Marinett, Matthew

Article

The new frontier of platform policy

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Marinett, Matthew (2021) : The new frontier of platform policy, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 10, Iss. 3, pp. 1-31,
https://doi.org/10.14763/2021.3.1570

This Version is available at:
https://hdl.handle.net/10419/245336

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 10 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

The new frontier of platform policy

Matthew Marinett *University of Toronto* matthew.marinett@utoronto.ca

DOI: https://doi.org/10.14763/2021.3.1570

Published: 13 September 2021
Received: 16 April 2021 **Accepted:** 10 June 2021

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Marinett, M. (2021). The new frontier of platform policy. *Internet Policy Review*, *10*(3). https://doi.org/10.14763/2021.3.1570

Keywords: Platform policies, Social media, Twitch, Platform governance, Off-platform behaviour

Abstract: Platform policies aimed at the misbehaviour of users that occurs off of the platform, especially offline abuse, are a relatively new and understudied phenomenon that may represent a new frontier of platform policy. However, policies of this nature raise unique problems in comparison to on-platform content moderation that exacerbate existing concerns about the accountability of platforms by demanding both heightened due process for the accused and strong privacy and safety protections for complainants. This article discusses these challenges through the case study of Twitch.tv, and argues that further steps must be taken to ensure accountability to users and the public when such policies are implemented.

Introduction

A great deal of academic literature has considered the policies and enforcement actions of online platforms with respect to content moderation (Klonick, 2018; Gillespie, 2018; Suzor, 2019; Douek, 2021), but there remains little, if any, literature addressing platform policies and enforcement actions targeting off-platform harassment and abuse by users. While few platforms explicitly have policies of this nature, live-streaming platform Twitch has become a pioneer in this space, having developed and enforced such a policy since 2018 (Twitch, 2018). [1] This policy was most notably put into effect during the summer of 2020 when the gaming world experienced an outpouring of sexual assault and harassment allegations that reached all corners of the industry, from major publishers and developers to broadcasters to community event organisers (Martens, 2020; Hall, 2020). While mainstream news media primarily focused on allegations about powerful men within large game development companies, a simultaneous outpouring of stories concerned Twitch streamers at various levels of popularity (D'Anastasio, 2020). In response, Twitch indefinitely suspended a number of streamers that had been identified in public sexual harassment and assault allegations (Kastrenakes, 2020; Hernandez, 2020).

While these were not the first enforcement actions taken by Twitch for off-platform abuse, the comparatively large number of bans over a short period of time associated with a wave of sexual abuse revelations suggests Twitch's policy may represent a new frontier of platform policy: enforcing policies against off-platform abuse. Twitch's active enforcement of policies against off-platform abuse goes beyond issues of content moderation to raise questions about the competence and accountability of platforms in imposing consequences on users for their behaviour regardless of where it occurs. In this article "abuse" is meant to signify a broad category of behaviour, whether online or offline, that targets and harms individuals, such as assault, harassment, bullying, communicating threats, or disclosing personal information about an individual, but does not include, for example, membership in a criminal organisation or terrorist group, or publishing hate speech or dangerous misinformation, even if some of the same concerns raised here may apply.

Policies against off-platform abuse may well be a positive step in producing healthier communities and ensuring consequences for harmful acts regardless of

1. While academic work on Twitch's policy enforcement is limited, a notable exception is Taylor's detailed account of the rise of the platform, which addresses on-platform policy enforcement, although not off-platform policy (2018). Other academic treatments of Twitch's policies often concern the use of policies to control women's attire and sexual content (Zolides, 2020; Ruberg, 2020).

where they occur. In Twitch's case, such policies follow long-standing problems with misogyny and harassment levelled at women on the platform (Taylor, 2018; Kastrenakes, 2020), and Twitch has stated that such policies are aimed at providing an environment that feels safe for all users (Twitch, 2018). Nonetheless, policies that undertake to investigate and sanction off-platform abuse even in the absence of legal sanction or widespread reporting raise unique problems in comparison to on-platform content moderation that increase the difficulty of balancing the positive aims of such policies with maintaining accountability and fairness. These problems do not suggest that such policies should not exist, but rather they suggest that where such policies are implemented, a significant commitment must be taken to balance the goals of such a policy with accountability and fairness in its enforcement. This article will discuss three such problems with respect to policies against off-platform abuse, like those of Twitch, with a special focus on sexual assault following the events of the summer of 2020.

First, platforms have little experience in setting policy for off-platform abuse and may lack competence in investigating and verifying such behaviour. Unlike with on-platform content moderation, platforms lack special access to the facts of the alleged conduct. Instead, platforms must rely on evidence provided by third-parties, but platforms lack expertise in obtaining, verifying, and weighing such evidence, which may increase the likelihood of error or undermine the effectiveness of a policy. Second, enforcing off-platform policies for abuse will often require engaging with highly sensitive information and events. Although not all investigations follow a direct report to Twitch—as many of those in the summer of 2020 began after public allegations rather than reports—many will nonetheless be based on direct reports. This heightens the need to be clear about the decision-making and investigatory process following such reports, and the privacy-invasive nature of such investigations heightens the dangers of failing to protect victim or complainant privacy and safety. Third, the impact of this kind of enforcement on sanctioned users is also potentially greater than in cases of content moderation enforcement actions. As the bans in the summer of 2020 demonstrate, a common enforcement action taken for off-platform behaviour is an indefinite account suspension, which can have significant and long-lasting impacts on a users' social connectedness, wellbeing, and ability to earn an income (Gillespie, 2018, p. 176). Where enforcement follows a public accusation, it may both increase public attention to the allegations and be seen as a confirmation of those allegations, which could exacerbate public stigma.

These three elements of enforcement for off-platform abuse exacerbate the core

problems of accountability and transparency that plague social media content moderation (Klonick, 2018; Balkin, 2018; Suzor, 2019; Douek, 2019) by increasing the challenge of balancing the objectives of the policy with heightened needs for due process, transparency, and confidentiality in enforcement given the increased stakes and risk of errors.

This article does not aim to fully resolve these tensions, although it suggests some areas where Twitch's policies could be more transparent and accountable than at present. It also points to the need for considerable future public discussion and research in this area as it remains possible that policies of this nature could become more common among platforms, especially those that create asymmetric relationships between content creators and their audiences. Crafting the proper balance between the safety-based objectives of policies against off-platform abuse and fairness and accountability in enforcement will require a broader public discussion about the proper role of platforms in sanctioning such behaviour. Regulators may want to take cognisance of the important difference between policies aimed at on-platform behaviour and those aimed at off-platform behaviour in fashioning regulations that impose procedural accountability obligations, including the EU's *Digital Services Act* (Proposal DSA).

This article proceeds as follows. Section 1 reviews Twitch's policy with respect to off-platform abuse and its history and practice of enforcement. Section 2 discusses the unique nature of policies against off-platform abuse, including the significant potential impacts upon both the reporting and reported individuals and the complexity and uncertainty introduced by the need to make factual determinations based on external evidence. Section 3 briefly explores the potential for policies against off-platform abuse to spread to other kinds of social media platforms. The article concludes by considering how lawmakers might respond to these challenges when crafting accountability requirements for platforms.

1. Twitch and policies against off-platform abuse

a. Policy background

Launched in 2011 as a spinoff of the "lifestreaming" website Justin.tv, Twitch is a video-streaming platform originally focused on the live-streaming of video game play, although it has since branched out to numerous other content areas. Twitch is now the 35th most visited site on the internet at the time of writing (Alexa, n.d.) and commands the largest share of the online live-streaming market (Iqbal, 2021). Content on Twitch is, like YouTube, provided by third-parties (who are often indi-

vidual users) rather than created by Twitch itself (Taylor, 2018). Almost anyone can create an account and, provided they have access to streaming software and the proper hardware, begin live-streaming on their own "channel". The company was purchased by Amazon in 2014, and Twitch's current business model relies primarily on advertising embedded in streams as well as optional subscriptions to either Twitch as a whole (via Twitch Turbo) or individual channels (Iqbal, 2021).

Streamers on Twitch are able to earn money through the platform's Affiliate and Partner programmes, with Partner being the more lucrative of the two. Partners and Affiliates are able to earn money from channel subscriptions, advertising on their streams, as well as a form of payment from viewers to streamers known as "Bits". Popular streamers may also have separate contracts with Twitch that may provide different benefits and terms in exchange for streaming exclusively on the platform (Gilbert, 2020). Twitch has over two million broadcasters and over 27,000 Partners (Twitch, n.d.d.). Twitch thus supports a sizable community of streamers, many of whom earn their entire living from streaming on the platform (Taylor, 2018; Wiltshire, 2019).

