YouTube content moderation

YouTube enforces many of its content policies algorithmically. The decisions of how to handle user–generated content have massive impact on the culture of today's online platforms [4]. On YouTube, machine learning systems classify whether each uploaded video contains content that is protected by an existing copyright [8], and whether it violates YouTube's "advertiser–friendly content guidelines" [18]. The content guidelines, for example, state that to earn ad revenue, videos must not feature controversial issues, dangerous substances, harmful acts, inappropriate language, or sexually suggestive content. YouTube's advertisers do not want their ads run on videos with this content, so if a video does not meet the content deadlines, it is *demonetized*, where YouTube does not show ads on the video and the content creator receives no income for it.

However, these demonetization algorithms have made highly–publicized errors. For example, YouTube began labeling videos uploaded by members of the LGBTQ community as "sexually explicit" [3], demonetizing them. When video titles included words like "transgender", they were demonetized; when the content creators removed "transgender" but left other aspects of the video as–is, the same videos were monetized normally [1].

Several issues here are worth dealing with separately. First, a system that categorizes a video so close to the margin of one decision versus another that the title makes a substantive difference ought to involve a human at that level of uncertainty. But that's beside the point. Sex and gender are different. While discussions — ranging from informal to legal — have conflated these ideas in the past [16], analyses and certainly systems that instantiate the notion that gender and sex reference the same concepts betray a failure to grow and learn in step with society. Finally, as people increasingly turn to and rely on YouTube to be a venue to earn a living — not unlike a public square — YouTube's algorithmic classification system effectively deprived this marginalized community of income for discussing the very issues of marginality that they face. YouTube eventually apologized, stating "our system sometimes make mistakes in understanding context and nuances" [17].

These issues also emerge in the opposite direction, so to speak. The same algorithms that mischaracterized discussions about *gender* (specifically, being trans) as *sexual* and consequently inappropriate for advertising failed to flag inappropriate content disguised in a large number of children's cartoons. Fraudulent videos of Peppa Pig being tortured at a dentist's office were left alone, and videos of cartoon characters assaulting and killing each other were passed over by the algorithm, in some cases being included in YouTube Kids, a subset of YouTube which offers to source content appropriate specifically for children. This failure to identify troubling content wrapped in ostensibly child–friendly animation again led to a refinement of YouTube's policies and algorithms.

It can be useful to think about YouTube's content moderation algorithms as analogous to the class of street–level bureaucrats who monitor and interact with street performers in offline urban contexts. Street performance, or *busking*, is usually monitored by police [13]. Many cities have laws which restrict busking in certain contexts. The police must identify when they should enforce laws strictly and when they should take a more permissive stance: the details of enforcement of those laws is necessarily left to police officers, affording them substantial latitude. The public record is full of instances of police ranging in behavior from aggressively managing to enjoying and engaging in the performance themselves [5]. As performance by nature often pushes the bounds of expectations and street performance in particular is inherently experimental [6], police have to be flexible about the application of their powers and make reasonable decisions in new circumstances. The challenge is to characterize the novelty of the situation and reflect on the decision they're being called to make. More importantly, they must make these decisions in the moment, applying the implications of that decision to a constantly–updating intuition about the effective policy they're creating through their selective enforcement of laws.

We can think of YouTube's monetization algorithm as akin to a sort of police force that encounters hundreds of thousands of new performances every day and is expected to navigate those situations appropriately. The challenge is that this algorithmic police force trains and updates its policies in batches: it is executing on today's performances based on yesterday's data. Yesterday, transgender people weren't speaking publicly, in a venue accessible all over the world, about deeply personal aspects of their lives in the ways that they now do. The algorithm is always behind the curve: at best, it gets feedback or negative rewards only after it has executed its decision, in this case when YouTubers appeal or gather media attention. By contrast, police reflexively construct an interpretation of the situation as soon as they encounter it, rather than merely match on learned patterns.

