STUDY

Requested by the JURI committee

European Parliament

The impact of algorithms for online content filtering or moderation

"Upload filters"

Policy Department for Citizens' Rights and Constitutional Affairs
Directorate-General for Internal Policies
PE 657.101 - September 2020

EN

The impact of algorithms for online content filtering or moderation

"Upload filters"

Abstract

This study, commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the JURI Committee, addresses automated filtering of online content. The report introduces automated filtering as an aspect of moderation of user-generated materials. It presents the filtering technologies that are currently deployed to address different kinds of media, such as text, images, or videos. It discusses the main critical issues under the present legal framework and makes proposals for regulation in the context of a future EU Digital Services Act.

This document was requested by the European Parliament's Committee on Citizens' Rights and Constitutional Affairs.

AUTHORS
Prof. Giovanni Sartor, European University Institute of Florence.
Co-authored by Prof. Giovanni Sartor and Dr. Andrea Loreggia, working under Prof. Sartors supervision.

ADMINISTRATOR RESPONSIBLE
Udo BUX

EDITORIAL ASSISTANT
Monika Laura LAZARUK KUBINSKA

LINGUISTIC VERSIONS
Original: EN

ABOUT THE EDITOR
Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact the Policy Department or to subscribe for updates, please write to:
Policy Department for Citizens' Rights and Constitutional Affairs
European Parliament
B-1047 Brussels
Email: poldep-citizens@europarl.europa.eu

Manuscript completed in September 20202

This document is available on the internet at:
http://www.europarl.europa.eu/supporting-analyses

DISCLAIMER AND COPYRIGHT

CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application programming interfaces
DSA	Digital Services Act
GDPR	EU General Data Protection regulation
ICT	Information communications technology
IP	Intellectual Property
SABAM	Société d'Auteurs Belge – Belgische Auteurs Maatschappij
URL	Uniform Resource Locator

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Background

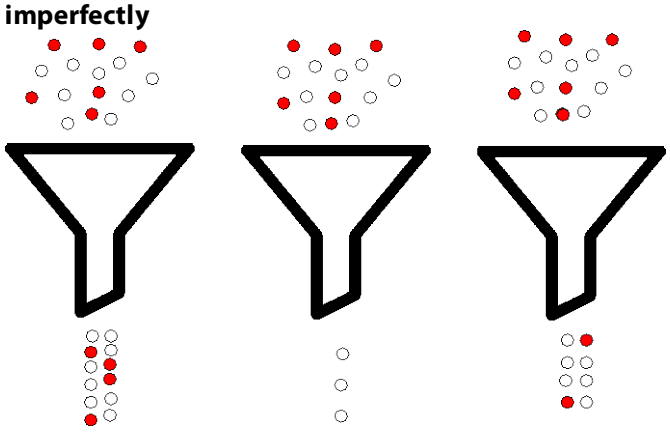
The 20 years old eCommerce Directive has played an important and largely positive role in the development of the digital economy and online information environment, but it is now being applied in a completely changed technological, economic and social context: a range of new technologies are available, from cloud platforms to AI systems; some Internet companies have become global players, with huge financial and technological resources; access to information and social interaction has moved online; economic and other activities increasingly exploit the integration of digital and physical resources.

The planned Digital Services Act (DSA), which will replace the Directive, should address this new technological, economic, and social context. A key issue concerns the regulation of digital services, in particular, those that fit the category of online platforms, namely, those digital services whose purpose is to facilitate digital interaction between their users (whether firms or individuals). In particular, platforms for user-generated content enable users to express themselves, to create, transmit or access information and cultural creations, and to engage in social interactions. Such platforms contribute to the fulfilment of legitimate individual interests, enable the exercise of fundamental rights (such as freedom of expression and association), and support the realisation of social values (such as citizens' information, education, and democratic dialogue), but they also provide opportunities for harmful behaviour: uncivility and aggression in individual exchanges, disinformation in the public sphere, sectarianism and polarisation in politics, as well as illegality, exploitation and manipulation.

To prevent unlawful and harmful online behaviour, or at least to mitigate its effect, *moderation* is needed, namely, the active governance of platforms meant to ensure, to the extent that it is reasonably possible, productive, pro-social and lawful interaction of the users. Unless moderation facilitates cooperation and prevents abuse, online communities tends to become dysfunctional, victims of spammers, vandals and trolls.

The present report addresses a key aspect of moderation today, i.e., automated filtering meant to classify, and consequently demote or exclude, user-generated materials. Automated filters are needed in order to monitor the huge amount of material that is uploaded online and detect (potentially) unlawful and abusive content. However, their use present risks, as it may lead to the exclusion of valuable content, and it may affect freedom of expression, access to information and democratic dialogue. The regulation of automated filtering should aim indeed at achieving two parallel, and often conflicting, objectives: (a) preventing, limiting and mitigating as much as possible the individual and social harm that can be caused by unlawful or inappropriate content and

Figure 1 Filtering out too little, too much, or still imperfectly



activities, while (b) allowing and even facilitating the delivery of beneficial content as well as free and civil interaction (see).

The Commerce Directive apparently adopts a hand-off approach to moderation, and in particular to filtering: it shields providers from liability for unlawful user-generated content, while it prohibits States from imposing general obligations to monitor upon providers. This approach has been challenged in recent years, as new technologies have become available for filtering, and such technologies have been widely used by leading content providers, such as search engines and social networks.

Aim

The purpose of this report is to provide an in-depth analysis of the technological and legal issues pertaining to the regulation of online filtering.

The report introduces the main filtering technologies that are currently deployed to address different kinds of media, such as text, images, or videos. Examining their strengths and weaknesses:

- Metadata searching, hashing, and fingerprinting are used to reliably identify copies of known digital works;
- Blacklisting is used to find unwanted expressions;
- Advanced techniques for natural language processing are used to address meaning and context;
- Multiple techniques, often based on AI, are used to identify unwanted images, or combinations of text and images, and to translate spoken language into text.

The report addresses the accuracy of filtering systems:

- Such systems are based on probabilistic methods, so all errors cannot be avoided.
- Determining what counts as a correct answer is problematic, given that the "ground truth" is provided by human assessments, in particular the human assessments which make up the training sets of automated classifiers.

Given the fallibility and subjectivity of automated filters, their functioning should be controllable and possible failures should be timely addressed. The report reviews methods to provide transparency and redress:

- Transparency requires that the individuals concerned, and society at large are informed about the filtering process;
- Appeal and redress mechanisms are needed to enable challenges by users, and to remedy failures.

The report also considers the accessibility and costs of filtering technologies. Such technologies are mostly used by large players, which often develop powerful inhouse systems, but some solutions are also accessible to smaller companies.

The report finally addressed the regulation of filtering. First it introduces some premises that should be taken into account:

- Automated filtering should not be discouraged, since it is an essential component of an efficient online moderation;
- Filtering is fallible, even when developed and implemented in good faith to address abusive content;
- Permissible pro-social filtering is not limited to unlawful content; it may justifiably address any kind of content that is objectionable in the context of a particular online community;
- While justified filtering is not limited to unlawful materials, it should not be based on capricious or arbitrary choices of platform owners and moderators;
- Contractual clauses that authorise filtering, or restrict remedies available to users have to comply with requirement of consumer protection;
- No set of mechanically applicable rules can definitely determine what counts as legally acceptable or unacceptable in a platform; both standards and rules are needed in the regulation of filtering;
- Due care/reasonableness standards can be used to assess providers' behaviour;
- Inappropriate regulation can induce excessive or insufficient filtering, so that valuable content is made inaccessible, or harmful content is remains online;
- Online diversity has a positive value, so that different filtering standards may justifiably be applied to different communities;
- There is uncertainty on what obligations to monitor content are prohibited under EU law.

On this basis some policy options are presented:

- An update of the general principles on providers' immunity contained in the eCommerce directive should be considered. It could be made explicit that providers' immunity for unlawful user-generated content presupposes that providers adopt reasonable measures to avoid unlawfulness and harm. It could also be newly specified to what extent EU law protects providers against the imposition of legal obligations to monitor, remove, or block content.
- It should also be clarified that engagement in moderation, and in particular in filtering out unlawful or abusive content should not affect whatever immunities or other advantages are granted to providers (Good Samaritan clauses).
- Procedural remedies should be introduced against online removals, so that the uploaders of filtered out, or demoted, content are informed on automated decisions concerning their content, are provided with explanations, and can challenge such decisions obtaining human responses.
- Public authorities should address online filtering, through specific regulations, and decisions on controversial cases. Existing authorities on media or telecommunications, could play a role. They could be coordinated by existing EU bodies or by a newly created authority.
- Small and medium enterprises should be supported. They should be responsible only for failing to adopt measures (including filtering) that are accessible to them, both

technologically and economically; incentives should be provided to the development of technically sound and affordable solution for automated filtering, possibly open source ones.

- An EU-broad approach to the regulation of filtering should be developed, while allowing for national diversity in the assessment of the lawfulness of online content. This will provide certainty and contribute to the digital single market.
- A broad debate on online moderation, and in particular, filtering is needed, involving not only political and administrative authorities, but also civil society and academia. This debate should put legal regulation in a broader context where the relation between human and social values and the online informational environment is addressed.

1. INTRODUCTION

The planned Digital Services Act (DSA) will replace the 20 years old eCommerce Directive,[1] which has so far provided the basic legal framework not only for the online economy, but also for online personal exchanges, as well as the online public sphere. This is due to the fact that certain private operators, so-called Internet intermediaries, provide the services, and indeed the infrastructure. on which online social interactions are taking place. Thus, the practices and business models adopted by Internet intermediaries — and consequently the regulations covering them— affect all digital interactions and hence the whole of the digital society.

The eCommerce Directive focused on the goal of promoting the development of the digital economy. It sought to expand the internal market, by including in it the emerging online services. Its short text includes key provisions for the development of the Internet economy, such as the freedom to provide digital services without the need for an authorisation, the obligation to provide information to the recipients of such services, the validity of digital contracts, and the liability exemptions for providers of mere-conduit, caching and hosting services.

The Directive has played an important and largely positive role in the development of the fledging digital economy in Europe, as did similar provisions in other countries, such as, in particular the 1999 Communication Decency Act and the 2000 Digital Millenniums Copyright Act in the US. However, the Directive is now being applied in a completely changed technological, economic and social environment, for which it appears no longer fully adequate, at least in certain regards.

First of all, the two decades since the adoption of the eCommerce directive have been characterised by a real explosion of multiple new information and communication technologies. On the hardware side, computational power, storage power, and communication capacities have grown exponentially. A global ICT infrastructure has emerged that links all kinds of information processing devices. This infrastructure todays connects not only 5 billion people, but also 30 billion automated devices, which incessantly create, transform and exchange information, interacting with humans and with other devices.

On the software side, a vast range of new systems has been developed, which exploit the newly available hardware, communication and data resources. In particular, cloud computing infrastructures, online platforms and personal digital devices have enabled a vast array of new services, both for businesses and for consumers: search, advertising, data analysis, optimisation, etc. A key role is currently played by AI[2]. In particular, machine learning technologies are using the computing resources today available —including vast computer power and huge data sets— to provide multiple new applications. The allocation of tasks between humans and computers is changing, as AI is entering multiple domains, sometimes substituting human intervention, and most often enabling new combinations of human and artificial capacities. No longer do computers only deal with executive and routine tasks, all creativity and discretion being preserved for humans: AI system, while not possessing human understanding and general intelligence, have the acquired the ability of executing a number

[1] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'). O.J. L 178/1 (2000).

[2] For a condensed introduction, see Sartor and Lagioia 2020, Section. 1

of tasks at human or even superhuman level.[3] This aspect, as we shall see, also concerns the management of online platform and in particular content moderation.

Concerning the economic environment, some of the start-up companies of 20 years ago have become large successful global companies, now drawing on huge technological and financial resources. A few such companies —the leading examples being the so called GAFAM, Google, Amazon, Facebook, Apple, and Microsoft— have acquired vast market power, which enables them to obtain large profits. Their success is based on technological and organisational excellence, but also on quasi-monopoly positions in key ICT services. This monopoly position is dependent on well-known aspects of the digital economy. First comes the so-called network effect: the greater a network, the greater the benefit that each individual can extract from it, due to the larger opportunities to interact, and to offer and access goods and services. Then there is the fact that in the digital domain development costs (the cost of developing software and creating a hardware infrastructure) are high, but marginal costs (the cost of serving an additional user) are low. Finally, automation and in particular AI facilitates scalability, as automated solutions can be distributed to all users, and centralised decision-making can profit from the automated analysis of vast amounts of data.