All streamers and users on Twitch, including Partners and Affiliates, are required to abide by Twitch's Community Guidelines (Twitch, n.d.a). As Twitch is a live-streaming service that primarily broadcasts ephemeral content, unlike text-based platforms it does not focus its enforcement of its Community Guidelines on the removal of content. While it uses machine-learning tools to prevent some content contrary to its Community Guidelines from being broadcasted, its primary method of enforcement is at the account-level (Twitch, 2020c). Account enforcements can range from warnings to temporary suspensions to indefinite suspensions. According to Twitch's first ever transparency report, concerning the year 2020, Twitch carried out over 1.1 million account enforcements during the second half of 2020 alone (Twitch, 2020c). However, the report did not specify how many of these enforcements were merely warnings compared with account suspensions.

Twitch made the decision in February 2018 to make enforcement for certain off-platform behaviours, including assault and harassment, an express part of its enforcement mandate in order to better protect its community (Twitch, 2018). This meant the inclusion of a relatively simple line in its policies that stated "[w]e may take action against persons for hateful conduct or harassment that occurs off Twitch services that is directed at Twitch users" (Twitch, n.d.a). Beyond potentially applying Twitch's broader hateful conduct and harassment policy to the off-platform behaviour of its users, when or how this policy would be enforced remained unclear.

Twitch stated that the reason for going after off-platform conduct was that "ignoring conduct when we are able to verify and attribute it to a Twitch account compromises one of our most important goals: every Twitch user can bring their whole authentic selves to the Twitch community without fear of harassment" (Twitch 2018). It's understandable that Twitch would want to prevent giving a platform to perpetrators of abuse and harassment, which may exacerbate the harms experienced by victims. This is especially true for a company at the centre of a gaming culture that has often been observed to be exclusionary and hostile to women's presence and involvement (Taylor, 2018). As Twitch CEO Emmett Shear later stated in an internal company email following the 2020 sexual harassment and abuse allegations, the company wanted to "set a higher standard for ourselves and those with power and influence on our service" (Shear, 2020).

It's unclear how often this policy has been enforced since 2018, as Twitch does not generally comment on enforcement actions. Twitch's 2020 transparency report contained no information about enforcement of policies for off-platform behaviour (Twitch, 2020c). Details of its enforcement are thus limited to sporadic news reports that, by their nature, skew towards more popular streamers. [2]

Despite some earlier cases, it wasn't until the summer of 2020 that the application of Twitch's off-platform policy to in-person interactions was truly tested. Over that summer, a "#MeToo" movement swept across the gaming industry, with an overwhelming number of individuals coming forward with personal stories of sexual harassment and assault in all areas of the industry, from game development to tournament organising to streaming (Schreier, 2020). While many stories did not identify perpetrators, many did, and dozens of men, and some women, were named as perpetrators. One streamer, Jessica Richey, created a Google Docs document that categorised over 400 personal stories (Lorenz & Browning, 2020).

In June and July, Twitch indefinitely suspended the accounts of several prominent streamers that had been identified in these stories without public comment. These included, for example, Gonzalo "ZeRo" Barrios, a major personality and former top competitor in the fighting game community. Barrios was banned shortly after several women within the community publicly alleged that Barrios had engaged in sexual misconduct, including sending sexually explicit messages to minors, and after Barrios admitted to some of that conduct (Galiz-Rowe, 2020). Others banned at

2. One example of the enforcement of off-platform policies prior to the summer of 2020 is the suspension of Gregory 'Onision' Jackson in January of 2020 following a series of allegations of abuse and grooming minors. His account was controversially restored in October of 2020 without public comment from Twitch (Colombo, 2020).

the time included popular streamer Brad "BlessRNG" Jolly; high-level fighting-game competitor Nairoby 'Nairo' Quezada; and at least several others, including those going by the monikers "iAmSp00n", "SayNoToRage", "DreadedCone", "Wolv21", and "WarwitchTV" (Hernandez, 2020; Kastrenakes, 2020; Walker, 2020). As Twitch does not comment on individual enforcements, it is not known if these actions followed reports to Twitch, or whether Twitch took action based solely on public disclosures.

On 24 June 2020 Twitch provided a general response to the revelations and allegations in a short blog post, which expressly referred to investigations for behaviour that took place off of the Twitch platform:

> We are reviewing each case that has come to light as quickly as possible, while ensuring appropriate due diligence as we assess these serious allegations. We've prioritised the most severe cases and will begin issuing permanent suspensions in line with our findings immediately. In many of the cases, the alleged incident took place off Twitch, and we need more information to make a determination (Twitch, 2020b).

Since then, Twitch has not made any public statements concerning its investigations or processes with respect to these cases, nor has it made any public statements explaining its actions in certain cases. Twitch did not elaborate on what 'due diligence' entailed, and it did not publicly provide reasons for why numerous other Twitch streamers accused of serious misconduct have not received account suspensions. It is also not clear if any of the accused were provided with the opportunity to respond.

This level of opacity is not unusual for the company. Twitch had already been the subject of much public commentary alleging it is inconsistent in content policy enforcement and fails to provide sufficient reasons for its enforcement decisions, even to those facing penalties (Asarch, 2019; Geigner, 2019). Indeed, Twitch's lack of clarity and consistency in content moderation was addressed directly by Twitch CEO Emmett Sheer during the company's annual convention, TwitchCon, in 2019 where he promised to increase transparency around policy enforcement processes broadly. He stated that "[n]o matter how good your decision-making process is, if people can't understand it they can't fully trust it. We're going to really focus on increasing that transparency so people can trust the process" (Shanley, 2019). However, while Twitch has improved transparency in some ways, such as through its new transparency reports, much remains hidden. It remains difficult or impossible

to find information concerning Twitch's internal policy development process, the size or operating procedure of its content moderation team, or how often it indefinitely suspends accounts.

b. The current policy

In an apparent response to the challenges of enforcing its policy that arose during the summer of 2020, Twitch significantly updated its policy regarding off-platform behaviour in April of 2021. Twitch announced that it had hired an unnamed outside law firm with expertise in workplace and campus sexual assault cases to carry out its investigations (Twitch, 2021). Whereas the policy previously contemplated offline harassment or hate broadly, this update limited the policy to enforcement against relatively egregious offline misbehaviours including:

- Deadly violence and violent extremism
- Terrorist activities or recruiting
- Explicit and/or credible threats of mass violence [...]
- Carrying out or deliberately acting as an accomplice to non-consensual sexual activities and/or sexual assault
- Sexual exploitation of children [...]
- Actions that would directly and explicitly compromise the physical safety of the Twitch community [and]
- Explicit and/or credible threats against Twitch (Twitch, n.d.-b).

In the blog post announcing the policy update, Twitch explained that "we only take action when there is evidence, which may include links, screenshots, video of off-Twitch behavior, interviews, police filings or interactions, that have been verified by our law enforcement response team or our third party investigators" (Twitch, 2021). However, official law enforcement or criminal justice action is not necessary for Twitch to enforce its policy. Instead, the company will take action where there is a "preponderance of evidence" that the behaviour took place. According to one reporter, Twitch stated that those under investigation will have an opportunity to respond (Newton, 2021), although that does not appear in the blog post or official policy. Additionally, contrary to the previous version of the policy, the new policy also states that it may take enforcement action for violations that target non-Twitch users. The policy will also apply to acts that occurred even before the policy violator was a Twitch user (Twitch, n.d.-b).

Twitch stated that its enforcement actions may include account suspensions, including "indefinite" suspensions. It's not entirely clear if there is a meaningful distinction between an indefinite and a permanent suspension. In the summer of 2020, Twitch said it had been issuing "permanent" suspensions in response to its

investigations (Twitch, 2020b). Presumably these terms are largely synonymous, and the term "indefinite" is used only to leave open the vague possibility that a suspension could be rescinded.

It's also not clear whether Twitch will now only begin an investigation following a direct complaint, but a channel to report directly to Twitch's new Off-Service Investigations Team was made available (Twitch, n.d.-b).

It should be noted that, while the high-profile bans for which any information is available have been directed at streamers, the policy is not limited to streamers, and theoretically all users, whether streamers or viewers, are subject to the policy. Whether any Twitch user that is not a regular streamer has been suspended under the off-platform policy is unknown, as Twitch neither publicly discloses any information about these suspensions, nor would such a user's suspension likely attract public notice.

