Check for updates

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June: 1–15
© The Author(s) 2020
DOI: 10.1177/2053951719897945
journals.sagepub.com/home/bds
SAGE

Algorithmic content moderation: Technical and political challenges in the automation of platform governance

Robert Gorwa[1], Reuben Binns[2] and Christian Katzenbach[3]

Abstract

As government pressure on major technology companies builds, both firms and legislators are searching for technical solutions to difficult platform governance puzzles such as hate speech and misinformation. Automated hash-matching and predictive machine learning tools – what we define here as *algorithmic moderation systems* – are increasingly being deployed to conduct content moderation at scale by major platforms for user-generated content such as Facebook, YouTube and Twitter. This article provides an accessible technical primer on how algorithmic moderation works; examines some of the existing automated tools used by major platforms to handle copyright infringement, terrorism and toxic speech; and identifies key political and ethical issues for these systems as the reliance on them grows. Recent events suggest that algorithmic moderation has become necessary to manage growing public expectations for increased platform responsibility, safety and security on the global stage; however, as we demonstrate, these systems remain opaque, unaccountable and poorly understood. Despite the potential promise of algorithms or 'AI', we show that even 'well optimized' moderation systems could exacerbate, rather than relieve, many existing problems with content policy as enacted by platforms for three main reasons: automated moderation threatens to (a) further increase opacity, making a famously non-transparent set of practices even more difficult to understand or audit, (b) further complicate outstanding issues of fairness and justice in large-scale sociotechnical systems and (c) re-obscure the fundamentally political nature of speech decisions being executed at scale.

Keywords

Platform governance, content moderation, algorithms, artificial intelligence, toxic speech, copyright

This article is a part of special theme on The Turn to AI. To see a full list of all articles in this special theme, please click here: https://journals.sagepub.com/page/bds/collections/theturntoai

Introduction

On 15 March 2019, a terrorist strapped a camera to his chest, began a Facebook live stream, and entered the Al Noor Mosque in Christchurch, New Zealand with an assault rifle, murdering more than 50 people. The video, initially seen only by a few hundred Facebook users, was quickly reported to Facebook by the authorities and taken down – but not before copies were made and re-posted on internet messaging boards. Within hours, hundreds of thousands of versions of the video (some altered with watermarks or other edits) were being re-uploaded to Facebook, as well as to YouTube and Twitter.

A few days after the attack, Facebook representatives stated that in the first 24 hours, versions of the

[1]Department of Politics and International Relations, University of Oxford, Oxford, UK
[1]UK Information Commissioner's Office (ICO); Department of Computer Science, University of Oxford, Oxford, UK
[3]Alexander von Humboldt Institut für Internet und Gesellschaft (HIIG), Berlin, Germany

Corresponding author:
Robert Gorwa, Department of Politics and International Relations, University of Oxford, Manor Road Building, Manor Road, Oxford OX1 3UQ, UK.
Email: robert.gorwa@politics.ox.ac.uk

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

video had been uploaded at least 1.5 million times, and some 80% of those videos, around 1.2 million, were blocked automatically before they could be uploaded (Sonderby, 2019). The tragic Christchurch incident thus became the highest profile test yet for the members of an organisation called the Global Internet Forum to Counter Terrorism (GIFCT), a group created by Facebook, Google, Twitter and Microsoft as part of a commitment to increase industry collaboration in the European Commission's code of conduct to combat illegal online hate speech (Huszti-Orban, 2017). As members of GIFCT, the four companies share best practices for developing their automated systems and operate a secretive 'hash database' of terrorist content, where digital fingerprints of illicit content (images, video, audio and text) are shared. Within hours of the Christchurch attack, Facebook had uploaded hashes of about 800 different versions of the shooter's video (Sonderby, 2019). In an incredible technical and computational feat, every single video and image uploaded by ordinary Facebook users (as well as YouTube and Twitter users) would now be hashed and checked against the database. If it matched, it would be blocked.

Turning to AI for moderation at scale

The more they grow, the less the mega-platforms of today resemble their social network predecessors. Where bulletin boards and forums were once meticulously managed by dedicated administrators who formed part of the community, platform companies operate at a scale that has led them away from traditional practices of community moderation (Lampe and Resnick, 2004) and towards what has been termed 'commercial content moderation' or 'platform moderation' (Roberts, 2018). Since the 2016 US election, there has been a substantial increase in public attention paid to content moderation issues – now widely seen as a crucial element of major tech and platform policy debates – as well as broader academic awareness of the problems with the platform governance status quo (Gorwa, 2019b). A growing body of scholarship has documented the multiple challenges with commercial content moderation as enacted by platforms today, ranging from labour concerns (about the taxing working conditions and mental health challenges faced by moderators, many of whom are outsourced contractors in the Global South); democratic legitimacy concerns (about global speech rules being set by a relatively homogenous group of Silicon Valley elites); and process concerns about the overall lack of transparency and accountability (see Gillespie, 2018; Kaye, 2019; Roberts, 2018; Suzor et al., 2019).

An important but still relatively under-examined feature of the rapidly evolving content moderation ecosystem is the use of technologies grouped under the generic term 'artificial intelligence' (AI). Amidst significant technical advances in machine learning (and the enormous amount of hoopla that has followed them), automated tools are not only being increasingly deployed to fill important moderation functions, but are actively heralded as the force that will somehow save moderation from its existential problems. As government pressure on major technology companies builds, both companies and legislators seem to hope that technical solutions to difficult content governance puzzles can be found. Under recent regulatory measures like the German NetzDG or the EU Code of Conduct on Hate Speech, platforms are increasingly being bound to a very short time window for content takedowns that effectively necessitates their use of automated systems to detect illegal or otherwise problematic material proactively and at scale.

These shifts should be scrutinised carefully and critically. It is clear that the use of various statistical techniques labelled as 'AI' has provided a major opportunity for firms to appease governance stakeholders while also presenting self-serving and unrealistic narratives about their technological prowess: Facebook CEO Mark Zuckerberg notably invoked 'AI' as the future solution to Facebook's current political problems dozens of times during Congressional testimony in 2018. But it is not all empty hype: the statistics trotted out in press materials and in company transparency reporting illustrate the significant role that automation is already playing in enforcing content policy. For example: after a major public controversy, Facebook improved its Myanmar-language hate-speech classifiers, leading to a 39% increase in takedowns from automated flags in only six months; YouTube now reports that '98% of the videos removed for violent extremism are flagged by machine-learning algorithms,' and Twitter recently stated that it has taken down hundreds of thousands of accounts that try to spread terrorist propaganda, with some '93% consist[ing] of accounts flagged by internal, proprietary spam-fighting tools' (GIFCT, 2019).

