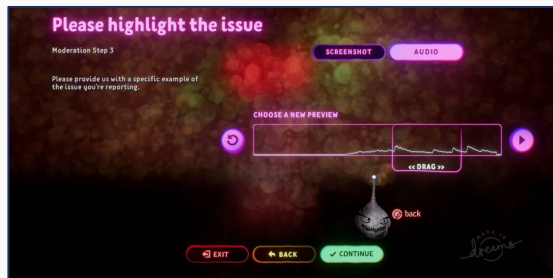
01



Overwatch

These 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.

02



Media Molecule's Dreams

This includes an in-game reporting mechanism for flagging user-generated creations. Each creation takes the form of a fully-fledged interactive experience comprising a vast amount of individually created objects and/or media. 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.

03



Twitter

Twitter's reporting mechanism which allows the reporter to specify up to 5 additional tweets from the reported user.

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.

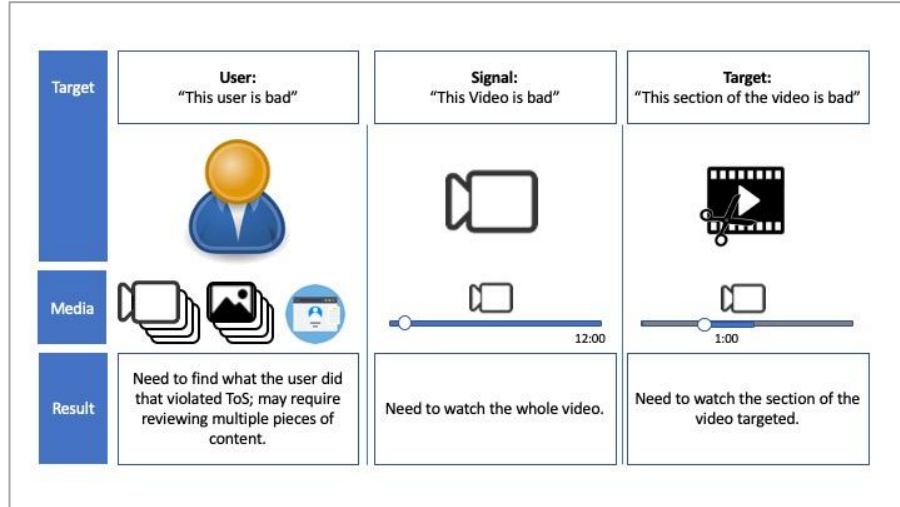
Targeted reporting

The next aspect of the ‘targeted reporting’ approach is detailed (left), where three potential options for a reporting mechanics are shown, each narrowing down the specifics of the complaint to an increasingly focused ‘target’.

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).

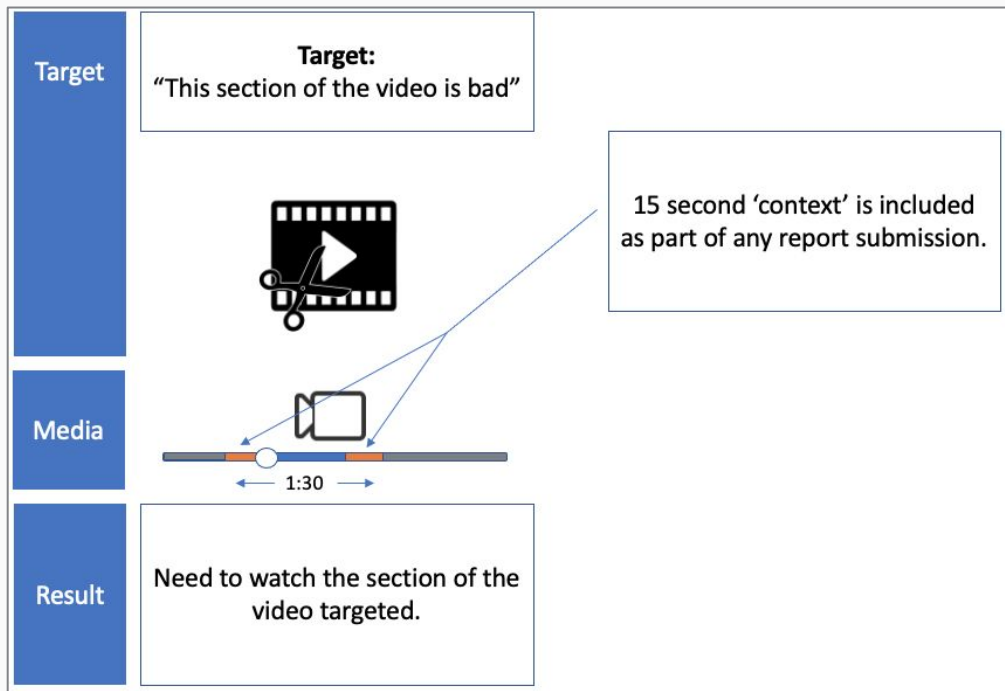
This will vary greatly depending on the nature of the report and/or social application or service in question, but is based on an assumption that the user may have created multiple pieces of content, including videos of 12 minutes in length (as denoted by the ‘media’ and ‘signal report’ sections).