While it is understandable that Twitch will not provide information on individual enforcement decisions due to privacy and confidentiality concerns (Twitch, 2021), there is also little information on what either those that report violators or those under investigation can expect. Twitch did not provide information on how evidence will be evaluated or verified, what degree of involvement either party will have in the decision, or the availability of appeals, if new evidence comes to light.

The policy as it stands is perhaps limited compared to its earlier incarnation, as the behaviours listed here appear to correspond with widely criminalised behaviours, whereas the previous policy could presumably apply to many non-criminal behaviours. It should be noted that the list of prohibited off-platform behaviours provided by Twitch and reproduced above is not exhaustive, and Twitch suggested that this policy may expand to include other behaviour beyond those listed currently in the policy (Newton, 2021).

Some of the current categories of prohibited behaviour in the new policy correspond to some existing policies that consider off-platform behaviour of major social media companies like Twitter and Facebook, such as with respect to engaging in terrorist activities or recruiting (Twitter, n.d.; Facebook, n.d.-a). However, the focus of this article is those policies aimed at discrete off-platform abuse, especially sexual assault, that may require the platform to independently investigate. It is this kind of policy that puts Twitch, as Twitch Chief Operating Officer Sara Clemens said when discussing this policy, in "uncharted territory" (Newton, 2021). At the time, she stated she was unaware of any other platform with a similar policy (New-

ton, 2021).

Indeed, what makes Twitch's policy unusual and raises unique accountability concerns is that it targets off-platform abuse with identifiable victims and that it undertakes to investigate behaviour even in the absence of law enforcement action, judicial action, or widespread reporting. The existence of these concerns does not depend on whether the conduct the policies target takes place on another platform or offline, nor on whether that conduct rises to the level of a criminal offense.

The policy, of course, has many potential benefits. Most obviously, as Twitch has stated, it is aiming to ensure a safe environment for its users and to protect individuals from abuse and harassment (Twitch, 2018). In this sense, the threat of policy enforcement serves as a deterrent to misbehaviour and enforcement of the policy serves as a means of incapacitation to protect Twitch users from being victimised on the platform, even if it cannot directly stop off-platform harm from occurring.

Perhaps more important than either of these functions is the expressive value of the policies: how they signal social norms and set expectations and perceptions. As Cass Sunstein notes with respect to law, the expressive value of rules lies in their ability to change behaviour and norms and to set the expectations of people (Sunstein, 1996). As mentioned earlier, Twitch has long-standing problems with harassment targeting women and other minority groups on the platform (Taylor, 2018; Kastrenakes, 2020). The policy may thus communicate that Twitch is a less abusive and more welcoming environment for all users and potential users. Such a communication of norms may set certain expectations of the environment on Twitch, which may itself change social behaviour (Reynolds, Subašić, & Tindall, 2015). Moreover, even to the extent that it does not change behaviour, it serves as a signal that Twitch is taking action and, thus, may create the perception that the platform is safer for users. This naturally aligns with Twitch's own interest in continuing to grow the platform. There is presumably far more potential for growth in a service that appears inviting to a broad spectrum of people than one that appears to prioritise a small subset of abusive individuals.

This is especially true for a live-streaming service. Twitch is somewhat unlike some other social media platforms in its affordances and modes of engagement. While the affordances of other social media platforms such as Facebook and Twitter superficially create a rough parity between users by providing all users with the same basic set of tools with which to interact, at any given time, Twitch inherently divides users between streamers and audience members, creating a relationship

more akin to that between a broadcaster and the public. This asymmetric relationship is exacerbated by the fact that popular Twitch streamers can have concurrent viewership in the tens, or even hundreds of thousands, and can earn lucrative incomes from the platform. While Twitch's policies apply to any user regardless of whether that user is a streamer, it follows that higher standards of behaviour would exist on a service that provides some streamers a privileged position of relative power and influence, and where broadcasters may be seen as representatives of the platform. By failing to prevent perpetrators of harmful conduct from continuing to broadcast—or at least failing to appear to do so—one may see Twitch as rewarding and enabling such conduct and allowing further traumatisation of victims.

Finally, Twitch presumably also wants to maintain a space in which advertisers feel comfortable promoting their products and services on the platform. Advertising play a key role in influencing the content policies of major platforms (Caplan & Gillespie, 2020; Klonick, 2018, p. 1627), and advertisers have increasingly pressured social media platforms to moderate harmful or controversial content (Caplan & Gillespie, 2020; Hsu & Lutz, 2020). As Twitch noted in its Transparency Report, "[a]dvertising is an important part of Twitch, and brands that advertise on Twitch want to know how we are making our users safer, and promoting a more positive and less harmful environment" (Twitch, 2020c). Ensuring brand safety for advertisers is likely another key motivator of Twitch's policy against off-platform abuse.

Thus, there are many reasons that policies of this nature can play an important role on Twitch and other platforms. The argument of this paper is not that such policies should not exist; rather it is that they raise new challenges that exacerbate accountability and transparency concerns, and that additional steps may be necessary to ensure accountability to users. I now turn to a discussion of some of those challenges.

2. Special concerns with policies against off-platform abuse

The academic platform literature over the past several years has identified numerous accountability deficits of traditional platform content moderation (Bloch-Wehba, 2019; Balkin, 2018; Suzor, 2019; Douek, 2019). As content moderation is increasingly being understood as a form of governance (Klonick, 2018; Gorwa, 2019) that typically attempts to balance the interests of users against a variety of content harms, an increasing consensus of academic commentators recognises that obligations with respect to transparency, error-correction, and fair decision-making processes attach to the development and enforcement of content moderation poli-

cy (Douek, 2019; Suzor, 2019; Bunting, 2018). Mark Bunting refers to these principles as "procedural accountability", which aims to "encourage intermediaries to embed a concern for all the relevant impacts of their governance into the processes by which that governance is designed and executed, without specifying what the right rules may be for their particular context" (2018, p. 176) Similarly, Hannah Bloch-Wehba calls for the application of global administrative law norms—including those of transparency, due process, and public participation—to content governance (2019).

Such norms include those based on the rule of law and due process or procedural fairness, such as having clear policies, consistent enforcement, publicly transparent enforcement, notice to those impacted by a decision, the right for those impacted to provide evidence and argument, provision of reasons for a decision, and the opportunity to appeal that decision (Fuller, 2000; Benvenisti, 2014). These norms perform a variety of functions, such as reducing error, mitigating arbitrariness, protecting fundamental rights such as freedom of expression, [3] and promoting public legitimacy and trust in the policy enforcement process (Douek, 2019; Suzor, 2019, pp. 144-7). A central principle of natural justice and due process (or procedural fairness) is that the greater the impact of a decision upon an individual, the greater the need for attendant procedural safeguards (*R v. Secretary of State for the Home Department,* 1993; *Mathews v. Eldridge,* 1976; *Baker v. Canada*, 1999). For example, it's an internationally-accepted principle of administrative law and natural justice that individuals be entitled to participate in decisions that affect their lives (Benvenisti, 2014, p. 161). The opportunity of the person affected by a decision to provide evidence and argument reduces error rates by ensuring that errors can be captured and all evidence considered (Mullan, 2001, p. 148; *R v. Secretary of State for the Home Department,* 1993), while also playing a role in increasing trust in the process for all impacted parties (Mullan, 2001, p. 148; Douek, 2019).

In the case of indefinite account suspensions—which is a common enforcement action for violations of Twitch's off-platform policy—the impacts on users can be great. As Gillespie explains, "removal from a social media account matters. For a

3. Note that my use of freedom of expression here does not refer solely to constitutional rights to freedom of expression against government limitation. Freedom of expression values can be engaged by private action even where no recognised right is infringed. Corporations are increasingly expected to comply with international human rights law, including social media companies with respect to freedom of expression (*Report of the Special Rapporteur*, 2018). As David Kaye, the former United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression put it, the idea that human rights law applies only to governments and not to companies is "rapidly becoming an archaic way of thinking about the structure of international governance" (Kaye, 2019, pp. 119–290).

user, being suspended or banned … can have real consequences—detaching her from her social circle and loved ones, interrupting her professional life, and impeding her access to other platforms" (2018, p. 176). In the case of Twitch, it can also disrupt or terminate an individuals' primary source of income. Such decisions should not, therefore, be made lightly.