Incidents like Christchurch clearly show that automated moderation systems have become necessary to manage growing public expectations for increased platform responsibility, safety and security; however, as has been repeatedly pointed out by civil society groups, these systems remain opaque, unaccountable and poorly understood. The goal of this article is to provide an accessible primer on how automated moderation works; examine some existing automated systems used by major platforms to handle copyright infringement, terrorism and toxic speech; and to

outline some major issues that these systems present as they continue to be developed and put into practice. Our main contention is that automated moderation systems, while often painted with the same broad 'AI' brush in public discourse, have varying affordances, and thus differing policy impact. In particular, we indentify an important distinction between hash-matching and predictive systems, with the potential harms to users (e.g on free expression grounds) varying considerably depending on the implementation. We argue that despite the promise of automated techniques, and the increasing pressure placed by governments on firms to deploy those techniques, what we call *algorithmic moderation* has the potential to exacerbate, rather than relieve, several key problems with content policy. In particular, some implementations of algorithmic moderation threaten to (a) decrease decisional transparency (making a famously non-transparent set of practices even more difficult to understand or audit), (b) complicate outstanding issues of justice (how certain viewpoints, groups, or types of speech are privileged), and (c) obscure or depoliticise the complex politics that underlie the practices of contemporary platform moderation.

What is algorithmic moderation?

Following Grimmelmann's (2015: 6) broad definition of content moderation as the 'governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse,' forms of content moderation have existed as long as group-based online communication has. In Grimmelmann's understanding, moderation includes not only the administrators or moderators with power to remove content or exclude users, but also the design decisions that organise how the members of a community engage with one another. Historically, automated systems appear to enter the community moderation toolbox when scale problems make manual curation or intervention unfeasible. On USENET and other bulletin boards, the growing proliferation of spam led some users to experiment with automated filters, such as the 'Automated Retroactive Minimal Moderation' system that was accidentally unleashed on USENET in 1993 (Brunton, 2013: 41). Later, as large scale peer-production communities like Wikipedia grew rapidly, automated 'bot' moderators enforced Wikipedia's rules, fought vandalism and monitored articles slated for deletion, playing a key role in the moderation process (Geiger, 2014). Past work has helpfully explored how systems of moderation deployed across a variety of communities integrate automated tools, from Wikipedia (Geiger, 2011) to Twitch and Reddit (Seering et al., 2019). We focus on 'commercial' content moderation as outlined by Roberts (2018) – distinct from the more 'artisinal' and contextual forms moderation exhibited within other online communities (Caplan, 2018) – and examine specifically the role of automation within the content moderation practices of Facebook, Twitter, YouTube and other major platforms for user-generated content.

We define *algorithmic commercial content moderation* (referred to as *algorithmic moderation* for brevity in the following sections) as systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown). This is a narrower definition of moderation than espoused by Grimmelmann (2015) and others: we investigate only systems that make decisions about content and accounts (exerting 'hard moderation') and exclude the litany of 'soft' moderation systems (recommender systems, norms, design decisions, architectures) that form the base of Grimmelmann's moderation taxonomy. This article is by no means a comprehensive overview of algorithmic moderation; rather, it is a general outline that we hope will be a useful guide for more focused research in the area. It is necessarily limited by our reliance on publicly available reporting and primary source material, such as company press releases, white papers, technical reports and investigative journalism. The platform companies are cagey about the details of how they conduct algorithmic moderation, and there are almost certainly systems that have not been reported or that we may have missed. We hope that future work can provide in-depth analysis into specific systems and specific companies, drawing on leaks, insider interviews, or other forms of data.

A primer on the main technologies involved in algorithmic moderation

Algorithmic content moderation involves a range of techniques from statistics and computer science, which vary in complexity and effectiveness. They all aim to identify, match, predict, or classify some piece of content (e.g. text, audio, image or video) on the basis of its exact properties or general features. However, there are some major differences in the techniques used depending on the kind of matching or classification required, and the types of data considered. One major distinction can be made between systems that aim to match content ('is this file depicting the same image as that file?'), and those that aim to classify or predict content as belonging to one of several categories ('is this file spam? Is this text hate speech?'; see Table 1).

Table 1. Simple typology of moderation, with examples.

	Identification: match	**Identification:** prediction
Consequence Hard (blocking, removal)	PhotoDNA	Perspective API
Consequence Soft (flagging, downranking)	Youtube content ID	Twitter quality filter

Matching

Systems for matching content typically involve 'hashing', i.e. the process of transforming a known example of a piece of content into a 'hash' – a string of data meant to uniquely identify the underlying content. Hashes are useful because they are easy to compute, and typically smaller in size than the underlying content, so it is easy to compare any given hash against a large table of existing hashes to see if it matches any of them. This is computationally much cheaper than comparing every bit for each pair. They are also generally expected to be (relatively) unique, such that it is very unlikely that two different pieces of content will share the same hash (what cryptographers call a hash 'collision').

Secure cryptographic hash functions aim to create hashes that appear to be random, giving away no clues about the content from which they are derived. It should also be very difficult to construct an input whose hash value will collide with that of another. Cryptographic hash functions are useful for checking the integrity of a piece of data or code to make sure that no unauthorised modifications have been made. For instance, if a software vendor publishes a hash of the software's installation file, and the user downloads the software from somewhere where it may have been modified, the user can check the integrity by computing the hash locally and comparing it to the vendor's.

However, cryptographic hash functions are not very useful for content moderation, because they are sensitive to any changes in the underlying content, such that a minor modification (e.g. changing the colour of one pixel in an image) will result in a completely different hash value. This means that while they could be used to moderate content that is exactly the same as previously identified content, they can be easily circumvented by minor perturbations (such as adding watermarks or borders, cropping, flipping, or any other modification). For this reason, other forms of non-cryptographic hashing are generally used. These alternative techniques, including fuzzy hashing, locality-sensitive hashing and perceptual hashing, aim to compute not exact matches, but rather 'homologies' – similarities between two inputs (Datar et al., 2004). For instance, an image that has had 1% of its pixels changed is more similar to the original than one in which 99% of pixels have changed.

This violates a principle of security in cryptographic hash functions, because the hash value reveals something about the underlying input. Similar inputs share similar hash values (a property called 'smoothness'), which means that matches are not designed to be exact and unique. (For this reason, some security experts regret that the word 'hash' is used in this way, believing it should be reserved for the cryptographic variety.).[1] This also means that someone could possibly guess the underlying input or construct a new input which would collide and be mistakenly identified as a previous input. There is a risk that a malicious user might 'poison' a hash database by deliberately modifying a piece of content to have a hash value matching that of a piece of benign content (e.g. an image which appears to be depicting banned material but actually has a hash value identical to that of a benign image such as the Coca-Cola logo), thus causing the benign content to be mistakenly flagged.[2] But the benefit is robustness; small changes to a piece of content will only result in a correspondingly small change in hash similarity, so non-exact matches can be found.

Of the non-cryptographic hashing techniques, the most suitable and robust for content moderation is 'perceptual hashing' (p-hash; Niu and Jiao, 2008). Perceptual hashing involves fingerprinting certain perceptually salient features of content, such as corners in images or hertz-frequency over time in audio. By picking up on these features, rather than the strings of bits used in cryptographic hashes, perceptual hashes can be more robust to changes that are irrelevant to how humans perceive the content. They aim to capture distinctive patterns that are relevant to semantic categories (e.g. shapes, colours, sounds), such that content remains identifiable even after perturbation. For instance, feature detection algorithms in computer vision ensure that even if an image is rotated or scaled, the same shapes can be identified when they are upside down or enlarged (Harris and Stephens, 1988). Given the task of constructing perceptual features, perceptual hashing benefits from many of the same techniques used for feature detection in deep learning prediction tasks. While it is unclear exactly how the GIFCT's Shared Industry Hash Database (or SIHD, discussed below) works, it is likely that it uses some form of perceptual hashing. Facebook's recently published PDQ and TMK + PDQF hashing technologies also fall into this category (Davis & Rosen, 2019).