Finally concerning the social environment, a vast domain of human interactions has moved online. Access to information, the performance of work, and social interactions are today often mediated, prevalently or exclusively, by digital infrastructures: these activities have become dependent on such infrastructure and are framed by the affordances and constraints that are provided in that infrastructure.[4] This migration has recently accelerated in the context of the COVID 19 epidemic, which has forced all citizen to rely on the digital infrastructure for most of their needs and activities

In this new technological, economic and social context a key issue concerns the regulation of digital services, in particular, those that fit the category of online platforms, namely, those digital services whose purpose is to facilitate interactions between their users (whether firms or individuals), via the Internet.[5] Online platforms have become key "infrastructures"[6] for the activities of their users. While enabling user's activities and indeed providing multiple affordances for the exercise of such activities, platforms direct and constrain users engagements, in consideration of the economic and other purposes pursued by the platform's owners. For instance, the leading platforms that enable users to share content online (such as Facebook or YouTube) obtain their revenue though advertising directed at their users. Therefore, they have an interest in keeping users on the platform, in order to expose them to advertising and they have an interest in collecting data on users, in order to effectively target them with ads (and use the data for further commercial purposes). To keep users online, a platform may incentivise the production of, and access to, the kind of content, connections

[3] See the White Paper on AI by the European Commission (2020). On how the allocation of tasks changes due to AI, see Brynjolfsson and McAfee (2014).
[4] Among recent work that addresses the interaction between technology, socio-political change, and legal regulation, see Cohen (2019).
[5] For a discussion on the concept of a platform, see OCDE (2019), which adopts however a more restricted notion, requiring the participation of different kinds of users.
[6] Following Frischmann (2012), an infrastructure can be viewed as is a resource such that: (1) it may usually be consumed non rivalrously, (2) Social demand for it is driven primarily by downstream productive activity that requires the resource as an input; and (3) it may be used as an input into a wide range of goods and services. This description applies to roads, bridges, airports, parks, and schools but also the environment, telecommunications, and computing resources.

and interactions in which its users are most interested. Moreover, the platform itself may cultivate users' interests (or even their quasi-addition) for what is provided by the platform [7].

As the Internet has grown, online platforms have progressively expanded their scope and their diversity: search engines and portals have provided access to the information resources available on the web; online marketplaces have supported the exchange of physical and digital goods; online repositories have enabled the sharing of multimedia contents; social networks have supported all kinds of online interactions; collaborative platforms have facilitated the creation of digital content, service platforms have managed and coordinated the delivery of a vast range of online and offline services such as short-term rentals, ride-sharing, pet care, food delivery, dating, etc.

While the scope of platform services has expanded to all kind of commercial and non-commercial interactions, in this report we shall focus on a specific platform service, namely enabling the creation of, or access to, user-generated content. This is the main goal of widely used platforms (such as social networks, wikis, and search engines), but online sites having other main purposes may also include this function. For instance, online marketplaces may host users' reviews or online newspapers may host user content.

Platforms for user-generated content —in short content-platforms— contribute to the fulfilment of legitimate individual interests and rights as well as of important social values. They enable their users to express themselves, to create, transmit or access information and cultural creations, and to engage in social interactions. Society profits from the extended interactions between individuals: users' discussion contribute to public debate and democratic engagement while their cultural activities provide for the creation and increased accessibility of information goods.

This remains true today, even though concerns are increasingly being raised, and rightfully so, on the extent to which online platform s—apart from contributing to civil and productive interactions between people, to the search for truth, and the formation of reasoned public opinions— may also provide opportunities for harmful user behaviour: uncivility and aggression in individual exchanges, disinformation in the public sphere, sectarianism and polarisation in politics, as well as illegality, exploitation and manipulation.

To prevent or mitigate such harm, *moderation* is needed, namely the active governance of platforms, meant to ensure, to the extent that it is reasonably possible, the productive, pro-social and lawful interactions of the users. The enforcement of mandatory legal rules cannot be a substitute for moderation. Not only the scale of online sharing, but also the need to provide diverse forums, require an indirect approach for governing user's behaviour on content-platforms: in addition to regulating users' behaviour, the law should direct moderation in such a way that it effectively contributes to preventing harm and implementing legal values. As moderation is being increasingly automated, or supported by automated tools, a key aspect of a regulation of moderation concerns addressing the opportunities and risks that come with such tools, especially when they rely on AI.

The present report addresses a key aspect of automation in moderation, the adoption of automated filters meant to exclude or classify user-generated materials.

Section 2 will set the stage for the analysis of algorithmic filtering, by introducing the concept of moderation and the role of automated filtering in moderation.

[7] See Courtwright (2019, Ch 7).

Section 3 will examine the main technologies for algorithmic moderation and filtering.

Section 4 will discuss keys issues concerning the regulation of filtering.

Section 5 will introduce some policy options in the context of the Digital Service Act.

2. MODERATION AND FILTERING OF USER-GENERATED CONTENT

Here we will set the stage for the analysis of the technologies and regulation of algorithmic filtering. In Section 2.1. online moderation will be introduced; in Section 2.2 the concept of filtering will be analysed, in Section 2.3 the regulation of filtering will considered in general terms; in Section 2.4 the normative framework provided by the eCommerce Directive and subsequent development will be described.

2.1. Moderation in online communities

In this section we shall examine some issues that may emerge in online communities, to which moderation may provide a remedy

2.1.1. Information goods and information bads in online communities

The dynamics of on line communities have shown that the Internet does not work as a perfect "marketplace of idea," where true claims, valuable ideas, and pro-social content ("goods") naturally tend to prevail over the competition of falsities, bad memes, and socially obnoxious content.[8] It is true that in many cases wrong ideas and obnoxious behaviour ("bads") can be effectively met with appropriate criticism,[9] but this is not always the case. On the contrary, certain instances of online speech not only may negatively affect other users, third parties, or society at large, but may also impair the functioning of online communities.[10]

First of all, *excessive* user activity can overwhelm the technical capacity of a platform, generating *congestion*, and consequently, the platform's inability to timely respond to the requests of all its users. Congestion may also result from malicious attempt to flood the platform with access requests (so called denial of service).

Secondly, *worthless* users' activity produces *cacophony* or pollution, that is, it may flood the platform with redundant or irrelevant content, namely, content that has no significance to the other users, or in any case a significance that is lower than the attention and effort that is required to access and read it. Spam —understood as any kind of unsolicited and unwanted messages, typically but not only, for commercial purposes—highly contributes to cacophony.

Thirdly, obnoxious user activity leads to *abuse, i.e., to* the platform being used to deliver "information bads", rather than "information goods", i.e., content that has negative social value, that does more harm than benefit. In particular, *harassment*, consist in producing or spreading information that negatively affects particular individuals or groups, without adequate justification. Activities such as bullying, revenge porn, and hate speech fall into this notion of harassment, broadly understood. The term *trolling* too is used to cover instances of harassment, especially when systematic or intentionally meant to offend and intimidate, or to disrupt a community. Online abusive behaviour can consist in informational behaviour that is unlawful or anti-social in various respects: it may involve violating the privacy or reputation of participants or third parties, promoting violence, racism, or terrorism, sharing violent or child pornography, etc. Abuse, broadly understood, also consist in the violation of any kind of legally

[8] The notion of a marketplace of ideas, made famous by John Stuart Mills, gained currency in the law owing to the US Justice Oliver Wendell Holmes, who in his dissenting opinion in Abrams v. United States (1917) argued that the "ultimate good desired is better reached by free trade in ideas -- that the best test of truth is the power of the thought to get itself accepted in the competition of the market.

[9] This idea was affirmed by another famous US justice Louis Brandeis, who in Whitney v. California 1927 affirmed that against falsehoods and fallacies, "the remedy to be applied is more speech, not enforced silence."

[10] See Grimmelmann (2015).

protected rights, including those rights whose legitimacy and scope are controversial, such as intellectual property rights.

Note that abusive content is not limited to what is prohibited by the law: the concept of "information bads" needs to be contextualised to the participants in an online community, to their interests and expectations. For instance, the distribution of certain images of nudity may be permissible by the law, but still negatively affect communication within certain online communities[11]. Similarly, the use of uncivil, rude, or aggressive language (which does not reach the threshold required for slander, libel, or prohibited hate speech) may be deleterious to the smooth interaction within many online communities. The same applies to images or videos depicting cruelty against humans or animals, self-harm, or generally any content that is violent or disgusting. The fact that an item of speech is legally permitted, and rightly so, as a matter of general laws, does not entail that distributing this item is valuable in the context of a particular community, or that its distribution in that community is beneficial to society. Stricter standards than legal prohibitions may be applied, and are indeed applied, by moderators of online communities. By distinguishing what is prohibited by general law from what is disallowed within platforms and communities, according to their policies, we can reconcile the need to preserve freedom of speech and at the same time provide for friendly and non-abusive online environment. In fact, if it is true that, as George Orwell famously said "'If liberty means anything at all, it means the right to tell people what they do not want to hear,"[12] it is also true that online users need to be provided with environments free from abuse and aggression, and that pro-social content and exchanges need to be supported and encouraged.

Online platforms specify in general terms what counts as unwanted speech in their terms of service, namely the unilaterally predisposed contracts that users accept to use the platforms. According to such contracts, users commit themselves to comply not only with the law, but also with the requirements establishes by the platform itself.[13]

For instance, Article 5 of Facebook's policy prohibits the posting of content "that infringes or violates someone else's rights or otherwise violates the law", but also specifies that Facebook can remove any content "if we believe that it violates this Statement or our any policies." In particular Facebook policy prohibits bullying, harassment, intimidation, and nudity, as well as the use of Facebook 'to do anything unlawful, misleading, malicious or discriminatory'. Note that, according to these rule, Facebook is empowered to remove content, not only when it violated the law or Facebook's policies, by more broadly, when Facebook believes that this is the case: only removal of content in bad faith (i.e., the removal of content that Facebook does not believe to be infringing) would be an abuse of the removal power, according to the policy.

In Twitter's terms of service, the power to remove content is affirmed in even broader term. Article 8 in Twitters' terms of service[14] reads as follows "We reserve the right at all times (but will not have an obligation) to remove or refuse to distribute any Content on the Services, to suspend or terminate users, and to reclaim usernames without liability to you"

2.1.2. The need for moderation

To maintain the benefits provided by users' productive engagement in online content-sharing platforms, while addressing the negative impacts it may have on other members of a platform,

[11] See Facebook community standards, available at https://www.facebook.com/communitystandards/.
[12] Orwell [1945] 19762
[13] See Leerssen (2020).
[14] https://twitter.com/en/tos/previous/version_8

on third parties and on society, a governance mechanism is needed, usually referred to as "moderation". Unless moderation facilitates cooperation and prevents abuse, online communities tends to become dysfunctional, victims of "spammers, vandals and trolls". [15]

The practice of moderation may be inspired by multiple goals, which may ultimately pertain to socio-political values or to economic profit. For instance, moderation on a non-profit platform such as Wikipedia is directly inspired by the need to deliver a valuable and unbiased online encyclopaedia, which is meant to make knowledge accessible to everybody. Moderation in a for-profit platform, such as Google, Facebook or Twitter may also be inspired by social values —protection from humiliation and abuse, free exchange of opinions, production of knowledge and pro-social interactions— as well as by the economic objective of attracting users with an environment that suits their preferences and expectations, so that they may be targeted with advertising (which is the main source of profit for such platforms). The two goals are not necessarily in conflict, as long as what users want is exactly an environment where they can freely communicate and be at ease, without experiencing aggression, abuse, or be exposed to content that they prefer to avoid. Diversity in online communities, also with regard to what is considered appropriate may allow users find places that meet their different normative expectations, as long as such expectations are consistent with legal requirements.[16]

In conclusion, moderation is needed to prevent and mitigate harm and support productive and pro-social behaviour, and this may remain the case even when platforms and moderators are ultimately motivated by economic goals. What matters is that such economic goals are achieved by providing users with the online environments that respects the law and corresponds to user preferences.

However, not in all cases engagement in managing and directing users' contributions and interaction aims at pro-social goals. Managers of sites devoted to the furtherance of racism, political violence, or terrorism may intentionally favour —by promoting or demoting users' contributions or by commenting on them —the delivery of obnoxious and illegal content, stoking intolerance and hatred, excluding alternative points of view, harassing unwanted participants, etc. Such "moderators" rather than being "good Samaritans", who prevent or remedy harm, act as "bad Samaritans", who facilitate harmful and antisocial behaviour.[17]

2.1.3. Dimensions of moderation

In the following we shall speak of platforms and communities by way of a very general term. By a *content-platform*, we generally mean any infrastructure that enables user to access and share content. This includes content-sharing repositories (such as YouTube), social networks (such as Facebook), wikis (such as Wikipedia), search engines (such as Google), blogs, etc.

The set of all users in a platform (e.g., the set of all users of Google, Facebook, or Reddit) constitute a community that shares the use of the platform and the access to the materials provided by the platform. The community encompassed by the whole platform may that be split into further smaller communities, each involving particular users, to achieve their specific goals. Each of these communities may be subject to the moderation mechanisms established

[15] Grimmelmann 2015, 44
[16] See Klonick (2018).
[17] See Citron and Franks (2019).

for the entire platform, as well as those that have been established for the particular community, usually by the users who took the initiative to create it.

Before moving into the analysis of filtering, it may be useful provide an account of the ways in which moderation may in general be exercised. Moderation has different dimensions, (*a*) the participation of users, (*b*) the organization of user generated content as well as (c) the definition and promotion of users' norms. The *participation* of users in an online community is governed through mechanism for *inclusion* (the community may be open to anybody, or there may be a more or less selective admission process) and for *exclusion* (typically as a consequence of abusive behavior).