Numerous civil society initiatives, including those of the Santa Clara Principles, Ranking Digital Rights, and the Electronic Frontier Foundation, aim to ensure that content moderation by platforms, and especially large platforms, meet basic standards based on rule of law and due process principles ("Santa Clara Principles," n.d.; Ranking Digital Rights [RDR], 2021; Gebhart, 2019), and a number of platforms have made significant commitments towards meeting these requirements (RDR, 2021). Additionally, governments are looking to legislate forms of procedural accountability for platforms: the UK Online Harms White Paper's approach to legal but harmful content is to require companies to enforce their terms consistently and transparently, and to provide redress mechanisms (UK Department for Digital, 2020). Meanwhile, the EU's proposed *Digital Services Act* prescribes graduated procedural requirements for online intermediaries depending on their type and size, including presenting clear public policies, providing reasons for enforcement actions, and offering appeal mechanisms (Proposal DSA). However, steps to improve procedural accountability come with numerous costs and trade-offs (Stewart, p. 192), and it follows that the procedural design of any given platform decision should be based upon a balancing of the impact of the decision, the risk of error, and other costs. This is certainly true for platforms engaging in policy enforcement, where heightened fairness entails financial costs, delays in responding to potential harm, a lack of flexibility, and possible impacts on user privacy, among others (Douek, 2019). The balance is difficult to strike with respect to content moderation; it is certainly more difficult with respect to policies aimed at off-platform abuse.

This is significant, because while the same concerns that motivate accountability for content moderation are present, including consistency and confidence in enforcement, remedying errors, and user interests, policies aimed at off-platform abuse raise at least three additional concerns. These additional concerns make it significantly more difficult to establish good policy, minimise error in enforcement, minimise the negative impacts of error, and establish public trust and the trust of victims.

a. Defining and verifying off-platform behaviour

The first concern relates to the competence of platforms: platforms typically apply their policies with respect to content that is carried on their services. However, policies such as Twitch's may apply to behaviour that neither manifests as content nor occurs on the platform, such as sexual assault.

The most important ramification of this is the factual uncertainty it generates. Platform content moderation typically does not face the problem of determining whether the behaviour in question actually occurred. Social media platforms often have full visibility over all content on their platforms, and can directly tie any user-generated content to the account holder that created it. Such platforms have access to the whole content, avoiding the possibility that it has access to only selectively chosen or edited content.

By comparison, when dealing with off-platform behaviour, the platform must ascertain what behaviour actually occurred or whether it differed from the reported accounts, as well as any surrounding context. In this, the platform faces considerable evidentiary problems—of the kind more commonly faced by trial courts—due to the inherent reliance on external evidence not within the possession of the platform. Such decisions demand acquiring evidence and determining the veracity of that evidence. This increases the decision-making complexity by requiring both investigation and evidentiary assessment: extra steps to the process of policy enforcement, and one that platforms like Twitch presumably lack expertise. In some cases, they also may lack access to the relevant evidence, or, as in Twitch's case, need cooperation from law enforcement or other platforms to complete an investigation (Twitch, 2021).

Consider, for example, the case of professional fighting game player Nairobi "Nairo" Quezada. Quezada was indefinitely suspended from Twitch in September of 2020 following accusations that he had engaged in a sexual relationship with a minor and after releasing a statement apologising for that conduct (Walker, 2020). However, shortly after, some accounts came to light that called into question the credibility of the accusations (Michael, 2020). Quezada then released a statement in October of 2020 denying the allegations against him, and alleging that he had, in fact, been the victim of sexual assault rather than the perpetrator. He claimed he hadn't understood what had happened to him at the time, and only realised that he was the victim following therapy (Quezada, 2020). He claims to have since filed an appeal with Twitch to restore his account (Quezada, 2021).

The full details of this situation remain obscure, and few credible news reports are available on the subject. Twitch, as per its policy, did not comment on the suspension, nor has it commented on the appeal. There is thus no way of knowing what evidence Twitch relied upon in issuing the suspension. If it was on the basis of Quezada's initial confession, it raises the question of what should happen when such a confession is retracted. Indeed, what evidence would be necessary upon appeal to have his account restored?

Regardless of the truth of either the initial allegations or Quezada's claims, the events surrounding Quezada's suspension reflect the considerable uncertainty introduced by the investigatory process and the reliance on third-party statements and reports. This increases the potential for error in decision-making by introducing new opportunities for error to arise. This is especially the case where, as with Twitch, the standard of proof is the "preponderance of evidence" (Twitch, 2021): a standard used in common law civil litigation (although often called the "balance of probabilities" standard in British English) that requires sufficient evidence that the prohibited behaviour was more likely to have occurred than not. Such a standard has no direct Continental European civil law equivalent as civil law typically requires the conviction of the judge (Schweizer, 2016, p. 218). Regardless, it appears Twitch will apply the preponderance of evidence standard globally. This relatively low standard is balanced in common law civil courts by considerable procedural safeguards, as well as the availability of appeals (Harper et al., 2017).

Indeed, where the potential for error is greater, the need for error reduction and correction mechanisms is greater (Gertmann, 2018). In addition to acquiring outside expertise, error reduction can be aided by ensuring the opportunity for all parties involved to provide evidence and be heard by an impartial decision-maker (Mullan, 2001). Error correction mechanisms for platforms typically involve the opportunity to appeal an adverse decision should the initial decision appear faulty on some basis (Douek, 2019). In the case of Twitch, it currently remains unclear the extent to which the accused can participate in the decision, and while appeals for account suspensions are available (Twitch, n.d.-c), the details of the appeal procedure remain hidden.

Twitch has certainly turned its mind to the extra evidentiary problems created by policies aimed at off-platform misbehaviour: Twitch stated that they have hired an outside law firm with experience investigating workplace and campus sexual assault to assist in investigations (Twitch, 2021). However, both the name of the firm and the investigatory process remain secret. While it may be preferable that Twitch is leveraging existing expertise to mitigate the evidentiary problems, it remains

impossible to know how much such expertise can actually mitigate these concerns without considerably more information. Further, given the heightened possibility of error, it may be reasonable for Twitch to offer robust due process to the accused. This may include a clear process and evidentiary standard for appeal as well as the opportunity to be heard and provide contrary evidence once the initial investigation is complete. Should that process not exist at present, taking steps to implement it would not appear to undermine the goals of Twitch's policy.

In addition to the considerable evidentiary problems such policies raise, policies of this nature also increase the challenge in determining their scope and content, as it may demand that platforms determine what kinds of non-speech behaviour warrant enforcement action. While platforms like Twitch have significant experience enforcing policies against various kinds of speech and content, which may similarly apply to off-platform behaviour in some cases, they presumably have little experience determining appropriate responses to off-platform activity. Twitch has mitigated this problem by limiting enforcement for off-platform activity to behaviour that generally amounts to criminal activity, but it will be a significant issue should they expand this policy to other misbehaviours, as they have suggested is likely (Casey, 2021).

Indeed, sanctions imposed against off-platform behaviour implicate different underlying normative concerns. When platforms moderate content, they can prevent harms from occurring directly through their moderation actions. This is especially true where action is taken against content *ex ante*. For example, with respect to hate speech content or harassing content, if a platform removes that content or reduces its spread, it directly decreases the harms arising from that content by preventing users either from coming into contact with the material or preventing their ongoing exposure to it. While the normative value of content moderation may include those familiar to criminal justice systems, such as rehabilitation (Jahver et al., 2019) and deterrence (Srinivasan et al., 2019), as well as incapacitation in the case of account suspensions, the benefits of immediate harm reduction is sufficient on its own to justify moderation activities. In contrast, policies aimed at off-platform behaviour are less likely to directly mitigate or prevent harms flowing from the incident(s) giving rise to the enforcement action. Platforms' enforcement actions here can only ever be ex post, and they have little direct control over the extent of any harm caused by the precipitating incident. Instead, their justification might lie much closer to criminal justice principles (Cohen, 1981), such as deterrence by warning of penalties for poor behaviour, denunciation by sending a signal to the community about what behaviour is not tolerated, and incapacitation by

preventing an offender from committing future harms on the platform and re-traumatising victims thereon. Indeed, incapacitation appears to be one of Twitch's primary motivations for its off-platform behaviour policy (Twitch, 2021). For victims, enforcement may also serve a retributive or vindicating role (Heydon & Powell, 2016). The potential for differing underlying values behind off-platform policies suggests that it may be preferable that any such policies be developed separately from those aimed at on-platform behaviour and involve considerable community input into what is necessary to create a safe environment.

b. Victims' and complainants' interests

A critical concern raised by policies targeting off-platform abuse is the heightened need to protect the interests of victims and/or complainants. As demonstrated in the summer of 2020, many instances of policy-violating off-platform conduct will involve highly-sensitive and potentially traumatising events, such as sexual assault. In these cases, serving victims' and complainants' interests will require both protecting their privacy and ensuring that, where a report is made directly to Twitch, those reports are received and reviewed through a clear process that promises accountability. This is especially the case since, as the new policy makes clear, those impacted by the off-platform abuse may not be Twitch users and may have little understanding of the platform.