Classification

The techniques discussed above all involve matching a newly uploaded piece of content against an existing

database of curated examples. Classification, by contrast, assesses newly uploaded content that has no corresponding previous version in a database; rather, the aim is to put new content into one of a number of categories. For example, while the GIFCT is primarily focused on matching through the SIHD, it also states that it is engaging in 'content detection and classification techniques using machine learning' (GIFCT, 2019).

Modern classification tools typically involve machine learning, i.e. the automatic induction of statistical patterns from data. One of the main branches of machine learning is supervised learning, in which models are trained to predict outcomes based on labelled instances (e.g. 'offensive'/'not offensive'). But historically, content classification has been based on manually coded features. Much of the early work focused on text. Automated systems for hate speech, personal attacks, 'toxicity' and related phenomena developed in response to advances in natural language processing (see Schmidt and Wiegand, 2017). Early approaches involved automatically screening comments for blacklisted keywords, where the blacklists are manually curated. Various collaborative open source blacklists exist, predominantly focused on the problem of enumerating curse words in multiple languages.[3] The simplest use of such blacklists is to find exact string matches.

Such approaches have clear limitations. Maintaining effective and up-to-date blacklists can be difficult, especially as norms develop and as users are able to reverse-engineer the blacklist and avoid exact string matches accordingly. They also risk over-blocking in cases in which the word may be acceptable in context. Some very simple examples, such as those in which profane words happen to be contained within benign ones (e.g. 'ASS-ociation') might be caught using regular expressions (a standard coding tool for matching patterns in text). However, beyond this, such systems inevitably miss the wide range of contextual clues used to determine whether or not a particular word is acceptable in a given sentence or comment. Early attempts to improve on keyword-based blacklists focused on manually crafted features (such as evaluated regular expressions or word lists), and detecting linguistic features, such as imperative statements ('get lost!') or particular noun phrases preceded by pronouns ('you jerk'), and detecting whitelisted 'polite' words, which might reduce how hateful the text is perceived (Spertus, 1997). Other approaches rely on domain-specific ontologies, e.g. of terms implicated in LGBT hate speech, to attempt to infer whether a statement might be hateful given certain domain knowledge (such as the gender of the recipient; Dinakar et al., 2012).

Recent work, along with much of natural language classification research, has focused on machine learning approaches. These generally involve training language classifiers on a large corpus of texts, which have been manually annotated by human reviewers according to some operationalisation of a concept like offence, abuse, or hate speech. A range of machine learning algorithms have been applied to a variety of feature sets. Simple approaches to feature selection include 'bag of words', which simply treats all of the words in a sentence as features, ignoring order and grammar (Chen et al., 2012). More complex approaches involve word embeddings, which represent the position of a word in relation to all the other words that usually appear around it. Semantically similar words therefore have similar positions (Mikolov et al., 2013).

Matching and classification have some important differences; while matching requires a manual process of collating and curating individual examples of the content to be matched (e.g. particular terrorist images), classification involves inducing generalisations about features of many examples from a given category into which unknown examples may be classified (e.g. terrorist images in general). But there are also systems which blur the lines between the two. For instance, a series of photos taken milliseconds apart might be something that a matching system ought to class as similar, even though the underlying images are different and therefore technically not matches. Facial recognition technologies may serve the dual purpose of inducing patterns from many faces and matching particular faces belonging to the same person. In these cases, the distinction between identity-matching and classification is a matter of degree.

An algorithmic moderation typology

The specific fashion in which these matching or predictive systems are deployed depends greatly on a variety of factors, including the type of community, the type of content it must deal with, and the expectations placed upon the platform by various governance stakeholders. Those expectations substantially affect not only the design of the system itself, but also the ways in which that system is used to then act upon and potentially moderate content. Following Gillespie's (2018) observation that content moderation is one of the core commodities provided by a platform – enabling it to serve advertiser, as well as user needs, and therefore be a viable business – algorithmic moderation is one of the central mechanisms through which that commodity can be realised in practice. Automated tools are used by platforms to police content across a host of issue areas at scale, including terrorism, graphic violence, toxic

Table 2. Publicly reported algorithmic moderation systems deployed by major platforms, by issue area.

	Terrorism	Violence	Toxic speech	Copyright	Child abuse	Sexual content	Spam & automated accounts
Facebook	Shared Industry Hash Database (SIHD), ISIS/Al-Qaeda classifier	Community standards classifiers	Community standards classifiers	Rights manager	PhotoDNA	Non-consensual intimate image classifier, nudity detection	Immune system
Instagram			Comment filter	Rights manager	PhotoDNA		Comment filter, false account detection
YouTube	SIHD, Community Guidelines (CG) ML classifiers	CG ML Classifiers	CG ML Classifiers	Content ID	Content safety API, PhotoDNA	CG ML Classifiers	CG ML Classifiers
Twitter	SIHD		Quality filter		PhotoDNA	Sexual content interstitial	Proactive Tweet and account detection, quality filter
WhatsApp					PhotoDNA		Modified immune system

API: application programming interface.[4]

speech (hate speech, harassment and bullying), sexual content, child abuse and spam/fake account detection (See Table 2).

Once content has been identified as a match, or is predicted to fall into a category of content that violates a platform's rules, there are several possible outcomes. The two most common are flagging and deletion. In the former case, content is placed in either a regular queue, indistinguishable from user-flagged content, or in a priority queue where it will be seen faster, or by specific 'expert' moderators (Caplan, 2018). In the latter case, content is removed outright or prevented from being uploaded in the first place. A host of other specific decisions are also available, depending on the desired governance outcome and the preferences of the governance stakeholders that have informed the design of system (Gorwa, 2019b).

For example, although intermediary liability provisions in the United States prevent YouTube from being held legally accountable for hosting copyright-infringing material, the company faced growing threat from legal challenges to the status quo like the suit filed by Viacom in 2007 (see the section on copyright below). YouTube's solution, an automated system called Content ID, is highly tilted towards the preferences of the copyright holders it intends to pacify. After uploading their audio or video content, copyright holders have the ability to select whether they wish to take down or receive a portion of the advertising revenue from content that matches the system's hashes (Soha and McDowell, 2016). The individuals uploading the video have little recourse, and reversing these decisions is extremely difficult (Perel and Elkin-Koren, 2015). In another example, after years of being critiqued for being insufficiently responsive to harassment, Twitter developed a 'Quality Filter' which tries to predict whether content may be low-quality, spammy, or automated (Leong, 2016). Due to Twitter's strong First Amendment stance on freedom of expression, and its general hesitation to significantly moderate, it designed the Quality Filter not to remove content, but rather to render it less visible (e.g. muting notifications for tagged users).

The location of human discretion within these systems is also deeply dependent on socio-political factors (Table 3). Civil society and academic human rights advocates have argued that fully automated decision-making systems that do not include a human-in-the-loop are dangerous (Duarte et al., 2017: 6). Facebook, when announcing its participation in the GIFCT hash database, insisted that matched content would not be blocked automatically, but rather flagged for further review (Facebook Newsroom, 2016). However, in its transparency reporting the company has also stated that it was able to automatically block millions of pieces of ISIS and Al-Qaeda content before it was

Table 3. A breakdown of notable algorithmic moderation systems.

