ARTICLE 19

Algorithms and Automated Decision-Making in the Context of Crime Prevention: A briefing paper

2016

Table of contents

Introduction

In this briefing paper, ARTICLE 19 evaluates the human rights impact of algorithmic or automated decision-making (algorithmic decision-making) in the area of crime prevention, with a particular focus on terrorism/extremism. We analyse recent policy developments in this area in the USA and Europe, where these issues have been particularly high on policy makers' agenda.

Algorithmic decision-making is widely used in various contexts. In particular:

- Several initiatives have been put forward to tackle **online "extremism"**
  - In the USA, the Obama administration has advocated for the use of "hashes" for the detection and automatic removal of extremist videos and images from the Internet. Additionally, there have been proposals to modify search algorithms in order to "hide" websites that would incite and support extremism from results pages. The hash mechanism has been adopted by Facebook and YouTube for video content; however no information has been released concerning the level of human input or the criteria used to establish which videos are "extremist;"

  - In Europe, while similar projects to the above are undergoing scrutiny, Interpol has created a regional organization monitoring online extremist content called the "Internet Referral Unit". The Unit identifies content in breach of the terms and conditions of each online platform. The companies can then voluntarily act upon the Unit's report. The system will be automated in the next few months with the introduction of the "Joint Referral Platform".

- Algorithmic decision-making is also currently being deployed in the USA for the **prevention of crime**. An algorithm is currently used to create a risk assessment predicting the likelihood of a person reoffending in the future. There have been reports on the inherent bias of the algorithm against people of colour and minorities, which are particularly troubling as the risk assessment might have far-reaching repercussions on the conditions of the sentence given. Other algorithmic-based software is currently used by the police in order to foresee where and when criminal acts are likely to take place. This software has also been criticised for reflecting and perpetuating bias: the algorithm is based on historical data which reflects systematically biased police practices.

- One of the most widely accepted uses of algorithmic decision-making is in the removal and filtering of **child sex abuse images/videos** by online platforms and ISPs. This filtering system is now automated in both the US and the UK, based on the use of "hash" technology to identify content for removal and filtering. This practice has been criticised due to the lack of judicial oversight mechanisms, lack of transparency and the risks of over-blocking.

- Algorithmic decision-making is also widely used in the context of **copyright removals**. In these cases there is usually human input in the process, as copyright owners are asked to upload their material within a specific program and to decide what consequence a breach of copyright should have. These programs have been widely criticised; in particular in the

context of YouTube, whose appeal system can be extremely complex and where the copyright owners are seen as "playing judge, jury and executioner".[1]

- Algorithmic decision-making has also been implemented in the context of **abusive messages on social media**. On Twitter, for example, "abusive messages" are filtered out of the recipient's notifications, while still remaining visible on the platform. Staff members then make a decision regarding possible bans and suspensions for the abusive user.

Overall, over-blocking and a lack of clarity on appeals procedures are recurrent problems in the context of algorithmic decision-making. Such initiatives are usually based on "self-regulatory" mechanisms which are, therefore, placed beyond the scope of the law and judicial oversight. Moreover, in practice, online intermediaries decide what available redress mechanisms should be, if any, and the level of human oversight over the automated decision-making.

This briefing paper provides an initial snapshot of some of the human rights implications of automated decision-making, in particular regarding the right to freedom of expression. ARTICLE 19 believes that it is important to ensure that human rights are protected in the context of algorithmic decision-making. Hence, this briefing paper provides an overview of some of the important issues and policy developments that should be considered when developing policy responses in this area. This will be addressed further in our future work.

[1] C. Hassan, What about all that copyright takedown abuse, YouTube?, Digital Music News, 29th February 2016, available at http://bit.ly/2fW8InR.

Key concepts

The term "**algorithm**" can refer to any computer code that carries out a set of instructions, and is essential to the way computers process data.[2] Algorithms are encoded procedures for transforming input data into desired output, based on specific calculations.[3]

Automated decision-making "generally involves large scale collection of data by various sensors, data processing by algorithms and subsequently, automatic performance."[4] It is an efficient means to manage, organise and analyse large amounts of data and then to structure decision-making accordingly.[5]

Algorithms and automated decision-making may involve partial or no human contribution. They are now being used in a growing number of contexts, from cameras issuing speeding tickets to the automated removal of inappropriate content on online platforms. Substantial research is being done to develop algorithms that can reason in a human-like fashion. The algorithms underlying "artificial intelligence" will play an increasingly important role in our societies. In the context of online content, algorithms are used in four main ways: prioritization, classification, association and filtering.[6]

Algorithmic law enforcement is increasingly being implemented by online intermediaries, which have acquired the role of managing online behaviour and enforcing the rights of Internet users.[7] They are seen as offering a point of control for monitoring, filtering, blocking and disabling access to content, and managing online content according to different laws, from security to defamation to intellectual property.[8] This development arguably gives intermediaries the role of adjudicators and enforcers, challenging a variety of human rights and freedoms.