These examples are intentionally simplistic and this type of content moderation can take many and varied forms, but this is a high-level illustration of the premise of targeted reporting. In practice there are multiple wrinkles to be considered in this approach such as ‘context’ and we will tackle this on the following slide.



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).



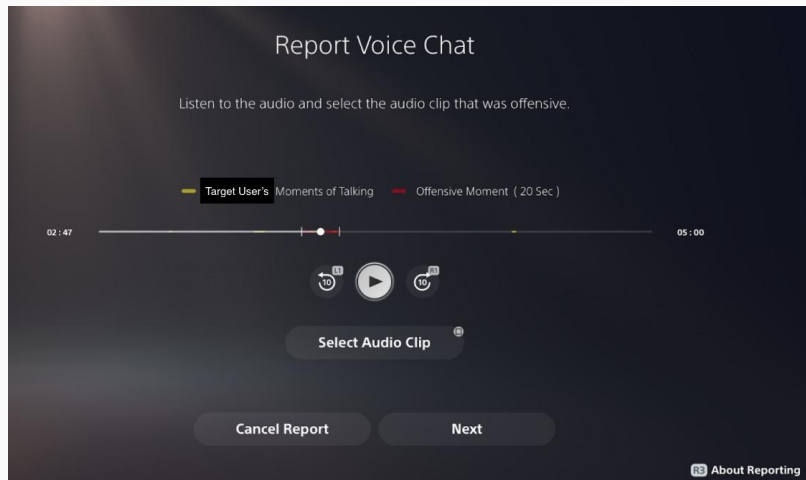
While the content that requires review has now grown by 50% versus the narrower ‘target’ approach—therefore reducing some of the operational gain described in the previous slide—it is still 8x shorter than the ‘signal’ approach (also shown on slide 6).

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

Context

Again, the example shown on the [previous slide](#) is both broad and intended purely as an illustration of the concept, but a very similar application of the concept can be found in existing online gaming reporting systems:

01

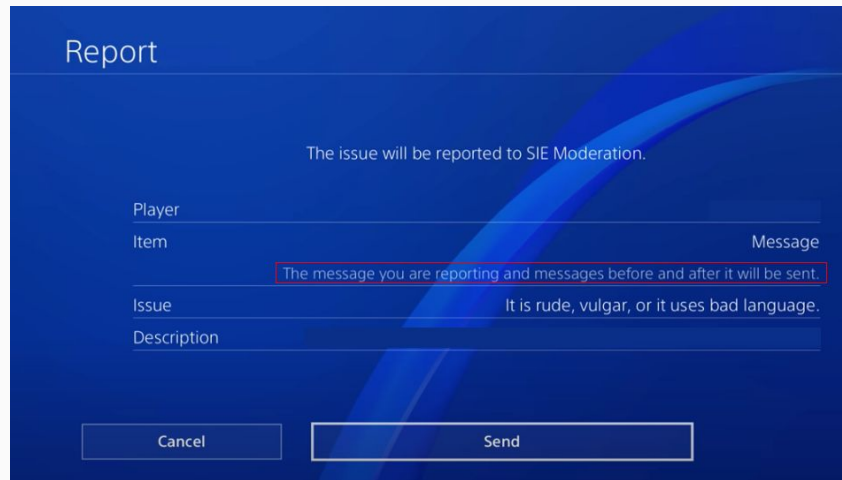


PS5 Voice Comms reporting

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.”

02



PS4 Messaging Reporting

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

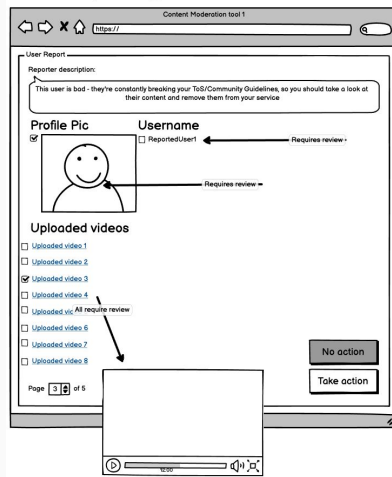
“The message you are reporting and messages before and after it will be sent [to SIE Moderation].”

Scalable Content Moderation

Using the examples of types of UGC reporting shown so far, we can see how targeted reporting can help scale 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.

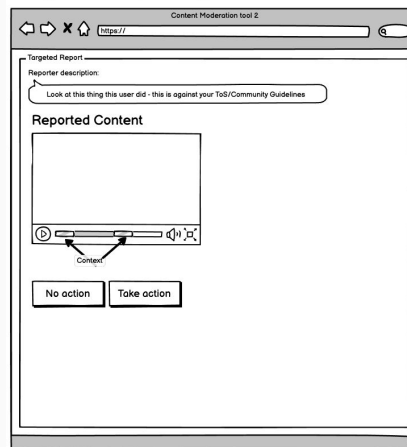
The tool can now simply 'queue' the specific pieces of content identified by the reporting users (plus any relevant context) and present it to the moderator. It also means that the moderator can perform consistent and replicable processes on the content they are reviewing.