Actor	System	Issue areas	Target content	Core tech	Human role
YouTube	Content ID	Copyright	Audio, video	Hash-matching	Trusted partners upload copyrighted content
Google Jigsaw	Perspective API	Hate speech	Text	Prediction (NLP)	Label training data and set parameters for predictive model
Twitter	Quality filter	Spam, harassment	Text, accounts	Prediction (NLP)	Label training data and set parameters for predictive model
Facebook	Toxic speech classifiers	Hate speech, bullying	Text	Prediction (NLP, deep-learning)	Label training data and set parameters for predictive model; make takedown decisions based on flags
GIFTC	Shared-industry hash database	Terrorism	Images, video	Hash-matching	Trusted partners suggest content, firms find/add content to database
Microsoft	PhotoDNA	Child safety	Images, video	Hash-matching	Civil society groups add content to database

Note that these systems often can be set to exert either hard or soft moderation based on the context, but we categorise them here based on their point of emphasis.

uploaded, indicating that it believes that it is acceptable – given the constant pressure from the US, EU and other major Western governments to combat radicalisation by those groups – to remove the human from the decision loop in certain cases.[5] (These reports do not provide information about whether content that supports more geopolitically complex and contested organisations, e.g. Kurdish or Kashmiri separatist groups, is automatically blocked by the hash database).

In the following section, we dive a bit more deeply into three main areas where algorithmic moderation has been deployed in the past decade: copyright, terrorist content and toxic speech.

Algorithmic moderation in practice

Copyright

Copyright has historically been one of the first, if not the first, domain where strong economic interests demanded technologies to match and classify online content. The aftermath of Napster and the file-sharing controversies in the early 2000s coincides with the rise of social media and platforms to host and share creative and cultural works, and video-sharing platforms, most notably YouTube, became a key target for industry lobbies and other rightsholders seeking to curb the unlicensed distribution of their content. Shortly after Google's acquisition of YouTube in 2006, the major media company Viacom sued the platform for massive copyright infringement by its users. While this case was finally settled in 2013, the long-running litigation increased pressure on platforms to monitor and police the content on their site even before recieving formal notice.

Anticipating the growing political and economic pressure, in 2006 YouTube started to experiment with content monitoring systems that were formally and procedurally independent of the obligatory notice-and-takedown-process in 2006 (Holland et al., 2016). These efforts evolved over time into the Content ID system that YouTube has now been running and iterating for more than a decade. Much like other platforms, YouTube remains secretive about the specific technological implementation of its proprietary algorithmic moderation systems.[6] Nonetheless, some characteristics can be discussed based on publicly available material. Content ID works along roughly the matching logics described above, but is unique in that it allows copyright holders to upload the material that will be (a) searched against existing content on YouTube and (b) added to a hash database and used to detect new uploads of that content. In the copyright context, the goal of deploying automatic systems is not only to find identical files but also to identify different instances and performances of cultural works that may be protected by copyright. These systems are not only able to find multiple uploads of a music video, but also of recordings of live performances of that song. Through perceptual hashing, the resulting fingerprints aim to reflect characteristics of the audio or video content: 'each note in a song could be represented by the presence or absence of specific frequency values; the volume of a particular note [...] by some amplitude at the frequency corresponding to that musical note' (Duarte et al., 2017: 14). As a consequence, the fingerprint is both more robust to technical modifications and better equipped to detect new variations and interpretations of a piece of content.

A key concern in the deployment of automated moderation technologies in the context of copyright is

systematic overblocking. While Content ID and other systems may improve from a technical standpoint, enhancing their ability to create quality fingerprints and then accurately detect those fingerprints, it does not necessarily mean that they become more adept at evaluating *actual copyright infringement*. Copyright law allows third-parties to create excerpts or use protected content through 'fair use', which varies across jurisdictions but creates important exemptions for educational purposes, parody and other important contexts (Patel, 2013). Burk and Cohen (2001) argue that the contextual factors needed to assess fair use standards cannot be programmed into automated systems – an argument supported by recent empirical studies of automated copyright enforcement that report substantial overblocking of content on video sharing platforms (Bar-Ziv and Elkin-Koren, 2018; Erickson and Kretschmer, 2018; Urban et al., 2017). Human and institutional oversight is thus essential. Although YouTube does provide remedy procedures for users that wish to challenge take-downs, these are slow and resource-intensive for the challenger, who is often at a major disadvantage compared to rightsholders. As Soha and McDowell (2016: 6) argue, users who have their content removed or demonetised have virtually no recourse, noting that 'even in clear cases of fair use, it can often require months as well as legal help and expert knowledge of copyright law to achieve a successful fair use claim.' Facebook (in 2016) and Instagram (in 2018) have followed YouTube by deploying the Rights Management platform, which features similar functionality to Content ID (Keef and Ben-Kereth, 2016).

Despite a large number of controversial takedown decisions,[7] regulatory and stakeholder pressure on the platforms has incentivised them to provide an easier path towards takedown than to thorough *ex-ante* investigations, even if specific cases might benefit from copyright exemptions. This imbalance could effectively be enshrined in much-discussed legislation like the EU Copyright Directive, which could affect the monitoring obligations from platforms for content uploaded by users, and lead to the greater deployment of matching systems at the point of upload.[8]

Terrorism

In December 2015, after preparatory meetings held in 2014 and 2015, the European Commission officially announced the creation of the EU Internet Forum, which brought together EU officials together with representatives from Google, Facebook, Twitter and Microsoft (Gorwa 2019a). After only six months and two meetings (that are publicly known; the entire process was deeply secretive, and notably excluded civil society; see Fiedler, 2016), the members of the Internet Forum announced the EU Code of Conduct on Countering Illegal Hate Speech Online, committing the firms to a wide-ranging set of principles, including the takedown of hateful speech within 24 hours under platform terms of service and the intensification of 'cooperation between themselves and other platforms and social media companies to enhance best practice sharing' (European Commission, 2016: 3). To comply with that commitment, the four firms announced the creation of the GIFCT in 2017. The organisation, which remains highly secretive, has a board made of 'senior representatives from the four founding companies' and publishes little about its operations (GIFCT, 2019). However, the organisation has been particularly focused on the improvement of automated systems to remove extremist images, videos and text.

The GIFCT maintains the SIHD of terrorist content, which is now used by 13 different companies, including Instagram, LinkedIn, Oath, Reddit and Snap. (It is not clear whether these new companies can add content to the database, or merely use the hashes uploaded by the founding members.) Each firm 'appl[ies] its own policies and definitions of terrorist content when deciding whether to remove content when a match to a shared hash is found' (Facebook Newsroom, 2017), suggesting that each hash is uploaded with a set of metadata, likely including the firm that uploaded it, the specific type of content/terrorist group, and information about the specific incident. Early GIFCT press releases emphasised that 'matching content will not be automatically removed' (Facebook Newsroom, 2016). However, statements following the Christchurch shooting revealed that hundreds of thousands of Facebook uploads were automatically blocked based on the SIHD (Rosen, 2019). This could mean that firms choose to selectively and automatically upload-filter only certain hashes based on target incident or group metadata (e.g. 'Christchurch Shooting'; 'Al-Qaeda'). It is also unclear how each firm decides whether to use hashes from other firms (given their differing definitions of terrorism – see Llanso, 2016), and whether hashes placed in the database by other firms (e.g. in the case of Christchurch, a total of approximately 800 different hashes) are approved by human reviewers in each firm before being used to remove other instances of matching content.