In the context of content linked to extremism, algorithms may be employed to delist, automatically remove or "hide" content.[9] The removals may originate from a filtering system, a blocking system or a take-down notice system.[10]

---

[2] Centre for Internet and Human Rights, The Ethics of Algorithms: from radical content to self-driving cars – final draft background paper, GCCS 2015; available at http://bit.ly/1D7lgTx.
[3] T. Gillespie, The relevance of algorithms, in T. Gillespie, P. Boczkowski & K. Foot, *Media* technologies: essays on communication, materiality and society, 2014, Cambridge MA:MIT Press, p.167.
[4] M. Perel & N. Elkin-Koren, Accountability in algorithmic copyright enforcement, Stanford Technology Law Review, Forthcoming.
[5] *Ibid.*
[6] Centre for Internet and Human Rights, The Ethics of Algorithms, *supra note.*
[7] Perel & Elkin-Koren, *op.cit.*
[8] P. Sánchez Abril, Private Ordering: A Contractual Approach to Online Interpersonal Privacy, 45 WAKE FOREST L. REV. 689 (2010).
[9] Delisting is currently, for example, used by Google where users search around the topic of suicide; see e.g. http://bit.ly/2ftErUq; the Obama administration has previously proposed delisting extremist material; Facebook and YouTube are currently removing videos categorised as extremist from their platforms; see below.
[10] Removals may involve human interaction to varying extents. For example, take-down removals are based on public scrutiny of online content. Filtering systems instead may be automated and human input may take place only at the review level.

Applicable international human rights standards

International human rights standards

While there is no international standard directed solely at the use of algorithms, artificial intelligence and automated decision-making, there are a number of soft law instruments covering their use.

The use of automated decision-making for the purpose of content removal and filtering can also have an impact on internationally recognised human rights and freedoms, in particular the right to freedom of expression, guaranteed in Article 19 of the International Covenant on Civil and Political Rights (ICCPR) and in regional treaties. Insofar as automated decision-making may interfere with this right, it is also vital that individuals have the right to challenge those decisions to prevent abuses and unrestricted censorship. The right to an effective remedy and the right to a fair trial (Article 14 of the ICCPR) are therefore equally relevant in this context.

There is a body of soft law international standards that are relevant to the use of algorithms and automated decision-making by intermediaries. The key principles include the following:

- Intermediaries should only implement restrictions on human rights after judicial intervention;[11]
- Intermediaries should be transparent with their users with regards to the removal and blocking of content;[12]
- Users should have the ability to challenge the blocking and filtering of content;[13]
- Users should have the right not to be subject to a decision which is based solely on automated processing which has legal impact for individuals or similarly significantly affects them;[14]
- States should not impose a general obligation on intermediaries to monitor the information that they transmit, store, automate or otherwise use.[15]

[11] Report of the Special Rapporteur to the Human Rights Council on the promotion and protection of the right to freedom of opinion and expression, 16 May 2011, A/HRC/17/27, para 47.
[12] *Ibid*,, para 47.
[13] Recommendation CM/Rec(2008)6 of the Committee of Ministers to Member states on measures to promote and respect for freedom of expression and information with regard to internet filters, s1.
[14] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) Article 22.
[15] Directive 2000/31/EC 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive).

European human rights standards

The Council of Europe

The Council of Europe has adopted several standards that are applicable to automated decision-making; these include:

- **Recommendation CM/Rec (2010)13 of the Committee of Ministers to member states on the protection of individuals regarding to automatic processing of personal data in the context of profiling:** the Recommendation relates to the question of privacy and non-discrimination. It deals with automatic data processing techniques that consist of applying a profile to an individual in order to make decisions concerning him or for the purpose of analysing or predicting his future behaviours.[16] The collection and processing of personal data in the context of profiling should be fair, lawful and proportionate, and for specified and legitimate purposes.[17] Additionally, the data controller should provide the data subject with information on this automated process, including highlighting that their data will be collected in the context of profiling; the purpose of the profiling; and the existence of appropriate safeguards and other information necessary for guaranteeing the fairness of recourse to profiling;[18]

- **Recommendation CM/Rec(2008)6 of the Committee of Ministers to Member states on measures to promote and respect for freedom of expression and information with regard to internet filters**: In this Recommendation, the Committee of Ministers has underlined that "users' awareness, understanding of and ability to effectively use Internet filters are key factors which enable them to fully exercise and enjoy their human rights and fundamental freedoms, in particular the right to freedom of expression and information, and to participate actively in democratic processes."[19]