The *organization* of user-generated content, namely *content-moderation* strictly understood, involves taking different actions:

- *Rejection/admission*, to which we shall also refer as "filtering out" and "filtering in", the exclusion or inclusion of content in the platform; exclusion may include permanent deletion, or temporary blocking;
- *Editing*, i.e., altering content, which may consist in the mere correction of typos, but also in more substantial changes such, as improving style or removing offensive language.
- *Commenting*, i.e., adding to an item of content further information, such as criticisms, endorsements, links to contextualize it, etc.;
- *Prioritizing*, i.e., promoting or demoting access to content, e.g., presenting it as at the top (or bottom) of searches, or including it in (or excluding it from) the lists of most relevant/interesting content;
- *Synthesizing*, i.e., extracting and combining information from different users' contributions.

Finally, concerning the *adoption of shared norm*, moderation involves communicating norms to the users, and encouraging them to learn, endorse and apply such norms, to comply with them and to react to other people's abuses. The communication and consistent application of norms is a key aspect of moderation. Unless a community knows what norms are applied and shares such norms, moderation will come across as the exercise of an arbitrary and capricious power.

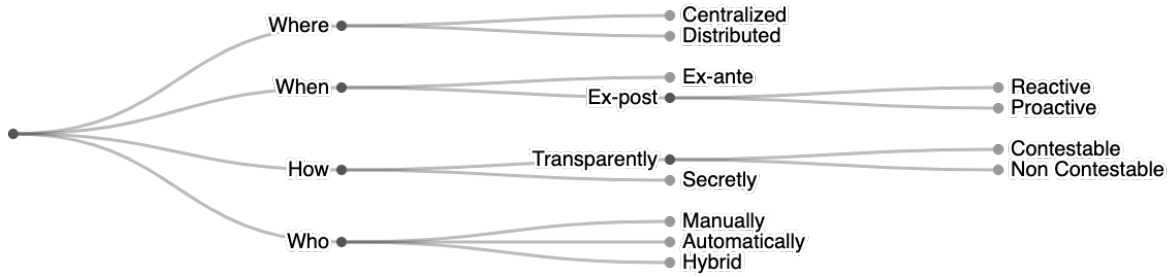
2.2. Filtering

Let us now focus on content moderation, and in particular, on filtering broadly understood, i.e., the classification of user-generated content for the purpose of determining its accessibility: excluding it from all users (or members of a specific community), making it inaccessible to certain categories of them (e.g. minors), upgrading or downgrading its priority.

2.2.1. Modes of filtering

Online filters may be classified according the modalities indicated in Figure 2.

Figure 2. Modalities for content moderation (filtering)



With regard to *where* filtering is applied and thus the type of organizational architecture, which is employed, we can distinguish between

- *Centralized filtering*, which is applied by a central moderation unit according to uniform policies, that apply across an entire platform
- *Decentralized filtering*, which involves multiple distributed moderators, operating with a degree of independence, and possibly enforcing different policies on subgroups within the platform.

For instance, a vast content sharing platform such as YouTube tends to apply its policies globally. On the other hand, Reddit provides for a minimum of common guidelines, and allows for different communities, each governed by the standards set by its own moderators:

> The culture of each community is shaped explicitly, by the community rules enforced by moderators, and implicitly, by the upvotes, downvotes, and discussions of its community members. Please abide by the rules of communities in which you participate and do not interfere with those in which you are not a member.[18]

With regard to *when* filtering is applied, we can distinguish between:

- *Ex-ante filtering*, which is applied before the content is made available on the platform; and
- *Ex-post filtering*, which is applied to the content that is already accessible to the platform's users.

Ex-ante filtering may be found in small moderated groups (e.g. a comment section in a blog), where a moderator reads all posts before making them available to the readers. It may also be adopted in those instances of filtering that are fully automated, such as the detection of certain clear violations of copyright (the posting of a full copyrighted file). Ex post filtering has the advantage of not delaying the publication of materials and not overburdening moderators, the obvious disadvantage is that harmful material remains available until it is removed.

With regard to the *initiative* on filtering we can distinguish between:

- *Reactive filtering*, which takes place after an issue with an item has been signalled by users or third parties.
- *Proactive filtering*, which takes place upon initiative of the moderation system, which therefore has the task of identifying the items of content to be acted upon.

[18] Reddit Content Policy, see https://www.redditinc.com/policies/content-policy.

In reactive filtering the identification of the items to be assess for filtering is done by users, making complaints or flagging items, while the decision on what items to exclude is taken by moderators. Obviously, the distinction between reactive and proactive filtering only concerns ex-post filtering; ex-ante filtering is necessarily proactive.

Reactive filtering has until recently been the only approach available in large communities, where the size of daily uploads would make human analysis of all posts impossible. This has led to the development of mechanisms of notice and take down in many online communities. In the US a notice and take down procedure is provided by the 2000 Digital Millennium Copyright Act: content is blocked when the alleged rightsholder sends a notice, but the alleged infringer is informed, and the content is made accessible again if the alleged infringer resists and the rightsholder does not initiate a lawsuit. In the EU a kind of notice and take down procedure has been established after the decision in the Google Spain case: [19] if a data subject requires that certain posts be delisted (i.e., that they not be presented in response to searches that use the data subject's name), the search engine is obliged to comply with the request, unless prevailing reasons exist for allowing such search results. In this procedure, however, no voice in the delisting procedure is given to the publishers of the content being delisted.

With regard to *how* i.e., to the *openness* of the filtering, we can distinguish:

- *Transparent filtering*, which provides information on the exclusion of items from the platform; and
- *Secret filtering*, which does not provide any information about the exclusion.

Transparency can come in degrees. Moderators (be they humans or automata) may just report the fact that an item has been excluded, or also on the reasons for the exclusion. This information may be provided to the uploader of the excluded content, or also to the community as a whole.

When the filtering is open, a further distinction can be drawn between:

- *Contestable filtering*, in which the platform provides uploaders with ways to challenge the outcome of the filtering, and to obtain a new decision on the matter; and
- *Non-contestable filtering*, in which no remedy is available to the uploaders.

A great variety of approaches exists concerning the extent to which filtering decision are made transparent and contestable. In general decisions to exclude users from a platform can usually be appealed by users, while for decisions to block or remove specific content often no procedure for redress is available.

With regard to *who* is doing the filtering —i.e., who analyses the content, classifies it as deserving inclusion or exclusion, and proceeds accordingly— we can distinguish between:

- *Manual filtering*, performed by humans;
- *Automated filtering*, performed by algorithmic tools;
- *Hybrid filtering*, performed by a combination of humans and automated tools.

For instance, individual bloggers usually engage manually with the contributions posted on their blogs and decide what to maintain or delate. On the other hand, fully automated filtering

[19] Judgment of 13 May 2014, Google Spain SL and Google Inc v. Agencia Española de Proteccíon de Datos (AEPD) and Mario Costeja Gonzalez, Case C-131/12.

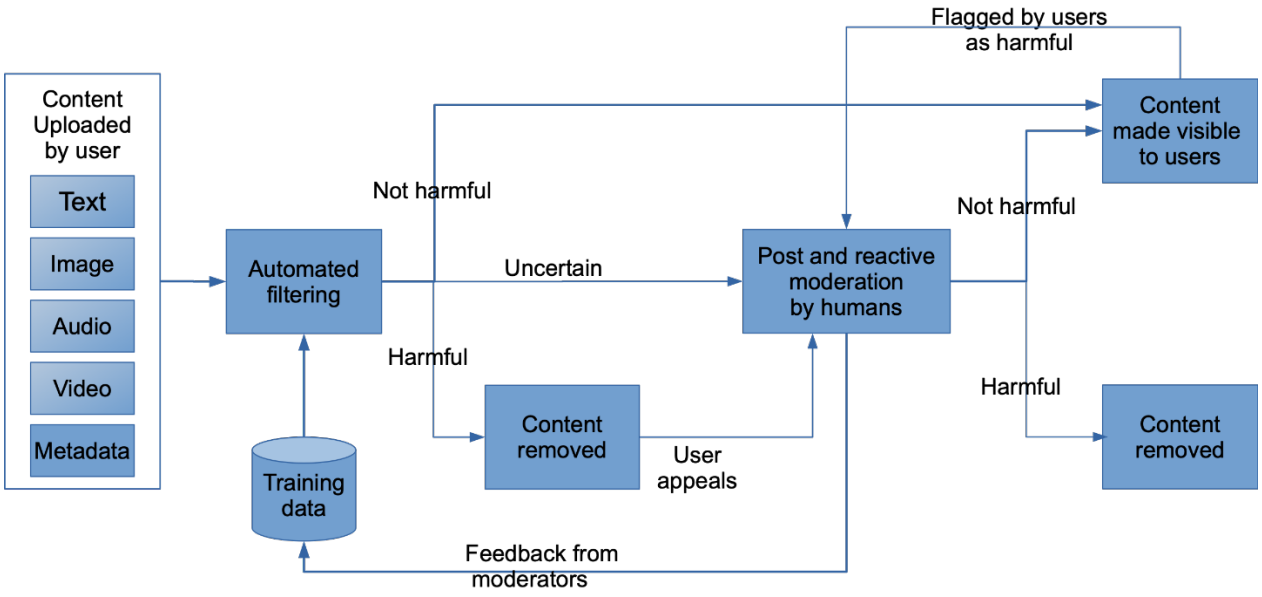
is often used in vast platforms, in particular to prevent the posting of copyrighted works. In other domains, such as hate speech, a hybrid approach may be adopted, typically identifying relevant cases automatically, and referring them to humans for a decision. In big platforms filtering is entrusted to a complex socio-technical system, including a multi-stage combination of human and machines that interact in complex ways.

2.2.2. Filtering in the context of socio-technical systems for moderation

Figure 3, shows the ways in which automated and human moderation can be integrated in a hybrid approach. Automated filters can take a first step in processing the user-generated data. The filter itself may classify certain items as definitely harmful, which leads to the automated removal of such items, or as definitely not harmful, which leads to the items be made available online.

The filter may also classify certain items as uncertain, i.e., as potentially harmful, in which case the items are subject to human review, according to which they can be blocked or published. An item may be subject to human review also subsequently to a user's complaint against automated removal. Human review not only corrects specific past outcomes of automated filtering, but it can also provide feedback to improve future performance. In particular, if a machine learning approach is adopted, the new human assessments may be added to the training set of the filter, so that the automated filtering system can update its model and learn to address accordingly the new cases.

Figure 3. The integration of automated filtering and human moderation



In conclusion, an automated filtering system should not be considered in isolation, but rather as a component of a socio-technical systems that includes technologies and humans as well as the organizational structure (guidelines, responsibilities, allocation of tasks, definition of workflows, etc.) that integrates them. Users may also play an important role, not only as originators of the content, but also when flagging harmful content, or when challenging the decisions of automated systems or human moderators. Thus, a filtering system should be judged in context, also taking into account the remedies available against its failures.

The distribution of unlawful or harmful content online —or the removal of legitimate and beneficial content— involves a failure of the platform as a whole, and this failure may involve the shared responsibility of users, moderators, and automated tools.[20] However, not necessarily all of these components are to blame for every failure, and in some cases none of them is to blame: given the difficulty of the task at hand, false negatives or false positives may occur even though human moderators carefully engaged with their tasks and state-of-the-art technologies are appropriately used.

2.3. Filtering in the secondary regulation of online speech

In this section we shall first introduce secondary regulation of online speech and then apply this idea to the regulation of filtering.

2.3.1. Secondary regulation of online speech

The platforms provided by digital companies enable users to upload content and make it available to online communities. However, as noted above, online content may have different legal and ethical quality. Certain content may be unlawful and severely infringe on the rights or individual and on social values. Other content may be lawful but inappropriate: while not being forbidden by the law it may be perceived as inconvenient by other users, or at any rate it can be detrimental to the interaction taking place. Finally, other items may be both lawful and appropriate, pertaining to the exercise or individual rights or to the pursuit of valuable social goals, and contributing to the interaction in the community concerned. We have also observed that moderation is needed to prevent online communities from become dysfunctional, in relation to the interests of participants, to the rights of third parties, and to social values. In particular, preserving lawful and pro-social interaction may require filtering out content that is unlawful or inappropriate.

In illustrating the issues involved in the regulation of filtering, it is useful to refer to the distinction between primary and secondary regulation of online speech.[21] The law may indeed act in two ways to prevent "information bads" or mitigate the harm they may cause.

On the one hand, the law provides for "primary regulations", i.e., rules and sanctions aimed at to the platform users, specifying what may or may not do, with what legal effect, and what the sanctions are. Participants in online communities are indeed subject to multiple legal rules, such as those addressing civil liability (torts), data protection, intellectual property, defamation, hate speech, holocaust denial, revenge porn, and cyberbullying. The enforcement of such rules against the individuals concerned is necessary to ensure lawful and responsible behaviour, but its effectiveness is problematic due to factors such as the vast size of many online communities, the difficulty in tracing the identity of participant, and the cost and uncertainty of legal proceedings.