The database appears to build on Microsoft's PhotoDNA technology, which is used by many platforms (including Facebook) to match uploads against the National Center for Missing and Exploited Children's hash database of child abuse imagery, as well as Google's open-source equivalent, the 'Content Safety API'. In November 2018, Facebook's policy

leadership stated that they had also begun adding audio and text hashes to the database (Bikert and Fishman, 2018). This appears to suggest that effectively every single piece of content uploaded by Facebook users – not just images and videos, but also ordinary status updates – is now being hashed and compared against the database for potential matches.

In addition, firms have begun using a host of proactive detection techniques to flag and remove potential terrorist content that makes it through the SIHD-filter. Facebook has begun using machine learning algorithms to try and surface content supporting certain groups, such as ISIS and al-Qaeda (Bikert and Fishman, 2018). These tools, trained on a corpus of training data, create a predictive score that tries to estimate how likely a post is to violate Facebook's terrorism policies. Depending on that score, that post will be flagged for human moderation, with higher scores placing them higher in the queue for priority review by 'specialized reviewers' (Bikert and Fishman, 2018). Interestingly, this system keeps a human in the loop for the final takedown decision in most, but not all instances, with the Facebook officials writing that 'in some cases, we will automatically remove posts when the tool indicates with very high confidence that the post contains support for terrorism. We still rely on specialized reviewers to evaluate most posts, and only immediately remove posts when the tool's confidence level is high enough that its "decision" indicates it will be more accurate than our human reviewers' (Bikert and Fishman, 2018, n.p.). Despite the strange assertion that the tool can have a confidence level higher than the specialised reviewers one would expect to provide ground truth, these systems have led to a massive increase in the amount of terrorism-related takedowns. In the first quarter of 2018, Facebook reported that it had removed 1.9 million pieces of ISIS and al-Qaeda content; in the second quarter, it took down more than 7 million (Bikert and Fishman, 2018).

It remains unknown if or how these systems are audited, and what kind of false positive rates are considered typical or acceptable for such tasks. Part of the challenge is that platform policies are deeply context dependent (Caplan, 2018): for instance, Facebook has traditionally allowed terrorist imagery if it is being used by a reputable news organisation or in order to express disapproval or condemnation of a group. Scholarship has documented the important role that 'witness videos' uploaded to platforms like YouTube have played in Syria and other contemporary conflicts (Smit et al., 2017). However, automated counter-terror content detection systems removed thousands of videos that had been uploaded to YouTube by civil society groups and activists to document atrocities conducted during the Syrian Civil War (Browne, 2017). Machine learning systems are famously poor at making such difficult context-dependent judgements, and, much like copyright, there is widespread civil society concern that such automated systems lead to over-blocking and curb important forms of expression (Duarte et al., 2017).

Toxic speech

Any platform that enables communication between users faces problems of potentially offensive speech, personal attacks and abuse that could harm users, distort conversation or drive certain contributors away. Recent discourse has cast these problems in terms of 'toxicity' of comments and 'conversational health' (Gadde and Gasca, 2018). These terms tend to be used as umbrellas for various concepts, including hate speech, offence, profanity, personal attacks, sleights, defamatory claims, bullying and harassment (despite differentiation between these concepts being necessary to build models for corresponding sub-tasks; Waseem et al., 2017). By training machine learning algorithms on large corpora of texts manually labelled for toxicity, they aim to create automatic classification systems to flag 'toxic' comments.

In 2017, Jigsaw, a Google/Alphabet subsidiary focused on 'global security challenges' announced a new project called Perspective. Perspective is an application programming interface (API) with a stated aim to make it 'easier to host better conversations'. According to the project description, a platform could use Perspective to receive a score which predicts the 'impact a comment might have on a conversation', which could be used 'to give realtime feedback to commenters or help moderators do their job'.[9] Similar efforts have been pursued by other platforms, including Twitter, and Disqus, a third-party comment plugin provider.

In the past few years, Facebook has responded to growing pressure around hate speech (especially from EU member states) by developing classifiers trained to predict whether text may constitute hate speech, and based on that score, flag it for human review. These efforts, which began in certain languages such as English and Portuguese, have now been scaled to others, including policy-vital languages like Burmese (Rosen, 2018). This work is increasingly drawing upon the research being carried out by Facebook's AI Research division, which has helped the classifiers cope with translation and other challenges. Instagram has also developed toxic speech classifiers (building upon Facebook's 'DeepText' platform) to identify comments for bullying and harassment, taking a different approach from Facebook by offering an opt-out filter that users can use to hide comments rather than referring them for moderation (Thompson, 2017).

YouTube has notably also moderated toxic speech as present in uploaded content, training machine learning classifiers that seek to predict the incidence of hate, harassment, as well as vulgar 'swearing and inappropriate language' in a video in order to de-monetise it and prevent advertisers from having their content embarrassingly paired with anything that could damage their brand (Internet Creators Guild, 2016).

Civil society has done the most to document the challenges facing the algorithmic moderation of toxic speech. The clearest problem is that language is incredibly complicated, personal and context dependent: even words that are widely accepted to be slurs may be used by members of a group to reclaim certain terms (York and McSherry, 2019). Insufficient context-awareness can lead crude classifiers to flag content for adjudication by moderators who usually do not have the context required to tell whether the speaker is a member of the group that the 'hate speech' is being directed against.

The technical limitations of publicly available toxic speech classifiers have become apparent in cases like Perspective, which has received significant academic attention and critique. The model that Perspective initially used was based on a research collaboration between Google and the Wikimedia Foundation (Wulczyn et al., 2017). The training data was derived from Wikipedia talk pages, where volunteer encyclopaedia moderators discuss content and debate edit decisions on particular pages. This data, consisting of individual comments, was labelled by workers on the micro-task site Crowdflower, according to a set of questions relating to 'personal attacks' and 'harassment'. While the models described in the original research article report high test accuracy, after the release of the publicly available API many observers on social media constructed and shared examples of over- and under-zealous toxicity predictions. For example, the single-term comment 'Arabs' was classed as 63% toxic, while the phrase 'I love führer' was only 3% toxic (Sinders, 2017). In addition to these manually constructed counter-examples, other researchers have demonstrated how to automatically construct adversarial examples (Hosseini et al., 2017). In the summer of 2019, Instagram announced that it would begin deploying a Perspective-like system to try and nudge users away from posting offensive comments (Mosseri, 2019).

Three political issues: Transparency, fairness and depoliticisation

Critical conversations about algorithmic moderation systems often emphasise the technical challenges that these systems face now and in the future (Li and Williams, 2018). In particular, there is outsized concern about overblocking: it is commonly (and correctly) pointed out that it is very difficult for predictive classifiers to make difficult, contextual decisions on slippery concepts like 'hate speech,' and that automated systems at scale are likely to make hundreds, if not thousands, of incorrect decisions on a daily basis. However, new matching techniques effectively search based upon known, manually curated ground truth, and as a result are less likely to lead to 'incorrect' takedowns. Algorithmic moderation is here to stay, now mandated either implicitly or explicitly in both legislation and informal platform regulation, such as codes of conduct (Gorwa, 2019a).