  The Committee highlighted that users should have the opportunity to challenge the blocking and filtering of content and to seek clarifications of the decisions and remedies available in cases of violations.[20] Particularly relevant for the current discussion, the Committee noted that blocking measures should only be carried out in compliance with Article 10(2) of the European Convention on Human Rights (ECHR), and that a state should only take action if the filtering concerns specific content, the legality of which has already been decided by a competent national authority.[21]

---

[16] Guide to human rights for Internet users – explanatory memorandum, 1197 Meeting 2014 Steering Committee on Media and Information Society, para 71.
[17] Recommendation CM/Rec(2010)13 of the Committee of Ministers to member states on the protection of individuals regarding to automatic processing of personal data in the context of profiling, s3(1).
[18] *Ibid.*, s4. Under s4(1)(f) additional information includes the categories of bodies to whom the personal data may be disclosed, the purposes for doing so, the possibility to refuse or withdraw consent, the conditions of exercise of the right of access, and the duration of data storage.
[19] Recommendation CM/Rec(2008)6, *op.cit.,* s1. ARTICLE 19 notes that Internet "filtering" is commonly associated with the use of technology that blocks pages by reference to certain characteristics, such as traffic patterns, protocols or keywords, on the basis of their perceived connection to content deemed inappropriate or unlawful; for a short introduction to filtering, see Open Net Initiative.
[20] *Ibid.*
[21] *Ibid.*, s3(2).

- **Recommendation CM/Rec(2012)4 of the Committee of Ministers to Member states on the protection of human rights with regard to social networking services**: In this Recommendation, the Committee of Ministers has recognised the fundamental role of social networks as human rights enablers and catalysts for democracy as they constitute a key tool for expression and communication.[22] It was identified that threats to human rights may arise from the lack of legal and procedural safeguards surrounding the exclusion of users.[23]
  States are encouraged to co-operate with the private sector to:
  - Provide users with concise explanations of the terms and conditions of social networking services in a form and language that is geared to, and easily understandable by, the target groups of the social networking services;
  - Provide users with clear information about the editorial policy of the social networking service provider in respect of how it deals with apparently illegal content and what it considers to be inappropriate content and behaviour on the network.[24]

  The capacity of users to understand both editorial policies and terms and conditions are fundamental since, in an increasingly self-regulated sector, these are often the benchmark against which the removal and blocking of content will take place.

- **Recommendation CM/Rec(2007)16 of the Committee of Ministers to member states on measures to promote the public service value of the Internet**: Here, the Committee of Ministers has recommended that member states should adopt policies to preserve and enhance the protection of human rights and respect for the rule of law in the information society.[25] In particular, it highlights that attention should be paid to:
  - The right to free expression, information and communication on the Internet and via other ICTs promoted by ensuring access to them;
  - The need to ensure that there are no restrictions to the abovementioned right (for example in the form of censorship) other than to the extent permitted by Article 10 of the European Convention on Human Rights, as interpreted by the European Court of Human Rights;[26]

  The Committee regarded the diversity, and the capacity of different communities and groups to participate in 'the information society to be of particular importance.[27] Online pluralism is key to the development of democracy and society and might be restricted by the use of automated decision-making tools, which may be biased against specific groups, in particular in the context of extremism.[28]

- **Recommendation CM/Rec(2012)3 of the Committee of Ministers to Member States on the protection of human rights with regard to search engines**: Here, the Committee of

[22] *Ibid.*, s1(1).
[23] *Ibid.*, s3.
[24] *Ibid.*, s1(3).
[25] Recommendation CM/Rec(2007)16 of the Committee of Ministers to member states on measures to promote the public service value of the Internet, s1.
[26] *Ibid.*
[27] *Ibid.*, s4.
[28] C Radsch, Privatizing censorship in fight against extremism is risk to press freedom, Committee to Protect Journalists, 16 October 2015.

Ministers recognised that search engines' mechanisms can impact upon freedom of expression as well as the right to seek, receive and impart information.[29] In particular, it was expressed that "challenges may stem from the design of algorithms, de-indexing … and lack of transparency about both the selection process and ranking of results."[30]

The Committee recommended states to, inter alia:

- Enhance transparency regarding the way in which access to information is provided, in order to ensure access to, and pluralism and diversity of, information and services;
- Enhance transparency in the collection of personal data and the legitimate purpose for which they are being processed;
- Encourage search engine providers to discard search results only in accordance with Article 10, paragraph 2, of the Convention. In this event, the user should be informed as to the origin of the request to discard the results subject to respect for the right to private life and protection of personal data.[31]

The European Union

At the European Union level, the following standards are applicable:

- **E-Commerce Directive**:[32] Under Article 15 of the E-Commerce Directive, member states shall not impose a general obligation on providers to monitor the information they transmit or store. Similarly, they cannot impose a general obligation to seek facts or circumstances indicating illegal activity.