Moreover, the line between unlawful behaviour, and behaviour that while lawful is inappropriate or anti-social is in many domains fuzzy: distinguishing harsh criticism from defamation, or radical political opinions from expressions of hatred and racism or incitement to violence may require contextual and nuanced analysis. Additionally, protecting individual freedoms, in particular freedom of expression, as well as supporting as democratic dialogue require that governments refrain from unnecessary limitations of speech, even when the

[20] Helberger et al (2028).
[21] See Balkin (2014).

information or opinions expressed may be me false, unjustified, unethical, or otherwise anti-social. Indeed, according to Article 10 of the European Convention on Human Rights, freedom of expression includes the "freedom to hold opinions and to receive and impart information and ideas without interference by public authority", a freedom that can be limited only when this is "necessary in a democratic society".[22]

Besides directly addressing online speakers, the law can affect online communities through secondary regulation, namely, the legal rules meant to induce providers of digital services to influence the activity of their users. In the latter case the direct target of the regulation are the intermediaries, but the final target are the users. Since human action, and in particular communication, today takes place through increasingly complex and flexible socio-technical systems, the behaviour of the users of such systems can often be most effectively regulated by directing legal norms towards the intermediaries who control those systems. Intermediaries are ideally induced to shape their services in such a way that the opportunities for unlawful or inappropriate user behaviour are restricted, and the negative consequences of such behaviour are mitigated, while enabling and encouraging beneficial user engagement Secondary regulation of intermediaries can take different forms. Intermediaries may be requested to adopt specific actions that prevent or mitigate unlawful actions by their users (security measures, antispam filters, etc.). They may also be required to monitor the activities of their users to facilitate law enforcement (as in regulation imposing data retention). Finally, intermediaries may be subject to secondary liability for the unlawful actions of their users — and in particular to the obligation compensate harm to third parties— to incentivise them to take whatever initiative is within their power to prevent or mitigate such actions.

Secondary regulation aimed at platform owners may use any mix of the just mentioned approaches, to influence the way in which moderation is performed. Such regulation may favour good and effective moderation, or it may lead to unwanted and unexpected results, which may include the elimination of useful and legitimate content (so-called collateral censorship), or, on the other hand, the proliferation of useless or abusive content.

2.3.2. The regulation of filtering

This general considerations developed in the preceding section can be applied to online filtering, the regulation of which indeed pertains to what we have called secondary speech regulation.

First of all, we have to observe that platform owners by filtering out certain content "regulate" the activity of their users, i.e., they influence their behaviour: uploaders are disabled from making available the content that is filtered out, and readers are consequently disabled from accessing the same content. Moreover, platforms may provide for additional measures against the user (in addition to the filtering out content), such as the expulsion of users who repeatedly infringe legal requirements or the platform's policies, posting unlawful or harmful content.

The law may regulate filtering by requiring platforms to filter out certain content, under certain conditions, or rather by prohibiting certain content to be filtered out. Such regulation qualify under the concept of secondary regulation being introduced above, as they ultimately aim to influence users' behaviour, by imposing obligations on platforms. Accordingly, the success of

[22] In the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

secondary regulations of automated filtering has to be assessed according to the extent in which such regulations lead providers to influence user behaviour in such a way to achieve two parallel objectives: (a) preventing, limiting and mitigating as much as possible the individual and social harm that can be caused by unlawful or inappropriate online content and activities, while (b) allowing and even facilitating the delivery of beneficial content as well as free and civil on line interaction. Moreover, the focus of the regulation must go beyond the platforms and look rather to the complex ecology of the so-called infosphere,[23] were many different platforms exist, responding to different audiences and needs.

Under EU law, a third layer of regulation can be identified, which we may call "tertiary regulation" of online speech. These are the rules that are meant to regulate the activity of regulators, who regulate providers, who regulate users. An important example of such "tertiary" regulation is provided by, Article 15 of the eCommerce Directive, which, as we shall see in the next paragraph, prohibits national authorities from imposing upon providers "general obligations to monitor content."

2.4. Moderation in the eCommerce Directive and beyond

In this section we shall consider the way in which moderation is addressed in the eCommerce directive and in the subsequent evolution of the EU law.

2.4.1. The eCommerce Directive's approach to user-generated content

The eCommerce Directive approach to user-generated content —and more generally to secondary regulation— expresses a hands-off approach. The relevant rules are indeed those which exempt internet providers from liabilities for the unlawful activities of their users. More to the point, the Directive specifically addresses three kinds of services:

- mere conduit, i.e., transmission over a communication; network of information, or access to a communication network;
- caching, i.e., automatic, intermediate and temporary storage of transmitted information, to increase efficiency; and
- hosting, i.e., storage of information.

The Directive states that such intermediaries are exempted from secondary liability when providing these services, this under the conditions specified in articles 12-15.

With regard to mere conduit, Article 12 indicates that providers are exempted when they do not initiate the transmission, do not select the receiver of the transmission, and do not select or modify the information contained in the transmission.

With regard to hosting, Article 14 specifies that host providers are exempted from liability as long as they do "not have actual knowledge of illegal activity or information" and, as regards claims for damages, are "not aware of facts or circumstances from which the illegal activity is apparent". Providers who obtains "such knowledge or awareness" are still exempted from liability if they act "expeditiously to remove or to disable access to the information".

A key provision on the secondary liability of provider is contained in Article 15: intermediaries may be ordered, by competent authorities, to terminate or prevent infringements by their

[23] Floridi (2013).

users, but they may not be put under any "*general obligation* to monitor the information which they transmit or store" nor to "actively to seek facts or circumstances indicating illegal activity".

Thus, on a literal reading of the eCommerce Directive it seems that automated filtering to detect unlawful content cannot be required by the law. Indeed, a legal norm requiring platform to filter on all incoming information would establish a general obligation to "monitor the information which they transmit or store," in violation of the prohibition in Article 15. If such an obligation cannot be imposed by law, providers cannot be deemed responsible for failing to do would be required by such an obligation. Thus, they should not be considered responsible for not reacting to unlawful content, when awareness of such content could only be obtained through automated analysis of incoming or posted content.

The main rationale for such a hands-off approach to the role of platforms in the distribution of abusive content is twofold: on the one hand, supporting the development of the internet economy (by shielding operators from the cost of preventive measure and liabilities); on the other hand, protecting users' freedoms, as the fear of incurring in liabilities could lead providers to censor user's content and block their activities, even when contents and activities are lawful and socially beneficial. This hands-off approach was also supported by a pragmatic ground: it was assumed that the size and speed of online user activity makes it impossible to effectively monitor online content and activities.

2.4.2. Later trends

A hands-off approach to user-generated content has come more and more under attack in recent years, as the amount of online information and activities has boomed, and the individual and social costs of unlawful and inappropriate online behaviour have become apparent. The very premises on which the hands-off approach was based have been questioned.[24] Is it still true that internet companies do have the financial resources needed to pay damages to parties injured by user activities enabled by those very companies? Why shouldn't they be asked to pay for the negative externalities of their business model? Given their financial and technological powers, can they still claim to be unable to exercise effective control over online content and activities? Can technological development, particularly in the domain of AI, make for efficient and cost-effective controls?

As noted above, the leading online intermediaries today are no longer small start-ups, facing strong incumbents with few resources and experimental technologies; they have acquired huge economic power, becoming dominant players in their respective markets; they have huge financial and technological resources at their disposal. For instance, Google now manages more than 90% of Web searches in Europe while Facebook handles more that 80% of social network usage. This dominance is reflected in changes in related markets, such as advertising: Google and Facebook attract about one-half of the expenditure devoted to online advertising, which now attracts the largest share of the total advertising expenditure (having overtaken TV advertising, while print newspaper advertising has collapsed). The economic power of the leading online intermediaries contributes to giving them the capacity to influence political decisions, through lobbying or by mobilising political opinion.

Another important development pertains to technologies for identifying and filtering out illegal content. While human monitoring of each item indexed by a search engine or made available on a large platform is unfeasible, software tools have become available that can

[24] See for US law Citron and Witte (2018), on EU law, Valcke et al (2017)

target illegal material with increasing accuracy. Such tools are far from perfect —they risk excluding a lot of legal and socially beneficial materials, alongside with illegal or harmful ones— but in some domains (e.g., recognition of unauthorised distribution of copyrighted works), they enable an effective and sufficiently precise control.

Some intermediaries are taking an increasingly active role: they contribute to framing the way in which third party content is created; they determine the way in which it is accessed, they combine further processing with that initiated by their users (e.g., indexing or linking). This is usually instrumental to the main function of the intermediary. For instance, social networks contribute to their mission —facilitating communication and social connection between their users— by providing templates through which users can prepare and organise the material they publish online; suggesting links to users having similar interests; presenting each user with the materials to which that user may be more interested, etc. Similarly, a search engine contributes to its mission —connect providers and users of online content— by indexing the content uploaded by content providers and presenting users with content that it likely to be relevant to them. A certain degree of initiative, or editorial discretion, is needed for a content intermediary to effectively exercise the function to "help end users locate and obtain access to content they find desirable"[25]. Intermediaries, however, also engage in activities that are not connected to their intermediation function. For instance, they may link advertising to queries or to content, and they process user data of their own initiative. They may also frame their services in such a way that —while still meeting to a sufficient extent the preferences of their users— they are geared towards other goals, commercial or not. This may concern favouring certain communications for private purposes (which may be unfair or even illegal, e.g., unfairly prioritising certain companies or services in search results) or also for social purposes that are endorsed by the intermediary.

Indeed, in some case intermediaries play an active role that assumes political significance; they undertake to control and manipulate certain interactions. For instance, Kent Walker, senior vice president at Google, claims that Google intervenes in the following ways to counter online extremism:[26]

- using technology to identify (and remove) extremist videos;
- employing experts (from NGOs) to make decisions;
- putting warnings and excluding monetisation, comments and endorsement from offensive but not prohibited videos (e.g., those being both inflammatory and religious or supremacist);
- redirecting potentially radicalised users to materials that can change their mind.

2.4.3. Legal responses to the new situation

As intermediaries frame and control the communications they enable, the law tends to use this capacity to regulate the activity of their users: the law establishes obligations and liabilities upon intermediaries, in order to induce them to prevent or restrain illegal or unwanted users' behaviour.

[25] Yoo (2012, Ch. 9).
[26] Walker (2017).

Particularly significant in this regard is a law enacted in Germany on 1 September 2017 called the Social Networks Enforcement Law (*Netzwerkdurchsetzungsgesetz*)[27]. Under this law, social media platforms will have to take down posts containing "obviously illegal" material within 24 hours of being notified of such posts. For less obviously criminal content, the compliance timeframe is seven days. If a platform repeatedly fails to meet those deadlines, it will be liable for fines of up to 50 million euros. At the EU level, various initiatives have been recently adopted that establish proactive duties for intermediaries:[28]

- The 2011 Directive on Child Abuse[29] provides for obligations to remove and block access against websites containing or disseminating child sexual abuses and child pornography.
- The 2017 Directive on terrorism provides for similar obligations against public online incitement to acts of terrorism.[30]
- The 2018 Directive revising the Audiovisual Media Service Directive includes new obligations for video sharing platforms to tackle hate speech and violence.[31]
- The 2019 Directive on Copyright in the Digital Single Market establishes that the liability exemption for host providers does not cover the unauthorised communication or making available to the public of material uploaded by their users, and establishes obligations for such providers (including best effort obligations to implement agreements with rightholders and remove and prevent access to works identified by them).[32]

Many judicial decisions require providers to actively counter illegal information or activities by their users, failing which these providers will incur liabilities. At the European level, we can mention the following.

- The 2014 Google-Spain decision of the European Court of Justice[33] requires search engines to comply with a person's requests to remove unwanted links to personal information from the results of searches made on the basis of that person's name.
- The 2017 Ziggo decision by the European Court of Justice[34] affirmed that "a sharing platform which, by means of indexation of metadata relating to protected works and the provision of a search engine, allows users of that platform to locate those works and to share them in the context of a peer-to-peer network" engages in the communication to the public of those works (and may consequently be subject to the corresponding sanctions).

[27] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken of 1 September 2017, (BGBl. I S. 3352).
[28] See de Streel et al (2019).
[29] Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA.
[30] Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA.
[31] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities.
[32] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.
[33] Google Spain SL and Google Inc v. Agencia Española de Protecciónde Datos (AEPD) and Mario Costeja Gonzalez, C-131/12
[34] Stichting Brein v Ziggo BV and XS4All Internet BV, C-610/15.

- The 2019 Glawischnig-Piesczek decision by the European Court of Justice affirmed the admissibility of injunctions ordering providers to remove and block unlawful information including any identical or equivalent information, possibly worldwide.[35]
- The 2015 Delfi decision by the European Court of Human Rights[36] upheld an Estonian decision punishing a journal that had failed to remove readers' online comments containing expression of hatred, even if there was no specific complaint.

2.4.4. Active and passive host providers

As the protection given to providers under the eCommerce Directive has been considered excessive, some moves had been taken to bypass the immunity so granted. In particular, it has often been argued that only "passive hosting" is covered by the concept of hosting, as used in Article 14 of the Directive. Accordingly, intermediaries that store and make accessible user-generated materials, but who also organise these materials, index them, link them to ads, remove unwanted items, etc., should not enjoy the protection that is granted to hosting services. This strategy is complemented by the view that toady the idea of passivity needs to be detached from the idea of automaticity: human intervention is no longer needed to make a service "active", since automated processing has become flexible and selective. As a consequence of this double move —considering that only passive intermediaries are protected, and that automated services may be non-passive— it has been denied that social networks (e.g., Facebook), content-sharing platforms (e.g., YouTube), and search engines (e.g., Google) fall under the protection of the eCommerce Directive. This approach has been followed by the case law of various European countries, though not by the EU Court of Justice, which still links the activity that excludes liability exemptions to knowledge of, or control over, specific items of content, as it stated in the 2011 L'Oréal vs eBay case:[37]

> Where, by contrast, the operator has provided assistance which entails, in particular, optimising the presentation of the offers for sale in question or promoting those offers, it must be considered not to have taken a neutral position between the customer-seller concerned and potential buyers but to have played an active role of such a kind as to give it knowledge of, or control over, the data relating to those offers for sale. It cannot then rely, in the case of those data, on the exemption from liability referred to in Article 14(1) of Directive 2000/31.