The use of automated techniques can potentially help firms remove illegal content more quickly and effectively, and firms will continue investing heavily down the moderation 'stack', optimising their systems in an effort to improve their precision and recall. But improving, for example, the quality of Facebook's PDQ photo-matching algorithm so that it is better able to find content that infringes on Facebook's Community Standards (Facebook, 2019) does not change outstanding accountability concerns about how those standards are created. Nor does it alleviate a host of ethical and political moderation problems that have the potential to be exacerbated by an increase in automated techniques.

These problems exist at the intersection of the recent literatures on content moderation, platform regulation, and fairness, transparency, and accountability in machine learning. In classic critiques of automated decision-making (and the recent scholarship that has refined those critiques), three arguments are often advanced that are particularly applicable to algorithmic moderation: decisional transparency, justice and de-politicisation.

Decisional transparency

Content moderation has long been a famously opaque and secretive process (Gillespie, 2018; Roberts, 2019; Suzor, 2019). Years of pressure by researchers, journalists and activists have recently led to notable efforts by companies like Facebook to make their moderation practices more transparent, such as the long-overdue publication of the 'Community Standards' that outline the bounds of acceptable behaviour on the site, the instigation of a formal appeals process, and an effort to create some kind of independent oversight mechanism into their policies (Kadri and Klonick, 2019). However, the rapid push towards algorithmic moderation in the past few years threatens to reverse much of this progress.

Much like in other areas, such as risk scoring (Oswald et al., 2018), automated systems for moderation introduce significant complexity that poses a challenge from an auditing perspective. A common critique of automated decision making is the potential lack of transparency, especially when claims of commercial intellectual property are used to deflect responsibility (Burrell, 2016). In content moderation, it will become significantly more difficult to decipher the dynamics of takedowns (and potential human rights harms) around some policy issues when the initial flagging decisions were made by automated systems, and the specific criteria by which those initial decisions were made remain unknown. From a user perspective, there is little transparency around whether (or to what extent) an automated decision factored into a takedown. The specific functionalities of these systems are left intentionally vague, and the databases of prohibited content remain closed off to all – including, worryingly, trusted third-party auditors and vetted researchers. As Llansó (2019) describes, there also appears to be scope creep around certain algorithmic moderation systems, where firms are experimenting with adding new functionalities – such as the recent announcement by the GIFCT companies that they were experimenting with sharing blocklists of URLs – without any apparent oversight and effectively zero transparency. This is not to say that transparency is any sort of panacea, either in general or when it comes to algorithmic systems (Ananny and Crawford, 2018; Gorwa and Garton Ash, 2019), but minimum standards of decisional transparency are essential to allow both ordinary users and critical experts to understand the patterns of governance within which they are embedded (Santa Clara Principles on Transparency and Accountability in Content Moderation, 2018). How are important decisions about free expression being made and enforced?

Justice

Recent years have seen substantial discussion about the potential for algorithmic decision-making systems to have unfair or discriminatory impacts on different groups, such as protected classes under anti-discrimination law (e.g. gender, race, religion, disability; see Barocas and Selbst, 2016). While such work has focused on fairness/discrimination in the distribution of outcomes in economic, welfare or justice domains (Berk et al., 2018; Chouldechova et al., 2018; Hardt et al., 2016), some of it has parallels in the area of algorithmic content moderation. Content classifiers in general, whether used for recommendation, ranking, or blocking, may be more or less favourable to content associated with gender, race and other protected categories (Blodgett et al., 2016; Ekstrand et al, 2018; Zehlike et al., 2017), and thus entrench forms of representational harm against such groups (Barocas et al., 2017; Binns, 2018). Even a perfectly 'accurate' toxic speech classifier will have unequal impacts on different populations because it will inevitably have to privilege certain formalisations of offence above others, disproportionately blocking (or allowing) content produced by (or targeted at) certain groups. To the extent that automatically blocking or allowing certain types of content can be seen as distributive harms or benefits conferred on protected groups, or privileging more or less 'deserving' individuals, they could be cast within the paradigm of algorithmic 'fairness' (Binns et al., 2017). For example, hate speech classifiers designed to detect violations of a platform's guidelines could be disproportionally flagging language used by a certain social group, thus making that group's expression more likely to be removed.

However, such framings have their own issues, and attempting to shoehorn problems of algorithmic content moderation into an algorithmic fairness framing may be misguided. Fairness critiques often miss broader structural issues, and risk being blind to wider patterns of systemic harm (Hoffmann, 2019). To take one well-known example, Facebook designed its policies at one point so that they would be 'fair' and ostensibly treat all racial categories equally. However, due to the positionality of Facebook's content policy employees, these rules were written in a way that failed to account for the massively disproportional impact of racial discrimination across much of the Global North, as well as the intersectional nature of disadvantage, thus failing to carve out hate speech protections for certain 'sub-categories' such as 'black children' (Angwin, 2017). While automated systems may help find and quickly take down more dehumanising racist attacks, they also risk entrenching unjust rules in a rapid, global, and inscrutable fashion.

De-politicisation

A third, and related concern, is about de-pollicisation, and the visibility of content moderation as a political issue. Thanks to the important work of journalists, researchers and civil society, moderation has become a site of political contestation in many countries. Governments, civil society groups and publics are demanding more say in how the speech rules governing our digital lives are created and deployed (Suzor, 2019). But what if this attention dissipates once automated systems are in place that can conveniently render unpleasant speech largely invisible? What if these moderation systems achieve their overarching aim by becoming an infrastructure that smoothly operates in the background, that is taken for granted, and that obscures its inner

workings, becoming hidden, much like the labour and practices of content moderation used to be (Roberts, 2019)?

Algorithmic moderation has already introduced a level of obscurity and complexity into the inner workings of content decisions made around issues of economic or political importance, such as copyright and terrorism. For example, companies like Facebook now can boast of proactively removing 99.6% of terrorist propaganda (Facebook, 2019), legitimising both their technical expertise and their role as a gatekeeper protecting a 'community'. However, this elides the hugely political question of who exactly is considered a terrorist group (Facebook only reports takedown numbers for Al-Qaeda and ISIS related content, and not for other types of terrorist content), and therefore what kind of data is trained and labelled for the classifiers, as well as the open question of the technical issues that these systems necessarily face. As Elish and boyd (2018) describe, much discursive work is done by the 'magic' of 'AI', which often combines 'technological inscrutability with a glossing over of technological limitations' (p. 66). When contrasted with the status quo – that is, the evaluation and removal of content by a 'fallible' human moderator – automation is associated with a 'scientific' impartiality that is inherently attractive to platform companies, one that additionally lets them keep their decisions 'non-negotiable' (Crawford, 2016: 67) and hidden from view. As algorithmic moderation becomes more seamlessly integrated into user's day-to-day online experience, human rights advocates and researchers must continue to challenge both the discourse and reality of the use of automated decision making in moderation, and not let firms hide behind the veil of black-boxed complexity as they seek to disengage from important content policy discussions.