  However, this does not prevent states from imposing obligations in specific cases, and the directive does not affect the possibility of courts ordering injunctions requiring ISPs to remove illegal content.[33] Notably, states can also apply a duty of care to information host providers requiring them to detect and prevent certain types of illegal activity.[34] Importantly, while the obligation must be reasonable and specified in national law, there is no definition of what constitutes "illegal content" under the directive.

- **General Data Protection Regulation**:[35] Under the new General Data Protection Regulation, coming into force in 2018, regulations are imposed on automated individual decision-making. Under Article 22, the data subject shall have the right not to be subject to a decision which is based solely on automated processing, including profiling, and which produces legal effects concerning him or her or otherwise significantly affects him or

---

[29] Recommendation CM/Rec(2012)3 of the Committee of Ministers to Member States on the protection of human rights with regard to search engines, s4.
[30] *Ibid.*
[31] *Ibid.*, s7-8.
[32] E-Commerce Directive, *op.cit.*.
[33] Directive 2000/31/EC 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, Article 15.
[34] A broad duty of care was imposed on a Dutch case. A file sharing site filtered pornography and viruses but not possible copyright infringing material. The judge concluded that the site was liable for this content as it had the capacity and means to control the site on illegal content, but refused to do so on content infringing intellectual property; see e.g. http://bit.ly/2fWJEYH.
[35] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

her.[36] Section 2 underlines that exceptions can be made "if necessary for entering into, or performance of, a contract", authorised by "Union or Member State law" or "based on the data subject's explicit consent."

Further protection is provided to users under Article 22(3), which prescribes that, even in the case of exceptions, data controllers shall "provide appropriate safeguards" including "the right to obtain human intervention...to express his or her point of view and to contest the decision." In particular, under Article 22(4), automated processing based on special categories of data is prohibited unless "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place."

While there is no further information on what the safeguards prescribed under Article 22(4) might be, under Articles 13 and 14, a subject has the right to "meaningful information about the logic involved" when profiling takes place. Importantly, it has been underlined that the General Data Protection Regulation rightly acknowledges that, when algorithms are used in society, their ethical design "requires coordination between technical and philosophical resources of the highest calibre."[37]

- **The Data Protection Directive**[38]: Under Article 15 of the Directive, states have to prevent everyone from being subject to a decision "which produces legal effects concerning him or significantly affects him" which is based solely on the automated processing of data for the purpose of a profiling application. Legal authorisation constitutes an exception, but the provision has however to lay down "measures to safeguard the data subject's legitimate interest".[39] Article 15 should be specifically interpreted as meaning that the data subject always has the right to know the logic of a decision made, and have this explained to them in particular in cases where automated decision-making is involved.[40]

[36] Regulation (EU) 2016/679, *op.cit.*, Article 22(1).
[37] B. Goodman & S. Flaxman, European Union regulations on algorithmic decision making and a right to explanation, 12 July 2016, arXiv:1606.08813.
[38] Directive 95/46/EC 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
[39] The Data Protection Directive, *op.cit.* Article 15(2)(b)
[40] A. Savin, Profiling and automated decision making in the present and new EU data protection frameworks, available at http://bit.ly/2fKjoxL.

Current information on automated decision-making and algorithms related to extremism

United States of America

Hash values removal

In June 2016, Lisa Monaco, President Obama's top counterterrorism adviser, supported newly created software designed to prevent the proliferation of extremist multi-media content online.[41] The algorithm - created by Hany Farid, a science professor, and supported by the Counter Extremism Project - is based on "hash" technology, which has already been adopted in the US for content related to child abuse."[42] A database would be created containing known extremism content, and social media and tech companies could run content through the algorithm. The algorithm would indicate if the content matched a hash identified as extremist, allowing the company to automatically flag or remove the content.

While a number of technology companies vigorously criticised the effectiveness of such mechanism, it appears that several social media companies have recently adopted the removal system. According to Reuters, YouTube and Facebook would be among the sites deploying Farid's software.[43] There is no information yet available on the involvement of human input or review in the removal process, or on the criteria which were used to identify the videos in the database as extremist.[44]

Automated ranking and filtering

In February 2016, a brainstorming meeting was organised between U.S. counterterrorism officials, technology executives, and entertainment representatives.[45] The Department of Justice stated during the meeting that the private sector, including content producers, social media companies, and NGOs, have "a crucial role to play in developing creative and effective ways to undermine terrorist recruiting and counter the call to violence."[46]