For our purposes, we need to consider whether, by proactively engaging in moderation, a provider plays an "active role of such a kind as to give it knowledge of, or control over, the data" uploaded by users, and would consequently be excluded from the protection by the eCommerce directive.

The key consideration against this idea, is that linking moderation to liability for unlawful user behaviour would incentivise providers to remain passive relative to online unlawfulness and abuse (in order to maintain their immunity), rather than actively intervening to counter them. This consideration inspired the so-called "good Samaritan" clause, included in the 1999 US Communication Decency Act, section 230, according which providers should not be considered liable for "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to

[35] Eva Glawischnig-Piesczek v Facebook Ireland Limited in. Case C-18/18.
[36] Delfi AS v. Estonia, no. 64569/09
[37] L'Oréal SA and Others v eBay International AG and Others. Case C-324/09, para 116.

be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected." This rule was enacted to counter the approach, adopted in some US judicial decisions at the time,[38] according to which providers who engaged in moderation were considered to exercise editorial control over user-generated content, and were consequently made liable for unlawful content posted by their users.

The Good Samaritan approach makes a lot of sense also in the EU. Host providers should not suffer prejudice—and in particular lose their immunities—for actively addressing online abuse. As we shall remark below, only a failure on their part to adopt reasonable measures to prevent user-generated harm to third parties should make them liable for such harm.

The Commission in its Communication of September 2017 on tackling online content[39], has indeed taken a positive position toward proactive moderation meant to address unlawful content. It has affirmed that

> Online platforms should, in light of their central role and capabilities and their associated responsibilities, adopt effective proactive measures to detect and remove illegal content online and not only limit themselves to reacting to notices which they receive. Moreover, for certain categories of illegal content, it may not be possible to fully achieve the aim of reducing the risk of serious harm without platforms taking such proactive measures.

According to the Commission, proactive measures adopted for this purpose should not deprive hosting platforms of the benefit of the liability exemption provided for in Article 14 of the E-Commerce Directive. Such measures

> do not in and of themselves lead to a loss of the liability exemption, in particular, the taking of such measures need not imply that the online platform concerned plays an active role which would no longer allow it to benefit from that exemption.

The Communication specifically addresses "detection and filtering technologies", as key tools to address unlawful online content:

> Given the volume of material intermediated by online platforms, as well as technological progress in information processing and machine intelligence, the use of automatic detection and filtering technologies is becoming an ever more important tool in the fight against illegal content online. […]

According to the Commission, the eCommerce provisions on liability "do not preclude the development and effective operation of technical systems of protection and identification and of technical surveillance instruments made possible by digital technology."

We agree with the approach adopted by the Commission. We would just add that legitimate proactive moderation may also address content which is not unlawful, but which is inappropriate or objectionable in relation to the public of the platform being considered.

[38] Such as, in particular, Stratton Oakmont, Inc. v. Prodigy Services Co, decided in 1995 by the 1995 by the New York Supreme Court.

[39] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms (COM(2017) 555.

2.4.5. Obligations to engage in active moderation

In some recent legislative instruments, active pro-moderation (and in particular filtering) is not merely considered as an opportunity for providers, but tends to become the subject matter of obligations.

The Audiovisual Media Services Directive,[40] at Article 28b (1), requires content providers to actively engage in moderations, in order to protect minors from content that may impair their development, and the general public from content that incites to violence or hatred, or whose dissemination is a criminal offence in connection with terrorism, child pornography, racism and xenophobia. According to the Directive, Member States have the obligation to ensure that video-sharing platform providers under their jurisdiction take appropriate measures to protect

> (a) minors from programmes, user-generated videos and audiovisual commercial communications which may impair their physical, mental or moral development in accordance with Article 6a (1);
>
> (b) the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter;
>
> (c) the general public from programmes, user-generated videos and audiovisual commercial communications containing content the dissemination of which constitutes an activity which is a criminal offence under Union law, namely public provocation to commit a terrorist offence as set out in Article 5 of Directive (EU) 2017/541, offences concerning child pornography as set out in Article 5(4) of Directive 2011/93/EU of the European Parliament and of the Council and offences concerning racism and xenophobia as set out in Article 1 of Framework Decision 2008/913/JHA.

A best effort obligation to engage in moderation, is also established under Article 17 (4) of the 2019 Copyright Directive[41], according to which online content-sharing service providers shall be liable for unauthorised acts of communication to the public, unless they providers demonstrate that they have:

> (a) made best efforts to obtain an authorisation, and
>
> (b) made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rightholders have provided the service providers with the relevant and necessary information; and in any event
>
> (c) acted expeditiously, upon receiving a sufficiently substantiated notice from the rightholders, to disable access to, or to remove from their websites, the notified works

[40] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), as amended by Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities.

[41] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

or other subject matter, and made best efforts to prevent their future uploads in accordance with point (b).

In particular, complying with Article 17 (b) may require, under today's technological conditions, the use of automated tools to detect copyrighted works unlawfully distributed, as specified by rightholders Complying with 17 (c) may require upload filters to exclude the re-publishing of unlawful content.

A key issue concerns determining the extent to which the obligations to be established by Member States according to the Audiovisual Media Services Directive and the Copyright directive are consistent with the eCommerce provisions, in particular with Articles 14 (liability exemption for host providers) and 15 (no general obligation to monitor). Article 28b (1) Audiovisual Media Services Directive explicitly specifies that the moderation measures to be imposed on providers should be "without prejudice to Articles 12 to 15 of Directive 2000/31/EC". The Copyright Directive at Article 17 (3), on the other hand explicitly derogates from Article 14 of the eCommerce directive, though not from Article 15.

2.4.6. Some open issues

The idea than online platforms should be required to take a more active role in moderation, so as to prevent abuse and unlawfulness raises some serious issues.

One issue is whether even smaller companies and start-ups have the financial resources needed to face liabilities and take preventive measures, and whether they have the technological capacities needed to this effect.

A second issue concerns the extent to which improvements in preventing and mitigating harm resulting from user activities may lead to interfering with lawful and useful user contributions, perhaps pertaining to the exercise of fundamental rights. Since filtering aims to identify instances of unlawful or inappropriate content or activity, such instance are "positives" for the purpose of the filtering (the possession the searched-for features, namely, unlawfulness or inappropriateness), while instances of lawful and appropriate content or activity are negatives. Unfortunately, for any given level of human and technological power deployed, increasing the so-called true positives (i.e., the amount of unlawful or inappropriate materials or activities identified as such) usually entails expanding the amount of false negatives (the amount of lawful and useful materials or activities that wrongly identified as being unlawful or inappropriate). Conversely, reducing false negatives entails increasing false positives (the amount of unlawful or inappropriate content that is wrongly considered to be lawful and appropriate). To give an example, hate speech base on ethnicity could be eliminated by blocking all content including words or phrases that may refer to ethnicity, but this would involve blocking any speech about ethnicity, even for the expressions that are meant to condemn racism and hate. Or similarly, all paedo-pornography could be eliminated by erasing all pictures that include uncovered parts of the human body, but this would eliminate much permissible expression including artistic or scientific content.

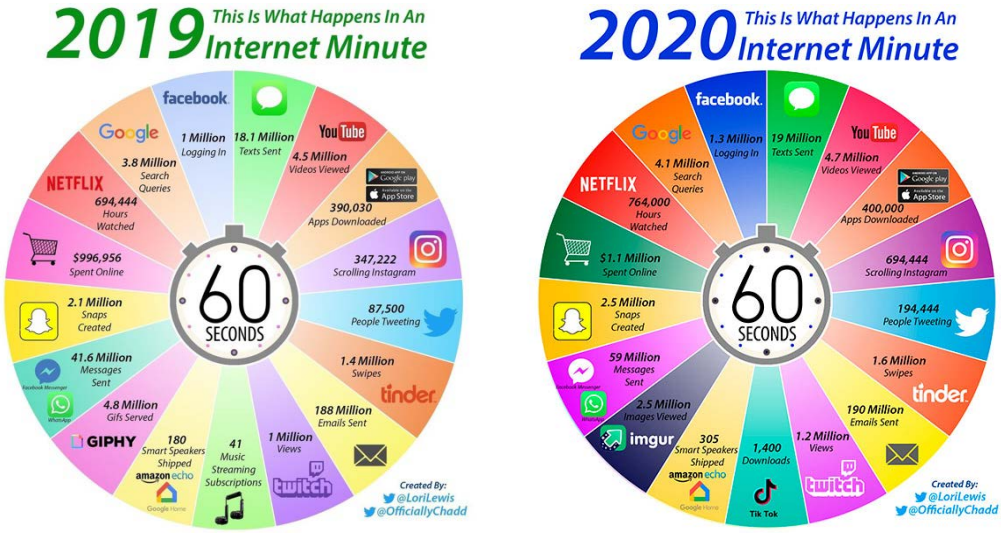
A third issue concern the interaction between controls meant to detect unlawful or inappropriate content or behaviour and civil liberties. Such controls necessarily involve data collection and surveillance of people's online activities, and consequently it involves an interference in the data protection and privacy rights of the individuals concerned. This raises the issue of the extent to which, on balance, the need to prevent online abuses justifies such an interference.

Finally, a key issue, which lies at the core of the present report, concerns the extent to which algorithmic moderation may integrate or even replace the activity of human moderators, providing solutions that not only are more cost-effective, but also better protect individual rights and legitimate interests as well as social values.

3. TECHNOLOGIES FOR FILTERING

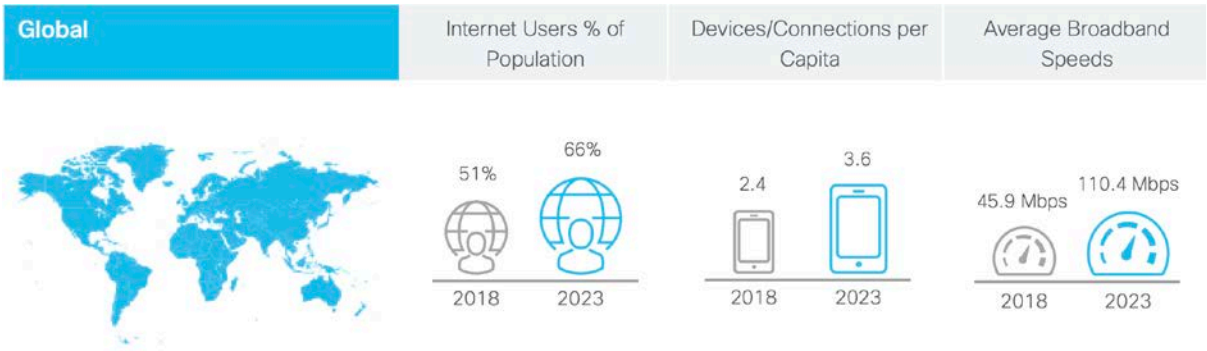
Every single minute an impressive amount of data is communicated over the Internet, as depicted by the infographic in Figure 4. The exchanged data increases every year. According to the infographics, the number of Facebook logins grows by about 20% as does the number of messages sent using Messenger or WhatsApp.

Figure 4 What happens in Internet in one minute



The same trend is also described in the Annual Internet Report by Cisco. In this extensive study, Cisco describes how both the number of users (Figure 5) and the traffic rate have grown and how they are predicted to grow over the next few years. This study highlights that the Internet becomes increasingly pervasive as the number of connected users increases: comparing statistics from 2018, the percentage of connected users is expected to grow by about 15% by the end of 2023 and the amount of devices per person by 50%, while the speed of Internet connections will more than double.