This case highlights a shortcoming with a commonly offered solution to these kinds of problems, that more training data would eliminate errors of this nature: culture always shifts. We argue that that more data will never allow us to anticipate and prevent these kinds of errors. Experimentation is often the point of performance and art, and certainly part of the nature of public performance art like YouTube [2, 11]. Transgender identity became a topic of discussion; cartoon characters became murderers. In statistical terms, we might say that the data in this system is a nonstationary process: the distribution (of words,

topics, and meanings) changes, sometimes abruptly. YouTube was, at one time, a uniquely empowering space for members of the transgender community to be candid and to explore their shared experiences [12, 14], but the culture has grown in ways that YouTube and its content classification algorithms have not. Reinforcement rewards and new training data can help an algorithm reconfigure its decision boundary, but even deep learning only gets this new training information *after* it has already made decisions, sometimes hundreds of thousands of decisions — *effecting* a policy deterring transgender YouTubers from discussing their gender identity, or risk being demonetized for discussing content the system erroneously classifies as "sexual".

The effects of bad street–level algorithms are farther–reaching than the immediate cases that are mishandled. Lipsky points out that people begin to work around street–level bureaucracies when they become unreliable and untrustworthy, or when the public decide that they cannot hope for bureaucrats to make favorable decisions [10]. We see this phenomenon unfolding on YouTube: as their demonetization algorithms progressively mishandle YouTubers' needs [7, 15], more and more creators have begun to circumvent YouTube entirely, encouraging audiences to support them through sites such as Patreon [9], and demonetizing their own channels.

REFERENCES

[1] James Bridle. 2017. Something is wrong on the internet. https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

[2] Jan Cohen-Cruz. 1998. *Radical street performance: An international anthology.* Cambridge Univ Press.

[3] Megan Farokhmanesh. 2018. YouTube is still restricting and demonetizing LGBT videos — and adding anti-LGBT ads to some. https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-demonetization-ads-algorithm

[4] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[5] Jan Grill. 2011. From street busking in Switzerland to meat factories in the UK: a comparative study of two Roma migration networks from Slovakia. *Global Connections and emerging inequalities in Europe: perspectives on poverty and transnational migration* (2011), 77–102.

[6] Sally Harrison-Pepper. 1990. *Drawing a circle in the square: street performing in New York's Washington Square Park.* Univ Pr of Mississippi.

[7] Erik Kain. 2017. YouTube Wants Content Creators To Appeal Demonetization, But It's Not Always That Easy. https://www.forbes.com/sites/erikkain/2017/09/18/adpocalypse-2017-heres-what-you-need-to-know-about-youtubes-demonetization-troubles/#31b00bb96c26

[8] Eugene C Kim. 2007. YouTube: Testing the safe harbors of digital copyright law. *S. Cal. Interdisc. LJ* 17 (2007), 139.

[9] Roope Leppänen et al. 2017. The State of Video Hosting Websites from the Perspective of Content Creators. (2017).

[10] Michael Lipsky. 1983. *Street-Level Bureaucracy: The Dilemmas of the Individual in Public Service.* Russell Sage Foundation.

[11] Bim Mason. 1992. *Street theatre and other outdoor performance.* Taylor & Francis.

[12] Matthew G. O'Neill. 2014. *Transgender Youth and YouTube Videos: Self-Representation and Five Identifiable Trans Youth Narratives.* Palgrave Macmillan UK, London, 34–45. https://doi.org/10.1057/9781137383556_3

[13] Julia Quilter and Luke McNamara. 2015. Long May the Buskers Carry on Busking: Street Music and the Law in Melbourne and Sydney. *Melb. UL Rev.* 39 (2015), 539.

[14] Tobias Raun. 2016. *Out online: Trans self-representation and community building on YouTube.* Routledge.

[15] Jocelyn Summers. 2018. What Professors Dont Tell You About Social Media. (2018).

[16] Francisco Valdes. 1994. *Queers, Sissies, Dykes, and Tomboys: Deconstructing the Conflation of" Sex,"" Gender," and" Sexual Orientation" in Euro-American Law and Society.* Ph.D. Dissertation. HeinOnline.

[17] Johanna Wright. 2018. Restricted Mode: How it works and what we can do better. https://youtube-creators.googleblog.com/2017/03/restricted-mode-how-it-works-and-what.html

[18] LLC YouTube. 2010. YouTube Community Guidelines.