Conclusion

The underlying sociopolitical questions explored through our short cases – where to draw lines between acceptable and unacceptable speech, which types of speech should or should not require a human in the loop – will never disappear, but they may be strategically buried. Talk of 'fixing' hate speech or misinformation by applying a dash of AI or algorithms might be easily dismissed by social scientists as merely perpetuating the myth of technological solutionism, but one should also consider the only-partially performative nature of such claims. Companies and governments *do* invest heavily in automated systems that are then implemented across a variety of contexts and quickly become taken for granted. Have you recently thought about spam? The technical answer to the non-trivial and context-dependent question of an email message being unsolicited or not seems now to be optimised to a level that produces convenience for the user and renders the underlying social questions invisible. What if this type of pattern is repeated in the context of speech classification? A perfectly 'accurate' algorithmic moderation system would re-obscure not only the complex moderation bureaucracies that keep platforms functioning, but also threaten to depoliticise the fundamentally political nature of speech rules being executed by potentially unjust software at scale.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Gorwa would like to thank the Social Science and Humanities Research Council of Canada and the Canadian Centennial Scholarship fund for supporting his studies.

ORCID iD

Robert Gorwa https://orcid.org/0000-0002-4891-5053

Notes

1. Facebook's former chief security officer Alex Stamos has said: 'I hate the term 'hash' because it implies cryptographic properties that are orthogonal to how these fingerprints work' https://twitter.com/alexstamos/status/928050441799196672
2. See e.g. https://towardsdatascience.com/black-box-attacks-on-perceptual-image-hashes-with-gans-cc1be11f277
3. See e.g. Shutterstock's 'List of Dirty, Naughty, Obscene and Otherwise Bad Words': https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words
4. The list of hyperlinked sources (press releases, blogs, white papers, etc) for this table is available at https://gorwa.co.uk/files/bds_table1.pdf
5. See https://transparency.facebook.com/community-standards-enforcement, Q4 2018 (https://perma.cc/A6PJ-E6GU?type=image)
6. In addition to the officially published material, a leaked 'YouTube Content ID Handbook' is circulating online that had apparently been prepared by YouTube for rightsholders. See the last available version, updated Q2/2014 at https://scribd.com/document/351431229/YouTube-Content-ID-Handbook.
7. See Ericksson and Kretschmer (2018) for empirical evidence on takedown decisions, and the Electronic Frontier Foundation's 'Take Down Hall of Shame' for some well-known anecdotal examples https://www.eff.org/de/takedowns.

8. See the excellent documentation of the policy process at the website of CREaTE at the University of Glasgow: https://www.create.ac.uk/policy-responses/eu-copyright-reform/
9. https://github.com/conversationai/perspectiveapi

References

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.

Angwin J and Grassegger H (2017) Facebook's Secret Censorship Rules Protect White Men, But Not Black Children. *ProPublica*. Available at: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Barocas S and Selbst A (2016) Big data's disparate impact. *California Law Review* 104(3): 671. https://doi.org/10.15779/Z38BG31

Barocas S, Crawford K, Shapiro A, et al. (2017) The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In: *9th Annual Conference of the Special Interest Group for Computing, Information and Society*, Philadelphia, PA. Retrieved from: http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf

Bar-Ziv S and Elkin-Koren N (2018) Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown. *Connecticut Law Review* 50: 339.

Berk R, Heidari H, Jabbari S, et al. (2018) Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 0049124118782533. https://doi.org/10.1177/0049124118782533

Bikert M and Fishman B (2018) Hard questions: What are we doing to stay ahead of terrorists? Facebook Newsroom. Available at: https://perma.cc/YRD5-P5HU

Binns R (2018) Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research* 81: 149–159.

Binns R, Veale M, Van Kleek M, et al. (2017) Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: *International conference on social informatics*, 2017, pp. 405–415. Berlin: Springer.

Blodgett SLL, Green B and O'Connor (2016) Demographic dialectal variation in social media: A case study of African-American English. In: *EMNLP 2016: Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, 1–5 November 2016.

Browne M (2017) YouTube removes videos showing atrocities in Syria. *New York Times*, 22 December. Available at: https://perma.cc/4RXP-GYV3

Brunton F (2013) *Spam: A Shadow History of the Internet*. Cambridge, MA: MIT Press.

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 2053951715622512.

Burk DL and Cohen JE (2001) Fair use infrastructure for rights management systems. *Harvard Journal of Law & Technology* 15: 41.

Caplan R (2018) *Content or Context Moderation?* New York, NY: Data & Society Research Institute. Available at: https://datasociety.net/output/content-or-context-moderation/ (accessed 17 December 2018).

Chen Y, Zhou Y, Zhu S, et al. (2012) Detecting offensive language in social media to protect adolescent online safety. In: *2012 International conference on privacy, security, risk and trust and 2012 international conference on social computing*, 2012, pp. 71–80. Piscataway, NJ: IEEE.

Chouldechova A, et al. (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: *Proceedings of Machine Learning Research 81: 1–15. Conference on Fairness, Accountability, and Transparency (FAT*)*, New York, NY, 23–24 February 2018.

Crawford K (2016) Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values* 41(1): 77–92.

Datar M, Immorlica N, Indyk P, et al. (2004) Locality-sensitive hashing scheme based on p-stable distributions. In: *Proceedings of the twentieth annual symposium on computational geometry*, 2004, pp. 253–262. New York, NY: ACM.

Davis A and Rosen G (2019) Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer. *Facebook Newsroom*. Available at: https://about.fb.com/news/2019/08/open-source-photo-video-matching/ (accessed 30 December 2019).

Dinakar K, Jones B, Havasi C, et al. (2012) Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems* 2(3): 18.

Duarte N, Llanso E and Loup A (2017) *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Washington, DC: Center for Democracy & Technology. Available at: https://perma.cc/NC9B-HYKX (accessed 17 January 2019).

Elish MC and boyd danah (2018) Situating methods in the magic of Big Data and AI. *Communication Monographs* 85 (1): 57–80.

Erickson K and Kretschmer M (2018) This Video is Unavailable: Analyzing Copyright Takedown of User-Generated Content on YouTube. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 9(1).

Ekstrand M, et al. (2018) Exploring author gender in book rating and recommendation. In: *Proceedings of the 12th ACM conference on recommender systems*. New York, NY: ACM.

Erickson K and Kretschmer M (2018) This video is unavailable. *Journal of Intellectual Property. Information Technology & Electronic Commerce Law* 9(75).

European Commission (2016) Code of conduct on countering illegal hate speech online, 31 May. Available at: https://perma.cc/3M7U-5AQY

Facebook Newsroom (2016) Partnering to help curb spread of online terrorist content. Available at: https://perma.cc/V8DZ-AZZ7

Facebook Newsroom (2017) Facebook, Microsoft, Twitter and YouTube announce formation of the global internet forum to counter terrorism. Available at: https://perma.cc/6QYS-ML76

Fiedler K (2016) EU internet forum against terrorist content and hate speech online: Document pool. EDRi. Available at: https://perma.cc/A6SM-Q3EA

Gadde V and Gasca D (2018) Measuring healthy conversation. In: Twitter Blog. Available at: https://perma.cc/3YAU-L2YQ

Geiger RS (2011) The lives of bots. In: *Wikipedia: A Critical Reader*. Amsterdam: Institute of Network Cultures.