User Reporting



This example shows how a content moderation tool might look if the reporting mechanism utilizes 'User' reporting (the first method shown on [slide 6](#)). This requires a significant amount of content to be reviewed for each individual report.

Targeted Reporting



This example, on the other hand, shows how a content moderation tool might look if using 'Targeted Reporting' (the third method shown on [slide 5](#)).

Now only the UGC identified by the reporter, plus a small amount of context, is required to be reviewed.

Scalable Content Moderation

As with all content in these slides, the examples shown are intentionally simplistic, but 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 next table (right), we demonstrate how this would translate to total workforce hours required across one thousand reports received by a service or game, calling back the [three potential reporting approaches](#).

	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 is 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 the [Community Guidelines section](#) 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, communicating the results of content moderation review to relevant parties is important in preventing future violation and/or misuse of reporting functions. We can think of this as **reinforcing the guidelines** that were initially communicated.

Did you know?

As far back as 2013, [Riot Games published](#) that by simply providing the information relevant to an account action via 'Reform Cards' helped to reduce future reports by as much as 13.2%.

Subsequent analyses have reinforced this finding. According to [a study on Reddit Moderation](#) undertaken by the Georgia Institute of Technology and the University of Michigan:

“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”.

Education

Two examples of this approach already in use:

01

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 to tell you more about how we define those things.

We don't allow things like:

- Descriptions or photos that degrade someone's appearance or character
- Targeting someone with threats

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

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

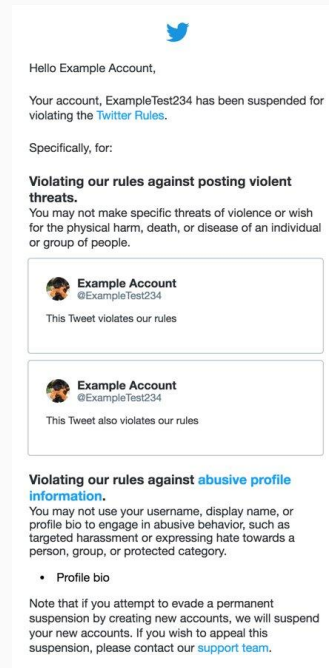
1 previous message

See Options

Facebook

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 in order to (i) Continue to reinforce guidelines (ii) improve the community's understanding of what is and isn't allowed (iii) improve the quality of future reports to more align with the rules of the community

02




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](#).

Twitter

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.

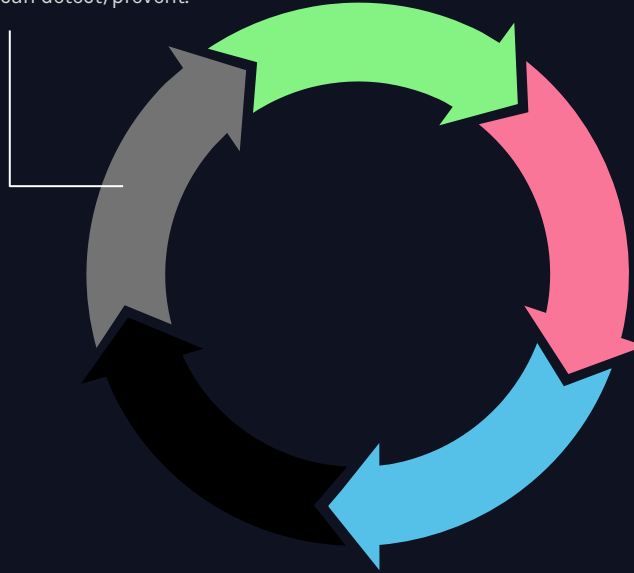
Technology

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

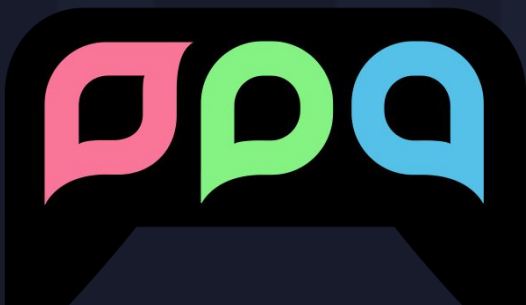
By enabling reporting mechanisms to target distinct and finite elements or moments of violating behaviour, we facilitate labelling via content moderators' actions of more discrete data sets. These can be used to feed machine learning and therefore improve our ability for proactive moderation.

Technology

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



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.



QUICK REFERENCE

Being 'Targeted' about Content Moderation

By Robert Lewington (Senior Director of Safety Operations @Twitch) & the Fair Play Alliance Executive Steering Committee.

If you enjoyed or were interested in the content of these slides, please read the [full whitepaper](#).

Fair Play Alliance: info@fairplayalliance.org