The February meeting was the culmination of a yearlong process involving discussions and meetings that have included the Department of Defence and State Department. In particular, a key proposal has been a request from the Pentagon to tweak the algorithms of several companies in order to promote certain types of content. However, the technology giants have publicly countered the proposal. A Google representative stated "this is a Pandora's box we

[41] E. Groll, Suppressing extremist speech: there is an algorithm for that!, 17 June 2016, Foreign Policy, available at http://atfp.co/29gHm5b.
[42] EDRi, Algorithms – censorship a la carte, 12 July 2016, EDRi, available at http://bit.ly/2fKr3vK.
[43] J. Menn & D. Volz, Exclusive: Google, Facebook quietly move toward automatic blocking of extremist videos, Reuters, 25 June 2016, available at http://reut.rs/28U55Vp.
[44] *Ibid.*
[45] S. Frenkel, Inside the Obama's administration attempt to bring tech companies into fighting against ISIS, BuzzFeed, 25 February 2016, available at http://bzfd.it/2ftwdLV.
[46] L. Stein, Tech giants, Madison Ave. meet with Justice Department to counter ISIS messaging, AdvertisingAge 24 February 2016, available at http://bit.ly/2gFaS7r.

won't open, because if we answer a request by the U.S. government to feature one search result over another, what's to stop other countries from requesting the same? What's to stop each country from tailoring the search results of their citizens to their agenda? It's not a path we are willing to explore."[47]

While algorithms are already employed by social media and search engines to shape rankings and news feeds, it seems that until now, the process has been mainly self-regulated. However, companies' use of automated filtering and review mechanisms has already attracted severe criticism. For example, Facebook was criticised in 2014 for allegedly "burying" content on the protests in Ferguson from its Newsfeed.[48] Additionally, it was reported that the major problem with algorithmic decision-making is the lack of appeal processes, the lack of transparency on the reasons for a removal and the limited channels to communicate with staff.[49]

Civil society responses

Global Network Initiative

The Global Network Initiative (GNI), a multi-stakeholder initiative including digital companies, civil society organizations and academics, intervened at the UN counter-terrorism meeting in Madrid to discuss the role of digital companies and algorithms.

The GNI stated its concerns over "the rush to adopt laws and policies that increase obligations of ICT companies to monitor and to restrict terrorist activities" which could have serious consequences for freedom of expression and the right to privacy.[50] The GNI also expressed concerns that various governments were currently influencing the content policies of digital companies and used these policies to remove content through informal mechanisms, which were outside the legal process. It found particularly troubling the requirements that digital companies block allegedly terrorist material without a court order, build backdoors into their products, and change their algorithms.[51]

---

[47] *Ibid.*

[48] G. Sullivan, How Facebook and Twitter control what you see about Ferguson, The Washington Post, 19 August 2014, available at http://wapo.st/2fwejve.

[49] K. Leetaru, The algorithms are taking over: who controls our online future?, Forbes, 2 January 2016, available at http://bit.ly/2fd7bFt.

[50] GNI Talking points at the UN counter-terrorism meeting in Madrid, 27 - 29 July 2015, available at http://bit.ly/2gr0k8W.

[51] The final Madrid Ministerial declaration only vaguely refers to the responsibility of ICTs by stating that "we underline the need to stop the criminal propaganda of the terrorist groups, the spread of the messages of incitement to violence and recruitment in social media networks and the internet, … and highlight that a closer dialogue with internet service providers is crucial in this regard;" Paragraph 8, available at http://bit.ly/1WcuYSK. However, as highlighted in the previous sections, it is clear that informal pressure has been put on ICTs to take action in the fight against extremism.

Use of algorithms and automated decision-making in other contexts

Child sexual abuse images

The removal of child abuse and sexual abuse material from the Internet has been largely uncontroversial.[52] Three approaches have been adopted to regulate the removal of this material: automated removal; intermediary based regulation (indirect enforcement targeting intermediaries rather than end users); and self and co-regulation (industry self- and co-regulatory schemes promoted by the government).[53]

States as well as private companies have taken steps in order to block child sexual abuse images. Blocking has taken place at both the ISP level as well as at the platform level.

In the **USA**, automated blocking of child abuse material has taken place at the platform and Internet Service Providers (ISP) level. It is based on the use of "hashes" which focuses on the file itself rather than its location, allowing for its automated removal in different locations of the web. The blacklist against which files are compared, however, is different depending on the platform.

In the US, AOL developed a blacklist entirely in-house, against which the algorithm compares the content, as it was "concerned that it would be treated as a state actor if it relied on a government provided list."[54] Conversely, at the state level, platforms like Facebook are currently blocking content based on hash values supplied by the New York law enforcement authorities.[55] While the removal system is completely automated, once content has been identified and removed, the company is "obliged by US mandatory reporting rules to notify the Cyber Tip Line by sending a report containing the image, username, email address and zip-code of the user".[56] The Cyber Tip Line will then review the information and decide if a law enforcement agency should be notified. At this point, companies may be obliged to disclose further details about the user; this system has resulted in numerous convictions.