While it does not appear that all investigations will be triggered by direct reports, as previous suspensions were in response to public allegations, Twitch's new policy appears to highlight a direct reporting option (Twitch, n.d.-b), and thus it seems that enforcement actions will often be in response to such reports. Reports may also, presumably, be initiated by those that were not themselves targets of the abuse, creating a potential distinction between victims and complainants.

Where a complaint is made directly to Twitch, fully protecting victims' or complainants' interests in cases of off-platform behaviour can prove difficult, because unlike with decisions made for on-platform conduct, the potential reliance on victim statements and evidence provided by complainants will typically be privacy-invasive. Victims may prefer to remain anonymous when reporting or having their case reported (Powell & Cauchi, 2011). In some cases, victims or other complainants may fear reprisals should the report become known to the perpetrator. At the same time, such anonymity may undermine the possibility of a full investigation, and it naturally inhibits the ability of a decision-maker to seek and receive meaningful input from the alleged policy violator. For these reasons, in criminal cases numerous jurisdictions and institutions make anonymous or confidential

sexual assault reporting available, such reports are typically used to assess criminal patterns and trends, and not to initiate individual investigations, which would require that the victim or complainant be identified (Heydon & Powell, 2016).

The criminal context demonstrates the tension, in some cases, between protecting victims' and complainants' privacy and safety and allowing the alleged policy-violator to participate in the decision-making process. The trade-offs create significant challenges for policy development as platforms that enforce rules against off-platform abuse may have to balance the privacy of the parties involved and fairness to the alleged policy violator. Twitch appears to have prioritised privacy, stating they ensure that all investigations remain confidential and only those involved will be notified of any decision. Unfortunately, it remains unclear what degree of notice or participation will be available for those subject to a decision prior to it being made.

This recalls some literature on campus sexual assault, where the lack of appropriate due process for the accused has received some scrutiny (Gerstmann, 2018; Harper et al., 2017). In that context, it has been argued that robust due process requirements themselves can play an important role in assisting victims in recovering from trauma by providing a responsive and thorough system and increasing the legitimacy of the reporting process (Harper, et al., 2017).

A tension may also arise with respect to providing reasons for decision-making. While public decision-making can increase trust in the system by demonstrating its effective operation, the privacy of the individuals involved may be undermined by any public statement on the matter. While identifying a specific victim in a decision would often be a considerable violation of privacy, even tying an enforcement action to a report for a specific breach may be enough to connect the victim or other involved parties to the incident. At the same time, failure to make public basic reasons for enforcement actions may weaken confidence in the system by making it appear capricious and arbitrary. This may undermine the feeling of safety of users as it may not be clear that the platform is following through on its policy, and it may undermine victims' and complainants' interests and chill reporting by failing to make it clear whether reports were seriously investigated. Ensuring a clear pathway to report policy violations with the promise that complainants will be heard and taken seriously is likely to be critical to ensuring that victims and other interested parties use reporting options and to ensuring that users feel safe on the platforms. Indeed, victims of sexual assault often report that participation, voice, and validation are central to their justice interests when reporting to police (Daly, 2014, p. 387). Victims or other potential complainants may choose not to re-

port if they see no evidence that such interests will be respected.

However, while some tensions exist between due process and privacy with respect to transparency in individual cases, complainants' interests are better served by fully explaining the process through which reports are received, investigated, and evaluated generally. In this, both parties stand to benefit from a well-articulated policy and process for the enforcement of policies against off-platform behaviour by increasing the confidence in the system, and therefore, increasing confidence in safety.

Unfortunately, there remains a dearth of specifics about how a report is handled by Twitch. There is little ground on which to conclude that it has a robust process that provides fairness or certainty to the victim or the accused. It remains unclear even whether all reports will receive a follow-up, let alone how investigations proceed, or what those who report violations, or those who are investigated, can expect from the process. Remedying these defects by more fully explaining the process may go some way in improving accountability.

c. Increased impact of adverse decisions

A third problematic difference between policies aimed at off-platform abuse and policies aimed at on-platform misconduct is the potential for the increased impact of an adverse decision. As Twitch's past enforcement actions make clear, the most common sanction imposed upon those found to have violated policies aimed at off-platform misbehaviour is an indefinite account suspension. Twitch has stated that it can take other actions in response to violations of its off-platform conduct policy, including removing a streamer's Partner status or preventing streamers from engaging in promotional activity (Shear, 2020), although it remains unclear how often such actions are taken. Given the ephemeral nature of content on Twitch, Twitch often engages in account level enforcements for content violations, but indefinite account suspensions appear to be the primary enforcement option for off-platform abuse. Indefinite suspensions have the potential to impact individuals more deleteriously than other content-level actions since they deny access to a vehicle for self-expression altogether and impose significant social and financial costs (Gillespie, 2018, p. 176). This is especially true where one's income may be largely based on access to a platform, as it is in the case of many streamers (Taylor, 2018).

Furthermore, where enforcement actions follow public allegations of harassing or hateful off-platform conduct, such as those made over the summer of 2020, this

enforcement may have the potential to be seen as a form of confirmation of those allegations, or at least to increase their visibility in the media. Either of these outcomes may contribute to public stigma. While such stigma may be justified, the possibility of it nonetheless increases the potential impact of Twitch's decisions. This possibility remains speculative, as it is impossible to know what impact, if any, Twitch's determinations have on public opinion. Certainly, where a public accusation is accompanied by significant corroborating accounts or evidence, or is admitted to by the perpetrator, Twitch's actions will likely have little to no effect. However, where the accusations are denied by the alleged perpetrator, but enforcement action is taken nonetheless, it is possible that individuals could conclude that Twitch conducted proper due diligence with the accusations, and that its findings are therefore credible, even if the basis for its decision is unknown.

For example, after Quezada filed his appeal with Twitch, one prominent fighting game community member said with respect to the possibility of Twitch rescinding the suspension: "[Twitch will] not make a decision without doing everything possible and exhausting all the information. If Twitch comes out and decides to unban Nairo, I guess that they've found enough evidence" (Chen, 2021). Regardless of the degree to which assumptions of this nature are made, it is perhaps likely that account suspensions following high-profile accusations will attract additional media attention. Indeed, it appears that the Twitch account suspensions during the summer of 2020 did attract the attention of news outlets that referred directly to various allegations (BBC News, 2020; Kastrenakes, 2020; Walker, 2020).

As previously discussed, the greater the impact of a determination upon an individual, the greater the need for procedural safeguards. The reality that enforcement actions for off-platform behaviour can only result in permanent or indefinite account-level actions and include the danger of increasing public stigma associated with publicised accusations militates towards increased due process and procedural fairness. This may include the opportunity for impacted parties to be informed of the evidence against them, to be heard, and to provide evidence. It may also include the provision of reasons for a decision to those parties involved and the opportunity to appeal that decision should it reveal errors, bias, or should it have failed to consider all of the relevant evidence (Gerstmann, 2018).

While Twitch allows for appeals in the case of account suspensions (Twitch, n.d.-c), it remains entirely unclear how they're considered, especially for those suspensions that relate to off-platform conduct. It is also not clear what Twitch communicates to the parties involved both before and after the relevant decision in a given appeal. To date, Twitch has not stated whether it provides reasons for enforcement

decisions to either party. Transparency around these issues would be a positive first step.

d. The need for open discussion about off-platform policies

As the relatively cursory discussion above indicates, there are important differences between platform content moderation policies and policies aimed at off-platform behaviour like sexual assault. Indeed, the latter heightens the impact of decisions to all affected parties, attenuating the tension between due process and privacy, while also increasing the possibility of error and calling into question the very purposes of platform policy. As difficult as developing content moderation policy is, it appears that crafting and enforcing policies aimed at off-platform behaviour in an accountable and fair manner is even more difficult. But the difficulty of establishing accountable processes and policies can obviate the need for careful consideration and robust processes. Where platforms do choose to create such policies, it is incumbent upon them to carefully balance the trade-offs and to make themselves accountable to their users.