Figure 5. Global - Internet Users

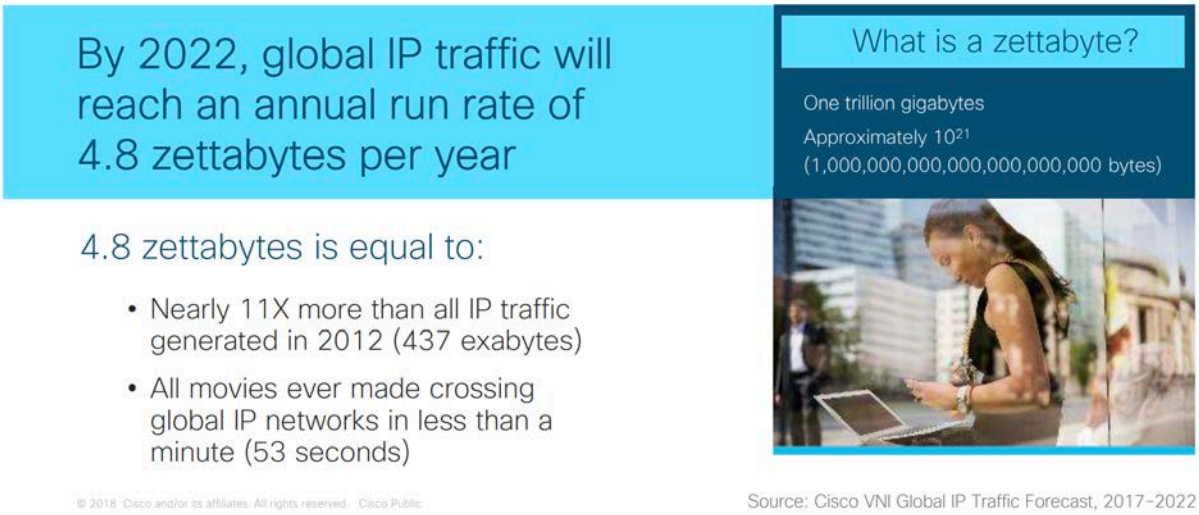


Source: Cisco Annual Internet White Paper 2020 [42]

[42] See https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

The Cisco report also makes projection concerning Internet Traffic, which by 2022 is expected to exceed by 11 times the total amount of traffic produced in 2012, as reported in Figure 6.[43]

Figure 6. Cisco Visual Networking Index (VNI) Complete Forecast Update, 2017-2022



Source: 2020 Cisco Annual Internet White Paper[44]

The sheer amount of information made accessible or available make it impossible to manually check each item. Therefore, automated filtering is increasingly used to classify content and take measures accordingly. Such measures may consist in excluding the content that is classified as unlawful or inappropriate, or in making it inaccessible to certain classes of user (e.g. minors).

The filtering may be complemented by a by subsequent human review, to be applied to all content classified as inappropriate, or only to instances where the classification is uncertain. This scenario applies to social network (e.g. Facebook, LinkedIn, Twitter, Redditt), a streaming service (e.g. Netflix, YouTube), messaging services (e.g. WhatsApp, Messenger, Hangout), and search engines (e.g., Google).

Different approaches can be employed based on the media that is used to share the content. Four basic media can be identified: text, audio, images, and video. Moreover, specific attention should be devoted to the merging of these media, e.g., text in images (e.g., memes published on social networks).

3.1. How filter algorithms work

The techniques used to filter content differ depending on the media to be analysed. A filter can work at different levels of complexity, spanning from simply comparing contents against a blacklist, to more sophisticated techniques employing complex AI techniques. As for AI in general, such techniques may be based on symbolic approaches to knowledge modelling and reasoning with it, or on statistics and machine learning, or on a combination of the two approaches. In machine learning approaches, the system, rather than being provided with a

[43] Source: Barnett et al (2018).

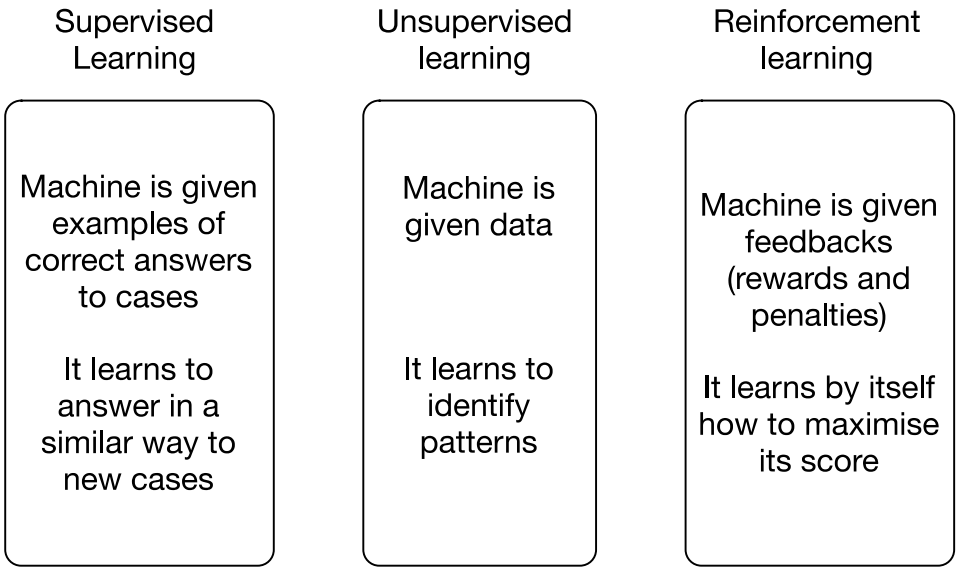
[44] https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

logical definition of the criteria to be used to find and classify content (e.g., to determine what counts as hate speech, defamation, etc.) is provided with a vast set of data, from which it must learn on its own the criteria for making such a classification.

3.1.1. Machine learning

In recent years, machine learning approaches have become dominant, and indeed AI system deployed in filtering are usually based on machine learning, They can adopt any combination of the three main approaches shown in Figure 7.

Figure 7. Kinds of machine learning



Supervised learning is currently the most popular approach. On this approach the machine learns through "supervision" or "teaching." It is given a training set, i.e., a large set of (probably) correct answers to the system's task, and learns to answer new cases in a similar way. For example, the training set for a system designed to detect spam consists of messages that humans have classified as instances of spam or not. The spam-detecting system learns to classify new messages in the same way as in its training set. Similarly, a system meant to detect hate speech (or pornography) may be trained though a set of posts, where instances of hate (or pornography) are distinguished from the rest of the posts.

In *unsupervised learning*, AI systems learn without receiving external instructions: they instead identify patterns in data. The techniques for unsupervised learning are used in particular, for clustering, i.e., for grouping the set of items that present relevant similarities or connections (e.g., documents that pertain to the same topic, people sharing relevant characteristics, or terms playing the same conceptual roles in texts). For instance, document sharing the same offensive language may be automatically grouped.

In reinforcement learning a system learns from the outcomes of its own action, namely, it observes the outcomes of its action and self-administers the rewards or penalties (e.g., points gained or lost) that are linked to the outcomes of such actions. Consider, for instance, a system that learns to prioritise or deprioritise news posts depending on the extent to which users shown interest in the posts presented by the system.

3.1.2. Neural networks

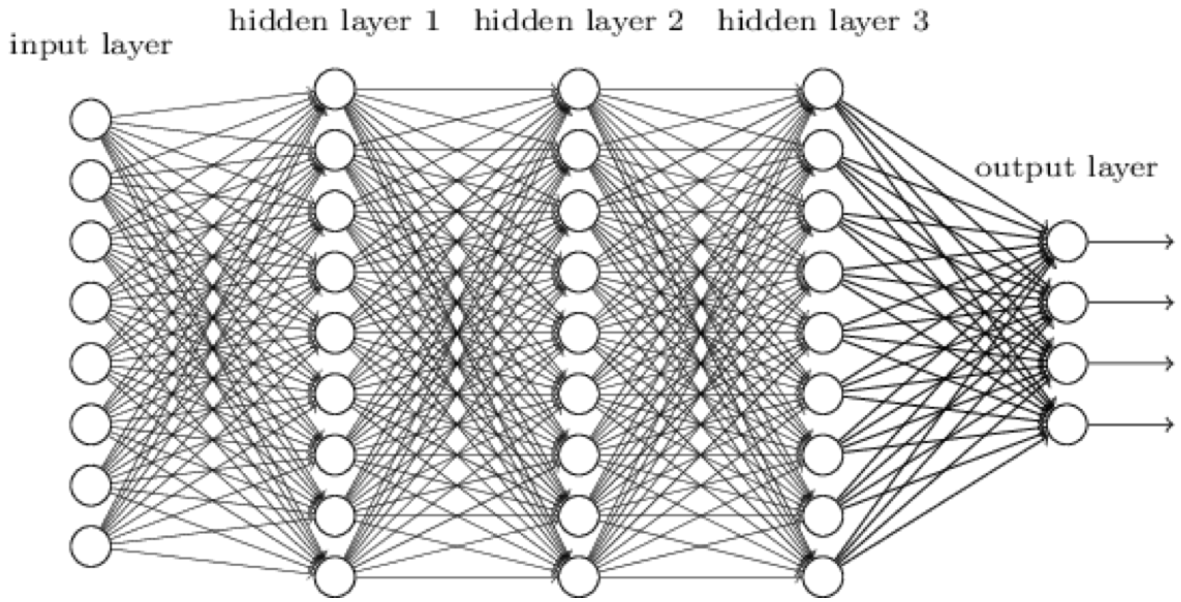
Many techniques have been deployed in machine learning: decision trees, statistical regression, support vector machine, evolutionary algorithms, methods for reinforcement learning, etc. Recently, deep learning based on many-layered neural networks has been very successfully deployed especially, but not exclusively, where patterns have to be recognised and linked to classifications and decisions, as in the case of filtering (e.g., classifying content as offensive of pornographic, and treat it accordingly). Neural networks are composed of a set of nodes, called neurons, arranged in multiple layers and connected by links. They are called that way because they reproduce some aspects of the human nervous system, which indeed consists of interconnected specialised cells, the biological neurons, which receive and transmit information. Each neuron receives signals (numbers) from connected neurons or from the outside. In the case of filtering, the input may represent points in an image, or words in the text. Each neuron applies some calculations to the input it receives, and if the result reaches the neuron's threshold, the neuron activates, sending signals to the connected neurons or outside of the network. The activation starts from nodes receiving external inputs and spreads through the network. The training of the network takes place by telling the network whether its answers (its outputs) are right or wrong. If an answer by the network is wrong, the learning algorithm updates the network —i.e., it adjusts the weights of the connections between the neurons — so that next time the network is presented with that input, it will give the correct answer.

Many neural networks exist. In recent years the structure of networks has become increasingly complex: networks have been developed that have many layers (so-called deep neural networks), or that take account of the proximity of inputs, typically points in images and features emerging from them (convolutional networks), or that have the capacity to store and process sequences of inputs taking their order into account (recurrent neural networks or long-short term memory networks).

Neural networks have the best performance in pattern-recognition tasks, such as those involved in filtering. Unfortunately, they do not provide explanations of their outcomes. It is possible to determine how a certain output has resulted from the network's activation, but this information does not provide a rationale that is meaningful to humans: it does not tell us why a certain response was given. Thus, neural networks are said to be opaque systems, or black boxes.

Figure 8 shows a simplified representation of a multi-layered neural network (actual networks may have many more layers of neurons.

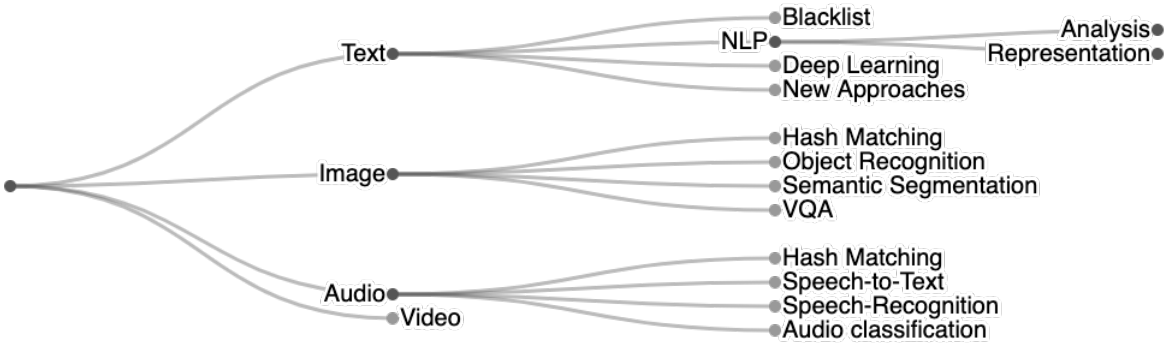
Figure 8. Multilayered (deep) neural network



3.1.3. Filtering and media

Figure 9 shows a taxonomy of different technologies based on the media they exploit. Notice that video media is usually a combination of audio and image media techniques.

Figure 9. Taxonomy of techniques



In the following subsections, we will analyse the different techniques that are used to exploit the knowledge embedded in different media in order to filter content.

3.2. Metadata searching, hashing and fingerprinting

Metadata searching, hashing, and fingerprinting can be used on any file, and therefore on any digital content, however the file is rendered. They are used in particular in the copyright domain, since they can be used to identify copies of given digital works.

3.2.1. Metadata filtering

Metadata filtering uses the information (metadata) that accompanies an item of content (the data), providing information about that item. Typical examples of metadata are the following: book's title, author and publisher; a song's title, performer, and length; a movie's title, performers; a standardised identifier (such as the DOI, digital object identifier). Metadata search is used in particular to find unauthorised copies of copyrighted works: a platform (or

new input data being posted to it) is scanned for metadata matching a target copyrighted work and matching files are filtered out or marked for removal.

Obviously, a metadata search is only possible if the online content is accompanied by its metadata, i.e., if the metadata are stored in the same file that host the data, or are linked to that file.

Even when metadata are available, metadata filtering of unwanted content is often inaccurate and easily circumvented, as a file's metadata may be incorrect, or can be easily manipulated to avoid detection. Metadata searches often misidentify content, because two pieces of different content can share the same metadata (two songs may have the same title, for example).