Geiger RS (2014) Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society* 17(3): 342–356.

GIFCT (2019) About the global internet forum to counter terrorism. Available at: https://perma.cc/44V5-554U

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

Gorwa R (2019a) The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review* 8(2). https://doi.org/10.14763/2019.2.1407

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

Gorwa R and Garton Ash T (2020). Democratic Transparency in the Platform Society. In: Persily N, Tucker J, eds. *Social Media and Democracy: The State of the Field and Prospects for Reform*. Cambridge, UK: Cambridge University Press, 2020. Available at: 10.31235/osf.io/ehcy2

Grimmelmann J (2015) The virtues of moderation. *Yale Journal of Law & Technology* 17: 42.

Hardt M, Price E and Srebro N (2016) Equality of opportunity in supervised learning. In: *30th Conference on Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain, 5–10 December 2016.

Harris CG and Stephens M (1988) A combined corner and edge detector. In: *Proceedings of the Alvey Vision Conference (AVS)*, Manchester, UK, 31 August–2 September 1988.

Hoffmann AL (2019) Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22(7): 900–915.

Holland A, Bavitz C, Hermes J, et al. (2016) Intermediary liability in the United States. Berkman Centre for Internet & Society NOC Case Study Series. Available at: https://perma.cc/2QAY-UTDY

Hosseini H, Kannan S, Zhang B, et al. (2017) Deceiving Google's perspective API built for detecting toxic comments. *arXiv* 1702: 08138.

Huszti-Orban K (2017) *Countering Terrorism and Violent Extremism Online: What Role for Social Media Platforms*? Rio de Janeiro: Fundação Getulio Vargas.

Internet Creators Guild (2016) YouTube de-monetization explained. In: Medium. Available at: https://perma.cc/R7AQ-MG2F

Kadri T and Klonick K (2019) *Facebook v. Sullivan: Building Constitutional Law for Online Speech*. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract = 3332530 (accessed 12 February 2019).

Kaye D (2019) *Speech Police: The Global Struggle to Govern the Internet*. New York, NY: Columbia Global Reports.

Keef T and Ben-Kereth L (2016) Copyright management. In: Facebook for Media Blog. Available at: https://perma.cc/YB5H-BEM5

Lampe C and Resnick P (2004) Slash (dot) and burn: Distributed moderation in a large online conversation space. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 543–550. New York, NY: ACM.

Leong E (2016) New ways to control your experience on Twitter. Twitter Product. Available at: https://perma.cc/Y69K-2VK3

Li, S and Williams J (2018) Despite What Zuckerberg's Testimony May Imply, AI Cannot Save Us. *Electronic Frontier Foundation Deeplinks Blog*. Available at: https://www.eff.org/deeplinks/2018/04/despite-what-zuckerbergs-testimony-may-imply-ai-cannot-save-us

Llansó E (2016) Takedown Collaboration by Private Companies Creates Troubling Precedent. *Center for Democracy & Technology Blog*. Available at: https://cdt.org/blog/takedown-collaboration-by-private-companies-creates-troubling-precedent/

Llansó E (2019) Platforms want centralized censorship. That should scare you. *Wired*, 18 April. Available at: https://perma.cc/PWB3-NU3Q

Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed representations of words and phrases and their compositionality. In: *27th Conference on Neural Information Processing Systems (NeurIPS)*, Lake Tahoe, NV, 5–10 December 2013.

Mosseri A (2019) Our Commitment to Lead the Fight Against Online Bullying. In: *Instagram Info Center*. Available at: https://instagram-press.com/blog/2019/07/08/our-commitment-to-lead-the-fight-against-online-bullying/ (accessed 30 December 2019).

Niu X and Jiao Y (2008) An overview of perceptual hashing. *Acta Electronica Sinica* 36(7): 1405–1411.

Oswald M, Grace J, Urwin S, et al. (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law* 27(2): 223–250. 10.1080/13600834.2018.1458455.

Patel R (2013) First world problems: A fair use analysis of internet memes. *UCLA Entertainment Law Review* 20: 235.

Perel M and Elkin-Koren N (2015) Accountability in algorithmic copyright enforcement. *Stanford Technology Law Review* 19: 473.

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Roberts S (2018) Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23(3).

Rosen G (2018) F8 2018: Using technology to remove the bad stuff before it's even reported. In: Facebook Newsroom. Available at: https://perma.cc/VN5P-7VNU

Rosen G (2019) A further update on New Zealand terrorist attack. In: Facebook Newsroom. Available at: https://perma.cc/2KYX-EKD3

Santa Clara Principles on Transparency and Accountability in Content Moderation (2018) Available at: https://perma.cc/8ZTS-89FR

Schmidt A and Wiegand M (2017) A survey on hate speech detection using natural language processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, 3 April 2017.

Seering J, Wang T, Yoon J, et al. (2019) Moderator engagement and community development in the age of algorithms. *New Media & Society* 21(7): 1417–1443. https://doi.org/10.1177/1461444818821316

Sinders C (2017) Toxicity and tone are not the same thing: Analyzing the new Google API on toxicity, PerspectiveAPI. *Medium*. Available at: https://perma.cc/R9BM-V638

Smit R, Heinrich A and Broersma M (2017) Witnessing in the new memory ecology: Memory construction of the Syrian conflict on YouTube. *New Media & Society* 19(2): 289–307.

Soha M and McDowell ZJ (2016) Monetizing a meme: YouTube, content ID, and the Harlem Shake. *Social Media + Society* 2(1): 2056305115623801.

Sonderby C (2019) Update on New Zealand. In: Facebook Newsroom. Available at: https://perma.cc/ZA85-2Y3X

Spertus E (1997) Smokey: Automatic recognition of hostile messages. In: *Proceedings of the Ninth Innovative Applications of Artificial Intelligence Conference (IAAI-97)*, Providence, RI, 27–28 July 1997.

Suzor NP (2019) *Lawless: The secret rules that govern our digital lives*. Cambridge, UK: Cambridge University Press.

Suzor NP, West SM, Quodling A, et al. (2019) What do We mean when We talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13.

Thompson N (2017) Instagram's CEO wants to clean up the internet – But is that a good @&#$ing idea? *Wired*, 14 August. Available at: https://perma.cc/AH3V-L5D9

Urban JM, Karaganis J and Schofield B (2017) *Notice and takedown in everyday practice*. UC Berkeley Public Law Research Paper (2755628).

Waseem Z, Davidson T, Warmsley D, et al. (2017) Understanding abuse: A typology of abusive language detection subtasks. arXiv:1705.09899.

Wulczyn E, Thain N and Dixon L (2017) Ex machina: Personal attacks seen at scale. In: *Proceedings of the 26th international conference on World Wide Web*, Perth, Australia, 3–7 April 2017.

York J and McSherry C (2019) Content moderation is broken. Let us count the ways. In: Electronic Frontier Foundation Blog. Available at: https://perma.cc/7FA6-WD6Z

Zehlike M, Bonchi F, Castillo C, et al. (2017) Fa* ir: A fair top-k ranking algorithm. In: *Proceedings of the 2017 ACM conference on information and knowledge management*. New York, NY: ACM.