How to do so remains an unresolved question. The fact that Twitch has narrowed the scope of its policy to apply only to those behaviours that are likely criminal, and the fact that it has hired outside expertise, suggests awareness of these difficulties. Nonetheless, Twitch has to date offered little public accountability in their enforcement of policies against off-platform behaviour, and the recent policy update does not suggest that this will change. The processes and decisions remain opaque. The lack of transparency with respect to the enforcement process undermines confidence in the system, and thus vitiates the feeling of safety that Twitch is attempting to create through its policies and enforcement actions. At a minimum, Twitch should more fully explain how it handles reports, investigates complaints, makes decisions, and considers appeals. Explanations should clarify what those who make a report can expect, who reaches the decision, whether there is an opportunity for an individual subject to a decision to offer evidence or challenge existing evidence, and how appeals will be considered. Depending on how robust these processes are at present, more may need to be done to create a reliable and trustworthy system.

Indeed, as the creation and enforcement of platform policies against off-platform behaviour is a new and little-studied issue, what is needed is a broader public conversation that can inform the creation and enforcement of these policies. Creating policy with little public consultation and no transparency, as Twitch is doing, is a recipe for poor development and implementation. The norms of content modera-

tion have changed enormously over the past decade, much of which is due to the open engagement of users, media, politicians, civil society, and academics (Klonick, 2018, pp. 1648–58). Twitch and other platforms considering or enforcing policies against off-platform behaviour should take the opportunity to begin a broader engagement process that takes into account the interests of victims, users, streamers, and makes use of the expertise of civil society and academia.

A final concern arises should policies against off-platform behaviour be widely adopted by other platforms. Should this happen, significant political and social disenfranchisement could be the result of a finding of harassing, abusive, or other harmful behaviour in any aspect of one's life. As Casey Newton put it jokingly in discussing Twitch's new policy, "[w]hat's next, a social credit score that follows you around the web the way it does the Chinese internet?" (Newton, 2021). Despite the humorous intent, questions like this reflect important questions about the role of non-state actors and the public in sanctioning individual conduct that are beyond the scope of this article. Indeed, answering these questions involves complex interrogation about the role of private enterprise in policing norms of behaviour, the risks to user privacy of records or allegations of behaviour following them across platforms, and the dangers of past behaviours leading to widespread de-platforming. This will not be an issue, however, if policies of this kind remain limited to a small number of platforms.

3. A new frontier of platform policy?

There may be good reason to be skeptical that policies against off-platform behaviour will spread widely beyond Twitch and similar services. As discussed earlier, as a live-streaming service, Twitch may be more similar to a broadcaster than other social media sites like Facebook or Twitter, and thus feel and project a greater responsibility for those it allows to broadcast. Thus it might be reasonable to suspect that policies against off-platform abuse are likely to be limited to platforms that similarly create an asymmetry between a content creator and an audience.

Indeed, YouTube does take enforcement action against some off-platform behaviour by video creators through its Creator Responsibility Policy (YouTube, n.d.). That policy, which does not apply to non-video creators, states that "if we see that a creator's on- and/or off-platform behavior harms our users, community, employees or ecosystem, we may take action to protect the community" (YouTube, n.d.). Examples of off-platform behaviour that may give rise to an enforcement action include participating in sexual abuse or violence, and enforcement can range from being removed from YouTube's recommendations to channel demonetisation to account

suspensions. While YouTube offers virtually no information about its complaint-handling, investigation, or decision-making process, other than to say a "team of experts" is involved (YouTube, n.d.), it has enforced this policy a number of times against creators, including for sexual misconduct and assault (Godwin, 2021; Crowley, 2021). Recently, for example, popular beauty influencer James Charles was "temporarily" removed from YouTube's Partner Program after he admitted to sending sexually explicit messages to sixteen-year old boys (Godwin, 2021).

Beyond video-based platforms, both Patreon and Medium currently have somewhat analogous policies. Crowdfunding platform Patreon's Community Guidelines expressly contemplate enforcement for bullying or harassment in "real-life interactions" (Patreon, n.d.). This policy has been enforced, for example, in banning one creator for revealing private information about another individual on a different platform (Kelly, 2019). Similarly, the Rules of the online publishing platform Medium currently state that it may "consider off-platform action in assessing a Medium account, and restrict access or availability to that account" (Medium, 2019). It is not clear how that policy has been enforced to date.

It's notable that the affordances of both Patreon and Medium similarly create clear asymmetries between users (i.e. between creators and patrons and between authors and readers, respectively). But too much stock should not be placed in a clear distinction between platforms that create such asymmetries and those that do not. While platforms like Facebook and Twitter appear to create functional parities between users, many of their modern affordances, as well as the simple reality that some users have far more reach than others, can create similar asymmetries. Twitter, notably, creates asymmetrical relationships by virtue of allowing one account to follow another account without requiring a reciprocal follow. This can allow for some individuals and organisations to amass large followings without following many accounts themselves (Paul & Friginal, 2019). And Twitter has recently begun to roll out various monetisation options for its users, including the ability for users with large followings to charge for extra content under its Super Follows programme (Koksal, 2021). And while Facebook's 'Friends' relationship has been categorised as creating a symmetrical relationship (Paul & Friginal, 2019), Facebook's Pages, for example, are designed to allow individuals to follow a single individual or business without the reciprocal relationship associated with being Facebook Friends. Facebook has also increasingly implemented content monetisation options, such as offering fan subscriptions, video advertising, and methods to allow fans to support Facebook creators (Facebook, n.d.-b). Further, Facebook has its own direct live-streaming service and Twitch rival, Facebook Gaming, although it ap-

pears to lack similar policies (Facebook, n.d.-b). Affordances across both Twitter and Facebook can thus create similarly asymmetric relationships, and the increasing monetisation options available to creators on these platforms increasingly position creators in similar relationships to those of streamers on Twitch. It may then stand to reason that even these platforms will eventually face similar pressures to those of Twitch in sanctioning off-platform abuse, at least with respect to these aspects of their services.

At present, major social media platforms such as Twitter, Facebook, and Reddit (Twitter, n.d.; Facebook, n.d.-a; Reddit, n.d.) do not currently have policies analogous to Twitch's off-platform abuse policy. But it should be noted that they do have policies against off-platform behaviour in some respects. For example, Facebook and Twitter, amongst others, prevent terrorist organisations or other violent criminal organisations from using the service for any purpose, while Facebook also prohibits any individual involved in mass or multiple murder, human trafficking, or organised crime from using the service. Naturally, enforcement of these policies involves consideration of acts and behaviour that occur beyond the enforcing platform, although they do not raise the same issues as discussed here.

Further, major platforms have begun looking to off-platform conduct in enforcing their existing content policies in order to determine the relevant context in which to understand potential violations. For example, in banning Donald Trump from their platforms, a number of platforms, including Twitter and Facebook, took into account the real-world impacts of Trump's statements both on and off of their platforms, including the violence of 6 January 2021 at the United States Capitol (Kelly, 2021; Twitter 2021). With the pressure to apply the same rules to other world leaders (Morrison, 2021), and growing support for de-platforming based on real-world impacts of harmful speech (Mystal, 2021; Bedingfield, 2021), it is likely that even if these companies do not establish explicit policies against off-platform abuse, investigating off-platform behaviour, and some of those difficulties associated with it, may increasingly become elements of their policy work.

Conclusion

This article has argued that policies against off-platform abuse are a relatively new phenomenon that raises unique challenges in balancing their community-safety objectives with maintaining accountability and fairness. While such policies may be justified on the basis of limiting the potential for future harm and signalling the standards of the community, they also create new challenges in ensuring accountability in platform policy enforcement. These include making factual

determinations, providing safe reporting mechanisms, and protecting victim privacy and safety all while ensuring transparency and fairness to all parties when meting out sanctions with potentially great impact on the sanctioned user. These challenges do not necessarily indicate that such policies should not exist, but rather that extra steps should be taken to balance the goals of the policy with fairness and accountability to users. This article has outlined some possible suggestions in the case of Twitch.

As countries around the world are increasingly attempting to regulate the creation and enforcement of platform content policy through requirements of transparency and due process, they may also want to consider to what extent these regulations do and should apply to policies aimed at off-platform misbehaviour and abuse. For example, the recently proposed *Digital Services Act* in the European Union would require internet intermediaries to ensure that they provide disclosures concerning their content policies and to publish transparency reports about content actions. Web hosts and platforms would have to provide notice and reasons for content removal decisions, while online platforms beyond a size threshold would have to provide internal appeals mechanisms (Proposal DSA). Notably, however, while some of these provisions might apply to the disabling of accounts for behaviour that occurred off of the platform in question (e.g. the requirement to provide reasons) [4], it does not appear that others, such as the requirement to provide an appeals process, would apply to policies aimed at off-platform behaviour. [5]

Similarly, the UK's Online Harms White Paper approach also considers transparency and user redress mechanisms, but is largely focused on ensuring complaint mechanisms for pieces of content and content removals, rather than account actions. The proposed *Online Safety Bill* based on the White Paper makes no mention of off-platform behaviour (Minister of State for Digital and Culture, 2021). In the United States, the bipartisan proposal for increased procedural accountability on interactive computer services, the *PACT Act*, focuses solely on content removal when it mandates transparency reporting and a complaint and appeals mechanism (2021). Policies aimed at off-platform conduct do not appear to be included.