3.2.2. Hashing

A hash function takes a file as input and generates a small record that is uniquely linked to the file (for a different file a different hash would be produced). Hash-based filtering presupposes that a database of the hashes of the unwanted items is available (e.g., copyrighted works that should not be posted without authorisation, child pornography videos, etc.). When the hash filter examines an item, the hash of the item is produced, and then it is matched against the database of the hashes. If a match is found between the hash of the examined items and the database, this means that the new item is exactly the same as an unwanted item whose hash was put in the database.

As any change in a data item (changing the format of the file, compressing it, deleting a few words from a text, or shortening a song by second) will produce an entirely different hash, hash-based filtering can only identify exact matches of unwanted files. Hash-based filtering can be circumvented by making small changes in the posted file (e.g. a copy of a movie), which make it different from the original that was used for creating the hash stored in the database.

3.2.3. Content fingerprinting

Fingerprinting tools, like hash functions, take a media item as input and deliver a record that uniquely identifies that item. The difference from hashing is that the record stores a set of relevant characteristics of the underlying item, such as frequencies in a song file, brightness levels or object movements in a video, or snippets from it. Fingerprinting-based filtering presuppose the availability of a database of fingerprints of unwanted items. When a new item is produced, its fingerprint is generated and matched against the database. In comparison to hash-based systems, fingerprinting can recognise an item of content, even if some changes have been applied to it (in colour, dimension, format, etc.), as long as the characteristics on which the fingerprint is based have not been affected. Thus, fingerprinting systems are more robust than ash-based systems, but they also are more complex and costly.

3.3. Filtering on text

Textual content which is available online includes any kind of literary work (reports, articles, books, etc.) or posted or exchanged message. In some cases, detecting unlawful content only requires matching textual sequences. This is the case, for instance, with violations of copyright over textual documents, which usually involve the exact reproduction of large portions of text. In other cases, an analysis of the text language and meaning is needed, as when instances of hate speech are to be identified. These diverse situations can be addressed on the basis of different approaches.

3.3.1. Blacklisting

The blacklist approach involves creating and maintaining a dataset of unwanted textual content. Incoming texts are then compared with the dataset to spot undesired texts, which are correspondingly rejected, deleted, or flagged.

Applications of this simple approach can take place many different scenarios. For instance, violation of copyright over textual documents can be identified by matching an uploaded document with a database of copyrighted textual works. Abusive content can be found by matching posted documents with blacklists of expressions that are often included in abusive messages (e.g. insults, racial slurs, swearwords). In such cases, it is possible to block or remove the entire document or only the blacklisted expressions. To identify varying combinations of words conveying the same abusive content, various techniques can be used, including regular expressions (formally defined search patterns).

In online forums, chat rooms, or video games, language filters —also known as *swear filter or profanity filter) are deployed to* modify text by removing expressions that are deemed offensive by the administrator or the community. For instance, Facebook offers the administrators of its pages the ability to apply various levels of such filters.[45]

Blacklisting approaches can be hacked by misspelling the undesired words or combining texts with graphics, such as emojis.

3.3.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is the subfield of computer science which studies how to equip computer systems to handle the language naturally spoken by humans. NLP techniques leverage linguistic studies in natural language and address the syntax, semantics, and pragmatics (the contextual meaning) of expressions in natural language.

In textual filtering, natural language processing is needed whenever the simple occurrence of certain word patterns is insufficient to classify the relevant textual items as needed. Consider, for instance, the case of hate speech, and assume that the purpose of the filtering is indeed to identify expressions of hatred. In the 2019 United Nations Plan of Action on Hate Speech, the following definition is provided:

> Hate speech is any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.

Illegal hate speech is defined in EU law as the

> public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin."[46]

These definitions make it clear that —regardless of whether we only focus on illegal hate speech (incitement to violence or hatred, as in the EU law definition), or rather address a broader notion of abusive expressions of hatred (as in the UN definition)— many instances of hate speech cannot be identified and distinguished from innocent messages by looking at

[45] See the Facebook Help Center: https://www.facebook.com/help/131671940241729.
[46] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

single words or combinations of them. In fact, expressions of hatred can be conveyed in many ways, and the same words typically used to convey such expressions can also be used for different purposes. For instance, such words can be used for condemning violence, injustice or discrimination against the targeted groups, or just for describing their social circumstances. Thus, to identify hateful content in textual messages, an attempt must be made at grasping the meaning of such messages, using the resources provided by natural language processing.

The automated analysis of a text for filtering purposes may involve multiple steps, taking into account the syntax of the text (the occurrence of word and of words combinations) and its semantics (meanings).

Syntactical analysis may involve the following: (a) tokenization, identifying the elements of a text (typically its words); (b) part-of-speech tagging, assigning every element to a part of speech (e.g., verb or noun); and (c) parsing, identifying the syntactic structure of whole clauses and sentences.

Semantic analysis may involve the following: (a) name entity recognition (identifying proper names), (b) lexical semantics (determining the meaning of single words), (c) topic categorisation (determining the subject matter), (c) natural language understanding (determining the meaning of chucks of text), (d) sentiment analysis (determining the positive or negative attitude or polarity expressed in the text).

In the domain of natural language processing the machine learning revolution has had a vast impact, successfully challenging systems based on human-made representations of complex grammar rules and semantic models. It has indeed been claimed by leading AI researchers that in domains such as speech recognition and machine translation, "invariably, simple models and a lot of data trump more elaborate models based on less data."[47] Machine learning systems based on simple representations input texts —e.g., focusing on the words occurring in a document (bag of words), in the textual contexts of such words (word embeddings), as well as on simplified syntactic structures (tree kernels)— have been very successfully trained on vast amounts of data. Neural networks have been applied to classify opinions in written text[48], but also to detect instances of hate speech.[49]

3.4. Filtering on images and multimedia

Visual information plays a key role online. In many domains and contexts, images alone or in combination with text are more effective than text alone. The dynamic combination of images, sounds (in particular spoken language), and text are used in videos, the most popular way of distributing content online today.

3.4.1. Hashes and fingerprints of images

Hash and fingerprints databases (see Section 3.2.2) are available for different kinds of visual content.

- Microsoft's PhotoDNA[50] aids in finding and removing known images of child exploitation using a hash mechanism. Microsoft donated PhotoDNA to the National Center for Missing & Exploited Children (NCMEC). The service is also available through

[47] See Norvig et al (2009, 9).
[48] Deriu and Cieliebak (2016).
[49] See Fortuna and Nunes (2018).
[50] https://www.microsoft.com/en-us/photodna.

CyberTipline, the US centralized reporting system for the online exploitation of children. Its Child Victim Identification Program has reviewed more than 313 million images and videos. In 2019, reports to the CyberTipline included 69.1 million files with 27,788,328 images, 41,280,816 videos, and 89,053 other files[51].

- Facebook, Microsoft, Twitter, and YouTube participate in the Global Internet Forum to Counter Terrorism (GIFCT). They share a database of hashes extracted from violent terrorist or terrorist recruitment videos which were previously removed from their platforms[52]. The database now contains more than 200,000 hashes of videos and can be used to block the re-posting of such videos.
- Content ID[53] enables copyright holders to identify YouTube videos that include content they own. Videos uploaded to YouTube are scanned against a database of files that have been provided by copyright holders. The latter can then choose whether to monetize, track, or block the video.

As noted in Section 3.2, hashing and fingerprinting based systems can be hacked by editing images so that their hash or their fingerprint differs from the original one.

3.4.2. Images and texts in memes

The so-called "memes" are short messages combining text and images or videos, which tend to replicate with a quick uptake and spread across the Internet.[54] A vast number of memes exists that express hate (based on race, gender, sexual orientation etc.) or that are meant to combat expressions of hate. Memes are in general a complex combination of cultural heritage, slang and idiomatic expressions, which makes it difficult to decode them automatically.[55]

The automatic identification of unwanted memes is particular difficult, because the meaning of memes is given by their combination of different media, and requires decoding contextual and socio-cultural references. Facebook has created a databased of 10,000 hateful memes it makes available to researcher, so as to enable them to develop ways to detect hate speech.

3.4.3. Combining text and videos

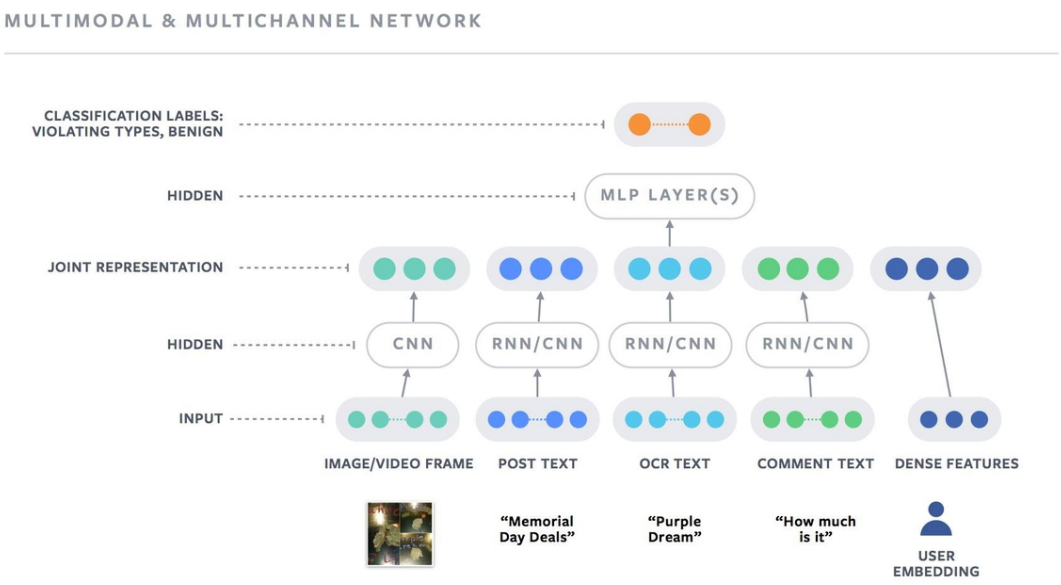
A number of approaches exist that address multimedia combinations. The Facebook's Rosetta From Ai project[56] aims to understand text which appears in images in order to screen readers for visually impaired persons. The technology is based on a broad machine learning approach which extracts text from billions of public images and videos posted on Facebook and Instagram.

Facebook has also developed a machine learning approach, called Whole Post Integrity Embeddings (WPIE), to deal with content violating Facebook guidelines. The system addresses multimedia content, by providing a holistic analysis of a post's visual and textual content and related comments, across all dimensions of inappropriateness (violence, hate, nudity, drugs, etc.). It is claimed that automated tools have improved the implementation of Facebook content guidelines. For instance, about 4.4 million items of drug sale content have been

[51] https://www.missingkids.org/footer/media/keyfacts.
[52] https://gifct.org/joint-tech-innovation/.
[53] https://support.google.com/youtube/answer/3244015.
[54] The notion of a meme has been coined from the biological term "gene", to denote units that cultural units who induce their own replication (thanks to the appeal they have on human minds), see Dawkins (1989, Ch 13).
[55] See Gomez et al (2020).
[56] Borisyuk et al. (2018).

removed in just the third quarter of 2019, 97.6 percent of which were detected proactively. Figure 10 shows the application of Facebook system to a post involving drugs (marihuana).

Figure 10. Whole Post Integrity Embeddings (WPIE), by Facebook



Source: Facebook Community Standard Report[57]

3.4.4. Understanding spoken language

In order to analyse the content of podcasts, soundtracks, and videos including spoken language, it is necessary to transforms sounds into textual sequences. This is the task of automated speech recognition, an area in which huge progress as taken place in recent decades: speech recognition systems are not available on any kind of devices, which a very high level of performance. Usually speech recognition is performed using supervised learning techniques implementing convolutional neural networks or recurrent neural networks. Once the speech has been transformed into text, it is possible to employ the previously described techniques for dealing with text or multimedia.

3.5. Accuracy of filter algorithms

Content filtering systems are evaluated on standard metrics in order to determine their accuracy. In this section we shall consider the criteria to be used to assess the performance of a filter, along with various issues pertaining the accuracy of automated filters.

3.5.1. Criteria for accuracy

Filtering can be viewed as a binary classification task whose purpose is to assess whether a given item belongs to a positive class or to a negative class (e.g., the message is harmful or non-harmful, infringing IP or not). A filtering system's positive or negative answer can be evaluated as follows:

[57] https://ai.facebook.com/blog/community-standards-report.

- true positive (TP): the system classifies an item as harmful, and the item is in fact harmful,
- true negative (TN): the system classifies an item as non-harmful, and the item is in fact non-harmful,
- false positive (FP): the system classifies an item as harmful, but the item is in fact non-harmful,
- false negative (FN): the system classifies an item as non-harmful, but the item is in fact harmful.

True positives and true negatives are correct answers, false negatives and false positives are wrong answers. Thus, the higher the proportion of TP+TN, among all answers given by a classifier (a filter), the better the performance. This is the so-called accuracy, but other metrics are also used to assess the performance of a filter, as shown in Table 1 below.