Since 1996, the **United Kingdom's** approach to child sexual abuse images online has been based on an industry-led response. A private body called the Internet Watch Foundation (IWF) took the role of a hotline, receiving public complaints and determining whether particular web pages potentially contained illegal material.[57] The IWF would then forward those complaints to the police and the hosting provider to have the material removed in cases where it was hosted in the UK.[58] At the same time, British Telecom (BT) adopted a system to block access to web pages also hosted outside the UK, called "Cleanfeed", operating a blacklist of URL developed

[52] *Ibid.*, pp.12.
[53] T.J. McIntyre, Child abuse images and clean feeds: assessing Internet blocking systems, in I. Brown & E. Elgar Research Handbook on governance and the Internet, 2012.
[54] *Ibid.*, pp.4.
[55] *Ibid.*, pp. 12.
[56] *Ibid.*, pp.12.
[57] *Ibid.*, pp. 5.
[58] *Ibid.*, pp. 5.

by IWF. Due to public and governmental pressure, almost all UK ISPs implemented a system similar to Cleanfeed.[59]

In 2015, tech companies including Google, Facebook and Twitter announced their collaboration with the IWF to remove child sexual abuse images through the use of "hashes".[60] The database will be created by IWF based on the images found on the URL it has blocked.[61] It is unclear, however, if an appeal process will be put in place, if the information related to the account will be reported to the police, or if takedown notices will be provided.

A number of criticisms can be made with regard to the current automated blocking systems used in the context of child sexual abuse images:

- Firstly, these systems bypass legal oversight mechanisms when private corporations or organization implement them.

- Similarly, the codes and regulations are in themselves opaque and removed from public scrutiny.

- Additionally, "the right to be heard before a decision is made is not offered in most schemes, despite the fact that blocking operates a prior restraint of speech".[62] This is particularly problematic as in the majority of cases no notice is provided before or after the takedown has taken place. In relation to the role of private companies and ISPs, a key problem is the incentive to over-block in order to protect themselves from liability. In particular, empirical evidence has been presented showing that internet intermediaries make decisions in a manner that can minimise their financial, legal and reputational risk.[63] This may be more likely outside of Europe, in jurisdictions where liability may be imposed on Internet intermediaries for the content they host.

Copyright removals

Facebook has currently adopted a semi-automatic copyright removal system. For both audio and videos, copyright holders are able to access a tool, either Audible Magic or Rights Manager, which allows them to identify content matching their copyrighted material.[64] Importantly, copyright holders can decide to "whitelist" certain pages that have permission to utilise their videos.[65]

---

[59] *Ibid.*, pp.6.
[60] D. Gilbert, Child abuse images to be digitally fingerprinted to stop paedophiles sharing on Facebook and Google, IBtimes, 10 August 2015, available at http://bit.ly/1IzWbYg.
[61] *Ibid*
[62] T.J. McIntyre, *op.cit.*
[63] C. Ahlert, C. Marsden & C. Yung, How "Liberty" Disappeared from Cyberspace: The Mystery Shopper Tests Internet Content Self-Regulation, 2004, available at http://bit.ly/2g9FfPI.
[64] H. Ungureanu, Facebook launches video Rights Manager to combat Freebooting and Copyright issues, TechTimes, 13 April 2016, available at http://bit.ly/1SMsRQx; see also M. Shields, Facebook introduces new tools to crack down on video copyright violations, The Wall Street Journal, 27 August 2015, available at http://on.wsj.com/1MXLZgz.
[65] *Ibid.*

Part of the features of Rights Manager is the ability to create "match rules" that allow automatic actions to be taken whenever Facebook finds a match for copyright material, instead of requiring manual review of the problematic content.[66] One of the options currently available is an automated report option set for all matching material within the parameters set by the copyright-holding user.[67]

According to Facebook's Rights Manager page, "When we receive a report from someone claiming that you've uploaded video content to Facebook that infringes their copyright, we may need to remove that content from Facebook without contacting you first."[68] The process of appeal then varies according to the country where the breach has occurred and can simply involve following up with the reporting party or recourse to a procedure established by national law. It is unclear, however, how Facebook will prevent large-scale abuses of its new tool by copyright owners.