4. The requirement for the provision of reasons in Article 15 applies to situations in which "a provider of hosting services decides to remove or disable access to specific items of information provided by the recipients of the service." Presumably, this could apply for reasons beyond content violations on the platform.

5. The requirement that platforms provide an internal complaint-handling system in Article 17 applies only to "decisions taken by the online platform on the ground that the information provided by the recipients [of the service] is illegal content or incompatible with its terms and conditions." This would not appear to apply to actions taken for off-platform behaviour as the provision specifies it applies only to actions based on the information provided by the recipients of the service.

If governments are concerned with ensuring not only that platforms remove harmful content, but that they protect the interests of users in continuing to use central platforms for modern discourse, they may want to consider the role of platforms in disabling access to individuals based upon conduct that took place off of the respective platform. To ensure effective regulation, it's critical that academics, civil society, platforms, and governments begin a wider discussion of how and when policies against off-platform behaviour should be developed and enforced. Platforms like Twitch should begin this process by being transparent about their processes and by seeking public and civil society input on their policy development.

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Ariel Katz and Jack Enman-Beech for early discussions on this topic. The author would also like to thank the editors and reviewers of *Internet Policy Review* for their comments, including evelyn douek, Andrew Zolides, Frédéric Dubois, and Balázs Bodó, whose insights greatly improved this article. Any mistakes remain with the author.

References

Alexa. (n.d.). *Top sites*. Retrieved 6 April 2021, from https://www.alexa.com/topsites

Asarch, A. (2019, August 13). Twitch's continuous struggle with moderation shines a light on platform's faults. *Newsweek*. https://www.newsweek.com/twitch-over-party-ninja-stream-porn-moderation-faults-1454148

Baker v. Canada (Minister of Citizenship and Immigration), 2 SCR 817 (Supreme Court of Canada 1999).

Balkin, J. M. (2018). Free speech is a triangle. *Columbia Law Review*, *118*, 2011–2056.

B.B.C. News. (2020, June 25). Twitch starts banning users over abuse. *BBC News*. https://www.bbc.com/news/newsbeat-53179288

Bedingfield, W. (2021). *Deplatforming works, but it's not enough to fix Facebook and Twitter. Wired*. https://www.wired.co.uk/article/deplatforming-parler-bans-qanon

Benvenisti, E. (2014). *The law of global governance*. Hague Academy of International Law.

Bloch-Wehba, H. (2019). Global platform governance: Private power in the shadow of the state. *SMU Law Review*, *72*(1), 27–80. https://scholar.smu.edu/smulr/vol72/iss1/9/

Bunting, M. (2018). From editorial obligation to procedural accountability: Policy approaches to online content in the era of information intermediaries. *Journal of Cyber Policy*, *3*(2), 165–186. http

s://doi.org/10.1080/23738871.2018.1519030

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 1–13. https://doi.org/10.1177/2056305120936636

Chen, J. (2021). *Tuesday 10.10—Justin Wong Talks Fatherhood And Becoming A Fighting Game God [Video]*. *YouTube*. https://www.youtube.com/watch?v=iS48c4qPUGk

Cohen, S. A. (1981). An introduction to the theory, justifications and modern manifestations of criminal punishment. *McGill Law Journal*, *27*, 73–91.

Colombo, C. (2020). Twitch sparks outrage after Onision is quietly unbanned. *Dexerto*. https://www.dexerto.com/entertainment/twitch-sparks-outrage-after-onision-is-quietly-unbanned-1430801/

Crowley, J. (2021, January 22). Why exactly has controversial YouTuber Onision been demonetized? *Newsweek*. https://www.newsweek.com/onision-demonetized-youtube-1563434

Daly, K. (2014). Reconceptualizing sexual victimization and justice. In I. Vanfraechem, A. Pemberton, & F. M. Ndahinda (Eds.), *Justice for victims: Perspectives on rights, transition and reconciliation* (pp. 378–395). Routledge. https://doi.org/10.4324/9780203094532-30

D'Anastasio, C. (2020, June 26). Twitch confronts its role in streaming's #MeToo reckoning. *Wired*. https://www.wired.com/story/twitch-streaming-metoo-reckoning-sexual-misconduct-allegations/

Douek, E. (2019). *Verified accountability: Self-regulation of content moderation as an answer to the special problems of speech regulation* (Paper No. 1903; Aegis Series). https://www.hoover.org/research/verified-accountability

Douek, E. (2021). Governing online speech: From "posts-as-trumps" to proportionality & probability. *Columbia Law Review*, *121*(3), 759–834. https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/

Facebook. (n.d.-a). *Community standards*. https://www.facebook.com/communitystandards/

Facebook. (n.d.-b). *Gaming community guidelines*. Facebook Gaming. https://www.facebook.com/fbgaminghome/creators/gaming-community-guidelines

Facebook. (n.d.-c). *How can I make money on Facebook?* Facebook for Business. https://www.facebook.com/business/learn/lessons/how-make-money-facebook

Fuller, L. L. (2000). *The morality of law*. Yale University Press.

Galiz-Rowe, T. (2020, August 2). Popular super smash bros. Streamer ZeRo has been banned from twitch. *GameSpot*. https://www.gamespot.com/articles/popular-super-smash-bros-streamer-zero-has-been-ba/1100-6480106/

Gebhart, G. (2019, June). *Who has your back? Censorship edition 2019*. Electronic Frontier Foundation. https://www.eff.org/wp/who-has-your-back-2019

Geigner, T. (2019, September 20). Content moderation at scale especially doesn't work when you hide all the rules. *Techdirt*. https://www.techdirt.com/articles/20190918/10465243018/content-moderation-scale-especially-doesnt-work-when-you-hide-all-rules.shtml

Gerstmann, E. (2018). *Campus sexual assault: Constitutional rights and fundamental fairness*. Cambridge University Press. https://doi.org/10.1017/9781108671255

Gilbert, B. (2020, September 20). Ninja just signed a multi-year contract that keeps him exclusive to Amazon-owned Twitch. *Business Insider*. https://www.businessinsider.com/ninja-signs-multi-year-exclusivity-contract-with-amazon-twitch-2020-9

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Godwin, C. (2021, April 20). James Charles: YouTube temporarily demonetises beauty influencer. *BBC News*. https://www.bbc.com/news/world-us-canada-56811134

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Grayson, N. (2020, June 26). Twitch finally starts banning streamers accused of sexual abuse. *Kotaku*. https://kotaku.com/twitch-finally-starts-banning-streamers-accused-of-sexu-1844164469

Hall, C. (2020, July 3). Evo Online canceled following accusations of sexual abuse against CEO. *Polygon*. https://www.polygon.com/2020/7/3/21312536/evo-online-canceled-joey-cuellar-mr-wizard-sexual-abuse

Harper, S., Maskaly, J., Kirkner, A., & Lorenz, K. (2017). Enhancing Title IX due process standards in campus sexual assault adjudication: Considering the roles of distributive, procedural, and restorative justice. *Journal of School Violence*, *16*(3), 302–316. https://doi.org/10.1080/15388220.2017.1318578

Hernandez, P. (2020, June 25). Twitch starts banning streamers over sexual abuse allegations. *Polygon*. https://www.polygon.com/2020/6/25/21302983/twitch-sexual-abuse-assault-harassment-bans

Heydon, G., & Powell, A. (2016). Written-response interview protocols: An innovative approach to confidential reporting and victim interviewing in sexual assault investigations. *Policing and Society*, *28*(6), 631–646. https://doi.org/10.1080/10439463.2016.1187146

Hsu, T., & Lutz, E. (2020, August 1). More than 1,000 companies boycotted Facebook. Did it work? *The New York Times*. https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html

Iqbal, M. (2021, March 29). *Twitch revenue and usage statistics*. Business of Apps. https://www.businessofapps.com/data/twitch-statistics/

Jahver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter?: User behavior after content removal explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, *3*. https://doi.org/10.1145/3359252

Kastrenakes, J. (2020, June 25). Twitch reckons with sexual assault as it begins permanently suspending streamers. *The Verge*. https://www.theverge.com/2020/6/25/21303185/twitch-sexual-harassment-assault-permanent-bans-streamers

Kaye, D. (2019). *Speech Police: The global struggle to govern the internet*. Columbia Global Reports.