YouTube utilises a semi-automated copyright system called Content ID to identify and manage copyright violations. Copyright owners upload their files against which videos on YouTube are then scanned.[69] The rights holders have several options for what action should be taken when Content ID flags a video; the most common is a complete block of the video or a takeover of monetisation for the video.[70]

When the uploader contests a flag, an appeal process begins during which he will not receive the revenue for the work.[71] The first stage is a dispute process during which the copyright owner has 30 days to respond. If it fails to do so, the next step is an appeal, executed solely by the rights holder who can either remove the flag or, as in the case of the US, submit a DMCA takedown notice to deny the appeal. There is also another option for uploaders whose content has been flagged, at least in the US, which is to submit a counter-notice.[72] The Counter Notice obliges YouTube to reinstate the video and, unless the rights holder files a lawsuit in a federal court within 14 days, YouTube will be required to restore the video.[73]

Content ID has been widely criticised for mistakenly flagging videos due to its automated mechanisms. Additionally, the appeal process has been described as "a confusing automated system"[74] whereby "content owners seem to be playing judge, jury and executioner."[75]

---

[66] J. Loomer, Facebook Rights Manager: no more freebooting, Jon Loomer, 22 April 2016, available at http://bit.ly/2fKyi77.
[67] *Ibid.*
[68] https://rightsmanager.fb.com/
[69] YouTube Help: how content ID works, available at http://bit.ly/1TfcIv7.
[70] D. Van Winkle, YouTube is finally fixing its terrible copyright appeal process, The Mary Sue, 29 April 2016, available at http://bit.ly/2fdfS2B.
[71] *Ibid.*
[72] S. McArthur, How to beat a YouTube ContentID copyright claim – what every gamer and MCN should know, Gamasutra, 24 June 2014, available at http://ubm.io/1keQC9f.
[73] *Ibid.*
[74] R. Brandom, YouTube's complaint system is pissing off its biggest users, The Verge, 1 February 2016, available at http://bit.ly/1PtZnth.
[75] C. Hassan, What about all that copyright takedown abuse, YouTube?, Digital Music News, 29 February 2016, available at http://bit.ly/2fdlc6a.

Abusive messages and offensive speech

In 2015, **Twitter** introduced a new automatic filter, which is applied to all users and which cannot be switched off.[76] The filter automatically determines whether a tweet is likely to be abusive based on its content and the context of the message. In case it is found abusive by the software, the recipient will not receive a notification in their mentions,[77] thereby not noticing its appearance in their "timeline". However, they will still be able to see it in their feed if they follow the abusive user's account or via research.[78]

Additionally, Twitter has expanded the scope of its rules: while "direct, specific threats of violence against others" have always been banned, the prohibition now includes "threats of violence against others or promoting violence against others."[79] To enforce the policy, Twitter may require abusive users to delete the offending tweet and supply a phone number to the company.[80] While the abusive account may receive a temporary ban or suspension, Twitter's staff will make the decision on the adopted course of action.[81]

In June 2016, **Yahoo** announced the development of a new algorithm capable of detecting abusive messages. The algorithm was allegedly able to identify that a comment was abusive in 90 per cent of test cases.[82] The algorithm was created through a combination of machine learning and crowd-sourced abuse detection.[83] It analysed comment length, number of insult words and punctuation, using a dataset comprised of abusive and non-abusive comments from Yahoo News and Finance articles.[84] While the algorithm has not yet been used, various experts are supporting the creation of a similar system on different online platforms.[85]

---

[76] A. Hern, Twitter announces crackdown on abuse with new filter and tighter rules, The Guardian, 21 April 2015, available at http://bit.ly/1SZTW7b.
[77] A mention is "a Tweet that contains another user's @username in the body of the Tweet". These messages appear on the timeline of the mentioned user and in his notifications; see http://bit.ly/1r03A7A.
[78] Naked Security, Twitter's new anti-abuse filter hiders harassing tweets from your mentions, 22 April 2015, available at http://bit.ly/1aTWdOn.
[79] A. Hern, *op.cit.*
[80] Naked Security, *op.cit.*
[81] *Ibid*
[82] M. Reynolds, Yahoo's anti-abuse AI can hunt out even the most devious online trolls, Wired, 29 June 2016, available at: http://bit.ly/2aneSCK.
[83] *Ibid.*
[84] W. Knight, Yahoo has a tool that can catch online abuse surprisingly well, MIT Technology Review, 26 July 2016, available at http://bit.ly/2aKISuI.
[85] *Ibid.*

Algorithms and crime prevention

The United States' criminal justice system

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm used in the US to determine when criminal defendants should be released.[86] The algorithm provides a "risk assessment" report, a score predicting the likelihood of a person reoffending in the future.[87] In some states, the scores are used to assign bond amounts, while in others they are given to judges during criminal sentencing, as in Arizona, Colorado, Virginia and others.[88]