Kelly, M. (2019, November 26). Controversial YouTuber banned from Patreon after alleged doxxing. *The Verge*. https://www.theverge.com/2019/11/26/20984785/onision-doxxing-patreon-deplatformed-twitter-youtube

Kelly, M. (2021, January 7). Facebook bans Trump 'indefinitely'. *The Verge*. https://www.theverge.com/2021/1/7/22218725/facebook-trump-ban-extended-capitol-riot-insurrection-block

Klonick, K. (2018). The New governors: The People, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Lorenz, T., & K, B. (2020, June 23). Dozens of women in gaming speak out about sexism and harassment. *New York Times*. https://www.nytimes.com/2020/06/23/style/women-gaming-streaming-harassment-sexism-twitch.html

Martens, T. (2020, July 9). Resignations and reckoning: Game industry's existential quest for a more inclusive space. *Los Angeles Times*. https://www.latimes.com/entertainment-arts/story/2020-07-09/game-industry-reckoning-sexual-harassment-ubisoft-chris-avellone

Mathews v. Eldridge, 424 U.S, (US Supreme Court 1976).

Medium. (2019, November 25). *Medium rules*. Medium Policy. https://policy.medium.com/medium-rules-30e5502c4eb4

Michael, C. (2020, September 15). CaptainZack allegedly lied about taking 'hush money' from Nairo. *Dot Esports*. https://dotesports.com/fgc/news/captainzack-allegedly-lied-about-taking-hush-money-from-nairo

Draft online safety bill, no. CP 405, 405 (2021). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

Morrison, S. (2021, January 20). Facebook and Twitter made special world leader rules for Trump. What happens now? *Vox*. https://www.vox.com/recode/22233450/trump-twitter-facebook-ban-world-leader-rules-exception

Mullan, D. J. (2001). *Administrative law*. Irwin Law.

Mystal, E. (2021, January 22). Twitter and Facebook just proved that deplatforming works. *The Nation*. https://www.thenation.com/article/politics/twitter-facebook-free-speech/

Newton, C. (2021, April 7). Twitch calls in the cavalry. *Platformer*. https://www.platformer.news/p/twitch-calls-in-the-cavalry

Patreon. (n.d.). *Community Guidelines*. Patreon. https://www.patreon.com/policy/guidelines

Paul, J. Z., & Friginal, E. (2019). The effects of symmetric and asymmetric social networks on second language communication. *Computer Assisted Language Learning*, *32*(5–6), 587–618. https://doi.org/10.1080/09588221.2018.1527364

Powell, M. B., & Cauchi, R. (2011). Victims' perceptions of a new model of sexual assault investigation adopted by Victoria Police. *Police Practice and Research*, *14*(3), 228–241. https://doi.org/10.1080/15614263.2011.641376

Proposal DSA. (2020). *Proposal for a Regulation of the European Parliament and of the council on a single market for digital services (Digital Services Act) and amending Directive 2000/31/EC COM/2020/825 final*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0825&from=en

Quezada, N. (2020, October 28). My statement [Medium Post]. *Nairo*. https://nairoby.medium.com/my-statement-9a091682fff3

Quezada, N. (2021, March 3). On the topic of Twitch, I do want to be clear that I'm not looking for a handout or anything [Tweet]. *Twitter*. https://twitter.com/NairoMK/status/1367222559826202630

R v. Secretary of State for the Home Department, ex p. Doody, No. 8 (UK House of Lords 1993).

Ranking Digital Rights. (2021). *2020 Ranking Digital Rights corporate accountability index*. https://rankingdigitalrights.org/index2020

Reddit. (n.d.). *Reddit Content Policy*. Reddit Inc. https://www.redditinc.com/policies/content-policy

Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. UNHRC, 38th Sess, UN Doc A/HRC/38/35. (2018).

Reynolds, K., Subašić, E. & Tindall. (2015). The problem of behaviour change: From social norms to an ingroup focus. *Social and Personality Psychology Compass*, *9*(1), 45–55. https://doi.org/10.1111/spc3.12155

Schreier, J. (2020, June 24). Video game industry rocked by outpouring of sexual misconduct allegations. *Bloomberg*. https://www.bloomberg.com/news/articles/2020-06-24/video-game-industry-rocked-by-outpouring-of-sexual-misconduct-allegations

Schweizer, M. (2016). The civil standard of proof—What is it, actually? *The International Journal of Evidence & Proof*, *20*(3), 217–234. https://doi.org/10.1177/1365712716645227

Shanley, P. (2019, September 30). Twitch CEO on war for streaming talent, transparent moderation plans. *The Hollywood Reporter*. https://www.hollywoodreporter.com/news/twitchs-emmett-shear-streaming-talent-wars-moderation-plans-1244171

Shear, E. (2020, June 22). There's been a lot of important conversation happening over the previous couple days, and I've heard your voices [Tweet]. *Twitter*. https://twitter.com/eshear/status/1275234049070526464

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community. *Proceedings of the ACM on Human-Computer Interaction*, *3*. https://doi.org/10.1145/3359265

Stewart, R. B. (2016). Global standards for national societies. In S. Cassese (Ed.), *Research handbook on global administrative law* (pp. 175–195).

Sunstein, C. (1996). On the expressive function of law. *University of Pennsylvania Law Review*, *144*, 2021–2053. https://doi.org/10.2307/3312647

Suzor, N. P. (2019). *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge University Press. https://doi.org/10.1017/9781108666428

Taylor, T. L. (2018). *Watch me play: Twitch and the rise of game live streaming*. Princeton University Press.

The Santa Clara principles on transparency and accountability in content moderation. (n.d.). https://santaclaraprinciples.org/

Twitch. (n.d.-b). *Off-service conduct policy*. Twitch Legal. https://www.twitch.tv/p/en/legal/community-guidelines/off-service-conduct-policy/

Twitch. (n.d.-a). *Community guidelines: Hateful conduct and harassment*. Twitch Legal. https://www.twitch.tv/p/en/legal/community-guidelines/harassment/

Twitch. (n.d.-c). *About account enforcements and chat bans*. Twitch Help. https://help.twitch.tv/s/article/about-account-suspensions-dmca-suspensions-and-chat-bans?language=en_US

Twitch. (n.d.-d). *Frequently asked questions*. Twitch Partnership Program. https://www.twitch.tv/p/en/partners/faq/

Twitch. (2018, February 8). Twitch community guidelines updates [Blog post]. *Twitch Blog*. https://blog.twitch.tv/en/2018/02/08/twitch-community-guidelines-updates-f2e82d87ae58

Twitch. (2020a, December 9). Introducing our new hateful conduct & harassment policy [Blog post]. *Twitch Blog*. https://blog.twitch.tv/en/2020/12/09/introducing-our-new-hateful-conduct-harassment-policy/

Twitch. (2020b, June 24). An update to our community [Blog post]. *Twitch Blog*. https://blog.twitch.tv/en/2020/06/24/an-update-to-our-community/

Twitch. (2020c). *Transparency Report 2020* [Report]. https://www.twitch.tv/p/en/legal/transparency-report/

Twitch. (2021, April 7). Our plan for addressing severe off-service misconduct [Blog post]. *Twitch Blog*. https://blog.twitch.tv/en/2021/04/07/our-plan-for-addressing-severe-off-service-misconduct

Twitter. (n.d.). *Rules and policies*. Twitter Help. https://help.twitter.com/en/rules-and-policies#general-policies

Twitter. (2021, January 8). Permanent suspension of @realDonaldTrump [Blog post]. *Twitter Blog*. https://blog.twitter.com/en_us/topics/company/2020/suspension.html

U.K. Department for Digital, Culture, Media & Sport & U.K. Home Office. (2020). *Consultation outcome: Online harms white paper: Full government response to the consultation*. https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response

U.S. PACT Act To Require Transparency, Accountability, and Protections for Consumers Online, United States Congress, 117 (2021).

Walker, I. (2020, September 11). Twitch bans Smash champion after he admits to having sex with minor. *Kotaku*. https://kotaku.com/twitch-bans-smash-champion-after-he-admits-to-having-se-1845027272

Wiltshire, A. (2019, November 11). What does it take to make a living on Twitch? *PC Gamer*. https://www.pcgamer.com/what-does-it-take-to-make-a-living-on-twitch/

YouTube. (n.d.). *Creator responsibility*. Google Support. https://support.google.com/youtube/answer/7650329?hl=en

Zolides, A. (2020). Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society*, 1–19. https://doi.org/10.1177/1461444820942483

Published by

in cooperation with