The Sentencing Reform and Corrections Bill introduced in 2015 proposes the introduction of a "post-sentencing risk- and needs-assessment system" for federal prisons.[89] The system would assign a score to inmates based on their likelihood to reoffend, effectively entrenching the use of algorithms like COMPAS in the federal prison system. The score would theoretically impact on housing assignments, telephone and visitation privileges, faith-based programming, and even sentence reductions.[90]

While the bill requires the department of justice to consult experts to ensure the metrics follows the best practices and provides for a three-year review, a study by ProPublica highlighted that the algorithm used was designed with bias.[91] The study found three key markers indicative of this bias:

- Only 20% of the people who were flagged as being at high risk of committing violent crimes within two years of being arrested actually went on to do so.
- Black defendants were twice as to likely be incorrectly flagged as being at high risk of committing crimes in the future.
- White people were more likely to be identified as being at low risk of committing crimes in the future.[92]

In particular, the study highlighted that while the questions posed during the risk assessment did not directly mention questions on race, they entrenched bias against black people (e.g. "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?").[93]

---

[86] R. Revesz, Criminal justice software algorithm used across the US is biased against black inmates, study finds, The Independent, 27 June 2016, available at http://ind.pn/28Y6JYE.
[87] Pulliam-Moore, An algorithm used by police to predict crime flagged black people twice as often as white people, The Fusion, 23rd May 2016, available at http://fus.in/2ayQ1Rd.
[88] J. Angwin, S. Mattu & L. Kirchner, Machine bias: there's a software used across the country to predict future criminals. And it's biased against blacks, ProPublica, 23 May 2016, http://bit.ly/1XMKh5R.
[89] C. Haugh, Prison by Algorithm, The Atlantic, 26th June 2016, available at http://theatln.tc/29uZQNN.
[90] *Ibid*
[91] Angwin et al, *op.cit.*
[92] Pulliam-Moore, *op.cit.*
[93] Angwin et al, *op.cit.*

USA's predictive policing

"Predictive Policing" (PredPol) is software used by police departments in the US which uses crime statistics from particular neighbourhoods to predict when and where crimes are likely to occur again.[94] It was introduced on the basis that if the police focused on an area where a crime was more likely to be committed, the level of crime in that location would be reduced overall.

Some studies have positively welcomed the use of PredPol, revealing that crime in the areas where it was deployed had substantially decreased, and noting that it would substantively reduce costs.[95] However, reporters have been highlighting the risk of PredPol becoming "an algorithmic justification for old-school racial profiling", placing more police in minority dense neighbourhoods.[96] The predictions are made based on historical data of crimes, and this is shaped by systematic police practices that have marginalised people from black and minority ethnic communities.[97] The bias embedded in the historical data is therefore transposed into the algorithm, which will generate further bias against minorities. According to the mathematician Cathy O'Neil, "algorithmic models can only amplify and expose insights based on the human behaviour that created the data in the first place."[98] Similar results have come out of other studies.[99]

[94] Pulliam-Moore, *supra note*.
[95] G. Mohler & M. Short, Randomized controlled field trials of predictive policing, Journal of the American Statistical Association V 110(512) 2015.
[96] A. Madrigal, Predictive policing: the future of crime fighting or the future of racial profiling?, Fusions, 24 March 2016, available at http://fus.in/2fcVqyQ.
[97] J. Smith, Minority Report is real – and it's already really reporting minorities, TechMic, 9 November 2015, available at http://bit.ly/2g9foY2.
[98] *Ibid.*
[99] L. Floridi, Big Data and Their Epistemological Challenge, Journal of Philosophy & Technology 25(4): 435–437 2012. See also Rosenblat et al., Discriminating tastes customer ratings as Vehicles for Bias, October 2016; available at http://bit.ly/2dSvNxW.

About ARTICLE 19

ARTICLE 19: Global Campaign for Freedom of expression (ARTICLE 19) is an international human rights organization that works globally to promote and protect freedom of expression and information. It was founded in 1987 and has an international office in London and regional offices in Bangladesh, Brazil, Kenya, Mexico, Senegal, Tunisia and Myanmar.

ARTICLE 19 advocates for the development of progressive standards on freedom of expression and access to information at the international level, and their implementation in national legal systems. It has produced a number of standard-setting publications which outline international and comparative law and best practices in areas such as defamation law, access to information and broadcast regulation.

On the basis of these publications and ARTICLE 19's overall legal expertise, the organisation publishes a number of legal analyses each year, comments on legislative proposals, as well as on existing laws that affect the right to freedom of expression, and develops policy papers and other documents. This work, carried out since 1998 as a means of supporting positive law reform efforts worldwide, frequently leads to substantial improvements in proposed or existing national legislation. All legal and policy materials are available at http://www.article19.org/resources.php/legal.