Table 1: Performance measures for classifiers

Term	Definition	Calculation
Sensitivity (recall)	Ability to classify positively what is positive (proportion of positive cases that have been correctly classified)	TP/(TP+FN)
Specificity	Ability to classify negatively what is negative (proportion of negative cases that have been correctly classified)	TN/(TN+FP)
Precision	Ability to classify positively only what is positive (proportion of positively classified cases that are really positive)	TP/(TP+FP)
Accuracy	Ability to classify cases correctly (proportion of cases that have been correctly classified)	(TP+TN)/(TP+TN+FP+FN)

Filtering technologies, and in particular those based on machine learning approaches, are based on probabilistic methods; errors cannot be completely avoided. At a given level of technical performance, we can usually reduce the false negatives rate (i.e., the likelihood of exposing users to content that should be rejected), only if we increase the false positives rate (the likelihood that acceptable content will be rejected). In other terms, we can improve sensitivity only by worsening specificity, or equivalently we can improve recall only by reducing precision.

3.5.2. Fallibility of filtering systems

Machine learning systems need big amount of training data that are representative of the domain being addressed. The datasets used to train such systems should contain enough samples for each relevant category of cases. If some categories of acceptable or inacceptable items under-represented in the training set, then the tool may be unable to recognize items of that category. This would affect the accuracy of the tool, leading to removal of items which are not harmful or, contrarywise, to publication of items which should instead be banned.

For some kinds of content finding datasets with appropriate characteristics it is not a problem. This is the case when the categories are well-known, and parameters for assessing data items are clear. Examples come from the copyright-infringement area and the child sexual abuse material. In such cases, it usually clear whether an item should be classified as infringing or not. For some other kinds of content, such as hate speech this distinction is often not easy. Just two examples to illustrate the complexity of the task.

On 15 March 2019, two mass shooting attacks took place in Christchurch, New Zealand, during prayers at mosques. The terrorist who carried out the attack livestreamed the mass murder of worshippers using a camera fixed to his helmet. The video was searchable on YouTube, Facebook and other platforms: neither automatic tools nor human moderators could detect the video and block it before it was made available online.[58] On the other side, YouTube has removed thousands of videos on atrocities in Syria. Those documents could be used as evidence of crimes committed against humanity during the war, and their removal could potentially jeopardize future war crimes prosecutions.[59] The videos at issue in these examples had a very similar visual content, but they were meant to transmit very different messages. To detect such differences in meaning is a really complex task. As curious examples in which automated classification failed to correctly identify certain items, we can mention the removal of image of Copenhagen's Little Mermaid statue for breaking nudity rules[60] as well as the ban of photo of the Neptune statue in Bologna for being explicitly sexual.[61]

Even more robust systems, such as Content ID by YouTube, can deliver weird results. For instance, in 2018 a musician uploaded 10-hour video of continuous white noise. The system notified the author for five copyright infringement claims, referring to white noise videos intended for sleep therapy.[62]

3.5.3. What ground truth? Subjectivity, bias and discrimination

In machine learning domain the term "ground truth" is used to refer to the correct outcome, as identified through standards external to the system, as opposed to the outcome that is proposed by the system. This expression, often used in machine learning apparently derives from cartography, and opposes the representation on a geographical map to the real situation on the ground, which provides the undisputable standard to determine whether the map is correct or wrong.

In online filtering, however, the ground truth is not provided by a physical reality but rather by human assessments on whether a certain item falls or not into a category of unwanted content, according to laws, guidelines and social norms, as interpreted by the individual assessors. Human assessments constitute indeed the training set of automated classifiers and provide the basis for evaluating and correcting the outcomes of such classifiers.

Thus, the working of an automated filtering system will reflect the attitudes, and possibly the biases of the humans whose behaviour the system is meant to emulate

The standard for making content accessible to the public also vary in different social and cultural context: language and content that is acceptable in certain communities, may be inacceptable to others. In large communities, different views may exist concerning what is

[58] Lapowsky (2019).
[59] Browne (2017).
[60] Bolton (2016).
[61] https://edition.cnn.com/travel/article/facebook-neptune-statue-photo-ban/index.html.
[62] Baraniuk (2018).

appropriate or inappropriate to the community concerned. For instance, there has been a discussion concerning the removal of images of female breasts on social networks, in general, or in connection with breastfeeding or other special contexts.[63]

Filtering may raise some issues concerning unfair discrimination, to the extent that the content delivered by or concerning certain groups may be excluded or deprioritised. For instance, a debate has taken place concerning the erasure of drag queens and LGBTQ+ books[64] and the removal of sex workers pages from social networks.[65]

Finally, content filtering can lead to not only to the identification of unwanted material, but also to the identification of the individuals that have published the materials been filtered. This will affect the privacy and data protection rights of such individuals, possibly to an extent that is disproportionate relative to the benefit that filtering out unwanted content may provide. This issue has emerged in the EU law, in connection with automated filtering for detecting copyright violations, as in the Promusicae and SABAM cases at the Court of Justice. Content filtering may also be used to identify and target political oppositions, as has happened in China and Egypt.[66] In such cases freedom of expression and association are seriously infringed.

3.6. Addressing failures

As filtering systems are both fallible (Section 3.5.2) and based on subjective judgements (Section 3.5.3), their possible failure should be anticipated, and mistakes should be identified and remedied. Therefore, two aspects are key: on the one hand transparency of the filtering process and on the other hand procedure to challenge its outcomes.

3.6.1. Transparency and accountability

To ensure that content is filtered out on reasonable and non-discriminatory bases, it is necessary that transparency is ensured both towards the individuals directly concerned, and toward society at large.[67]

With regard to such individuals, mechanisms should be provided to inform them that content they uploaded has been filtered out or blocked, and to lodge complaints in case they do not agree with the assessment of the filtering system.

[63] https://help.instagram.com/172319602942927.
[64] https://www.huffpost.com/entry/rejection-of-lgbt-christian-ads-social-media_b_9525612.
[65] http://www.mtv.com/news/3143241/sex-work-censorship-effects/-
[66] https://www.opendemocracy.net/en/north-africa-west-asia/how-twitter-gagging-arabic-users-and-acting-morality-police/.
[67] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

Figure 11. Facebook removals statistics

Source: Barrett (2020)

With regard to society, information should be provided on individual cases (preserving the privacy of the individuals concerned) as well as on aggregate results.

Among the initiative meant to provide transparency, we can recall the Santa Clara Principles[68], the Corporate Accountability Index[69], and the Electronic Frontier Foundation principles on filtering.[70].

The Santa Clara Principles[71] address companies engaged in content moderation. These companies should:

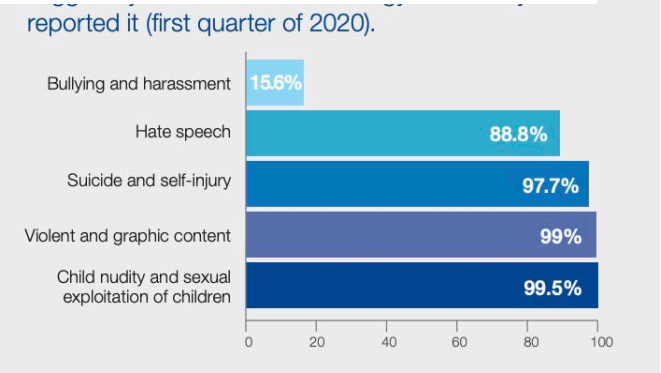
- publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines;
- make sure that each user whose content is taken down or whose account is suspended is given notice of the reason for the removal or suspension;
- provide a meaningful opportunity for timely appeal of any content removal or account suspension.

Twitter's latest Transparency report implements the Santa Clara rules providing detailed data on its filtering operations. In the first semester of 2019. 15,638,349 unique accounts were reported for possible violations of Twitter Rules. During this period, Twitter suspended a total of 244,188 unique accounts for violations related to child sexual exploitation. Of those unique suspended accounts, 91% were flagged by a combination of technology (including PhotoDNA and internal, proprietary tools)[72]. Moreover, Twitter suspended 115,861 unique accounts for violations related to the promotion of terrorism. About 87% of them were suspended using internal, proprietary tools. These numbers suggest how the hybrid approach for content moderation seems to the emerging solution for mitigating some of the aforementioned issues.

Automatic and hybrid content moderation is also employed by YouTube. The content moderation is applied centrally both ex-ante and ex-post. According to YouTube's Transparency Report[73] during the first trimester of 2020, YouTube moderatos manually

[68] https://santaclaraprinciples.org/.
[69] https://rankingdigitalrights.org/index2019/.
[70] https://www.eff.org/wp/who-has-your-back-2019.
[71] https://santaclaraprinciples.org/.
[72] https://transparency.twitter.com/en/twitter-rules-enforcement.html.
[73] https://transparencyreport.google.com/youtube-policy/removals.

Figure 12. Facebook statistics on the usage of AI tools for filtering



Source: Barrett (2020)

flagged and removed 399.422 videos from the website, while automatic procedures flagged and removed 5,711,586 videos[74]. This is more than 14 times the number of videos flagged by humans. Apart from the difference in numbers, it is interesting to appreciate that the automatic procedure makes it possible to remove 53% of videos before they are viewed at least one time and 28% of videos with a maximum of 10 views. Moreover, YouTube removes possible harmful comments. During January-March 2020 the platform flagged 693.579.605 comments as possible spam.

In its transparency report, Google indicates the number of links (URLs=) requested to be delisted from the results of its search engine. The number is enormous, more than 4 billion URLs[75] have been requested to be delisted upon notification due to copyrights infringements. Unfortunately, the webpage does not report whether the delisting is done automatically or manually. It seems plausible that a hybrid approach is adopted.

Google sends the details of approved delisting request to the Lumen database,[76] a project of Harvard's Berkman Klein Centre for Internet & Society which collects and analyses legal complaints and requests for removal of online materials.

A new report from New York University's Stern Center for Business and Human Rights[77] suggests that Facebook should stop outsourcing to third-party vendors and should instead bring content moderators in-house, make them full employees, and double their number of content moderators to 30,000. Figure 11 depicts the amount of removals for different categories. Figure 12 shows the percentage of Facebook content that has been removed after being flagged through though AI tools.

3.6.2. Appeal and redress mechanisms

As discussed in previous sections, content filters can make mistake: unwanted materials can be accepted by the filter (false negatives), or acceptable materials can be rejected (false positives). Mistakes can take place in all filtering models: when the filtering is based on human intervention, on a fully automated process or a combination of the two. Given the fallibility of filtering, every socio-technical system for moderation should include measures to identify mistakes and react to them.

The most obvious and significant way of addressing mistakes is to empower users: users who upload content, as well as users who access it. First, users who post content should be warned whenever such material is blocked or removed by filters. They should have the opportunity to contest such decisions, presenting their reasons against filtering out their posts. Second, all

[74] https://transparencyreport.google.com/youtube-policy/removals?total_removed_videos=period:Y2020Q1;exclude_automated:human_only&lu=total_removed_videos.
[75] https://transparencyreport.google.com/copyright/overview?hl=en&copyright_process=p:2&lu=copyright_process
[76] https://www.lumendatabase.org/.
[77] Barrett (2020).

users should have the opportunity to flag items they consider to be unlawful or inappropriate, and to present their reasons in favour of removal. In both cases, simplicity of the procedure is most important, as is a timely response by moderators.

YouTube provides a simple mechanism allowing users to appeal against the removal or their posts.[78] In its transparency report, YouTube gives evidence of the scope of this procedure: during the first trimester of 2020, a total of 6,111,008 video were removed for violating its Community Guidelines. More than 2% of these removals were appealed, for a total of 165,941 videos. The appeal requests were reviewed by senior moderators, who can uphold or reverse the decision. Of the reported appeals a total of 41,059 (almost 25% of the received appeals) were successful, leading to reinstatement of the videos.

Twitter users can submit a form to contest a suspended or blocked account[79], the procedure is manually validated and usually takes up to 7 days to be validated. Unfortunately, the process lacks transparency, as the number of filed appeals is not mentioned in the transparency report, nor is the number of processed or reinstated accounts or tweets. Instead, the report describes the number of Tweets and accounts that are withheld[80]. As reported by the platform "Many countries, including the United States, have laws that may apply to Tweets and/or Twitter account content. […] if we receive a valid and properly scoped request from an authorized entity, it may be necessary to withhold access to certain content in a particular country from time to time." In this regard, in comparison with the previous reporting period, Twitter received roughly 67% more requests to remove content, these requests originating from 49 different countries. [81]

Facebook's content appeal and restore process is described in Facebook website[82]. Facebook's transparency report indicated the number of appealed decision and also gives a statistic about restored content. In particular it reports how many items were restored without an appeal and how many after it (see Figure 13).

More complicated is the procedure for appeals for a blocked account. Due to several concerns about identity, this requires users to submit proof of identity in order to validate the request.[83]

[78] https://support.google.com/youtube/answer/185111.
[79] https://help.twitter.com/forms/general?subtopic=suspended.
[80] https://transparency.twitter.com/en/removal-requests.html.
[81] https://blog.twitter.com/en_us/topics/company/2019/twitter-transparency-report-2019.html.
[82] https://transparency.facebook.com/community-standards-enforcement/guide
[83] https://www.facebook.com/help/contact/269030579858086.

Figure 13. Statistics about Facebook appealed and restored content

Source: Facebook transparency report (https://transparency.facebook.com/community-standards-enforcement

3.7. Challenges to filtering

Filtering algorithms are subject to multiple malicious attacks, in ways that depend on the filtering technology being used.

Blacklist-based filters are particularly vulnerable. Small changes in the typesetting of an expression are sufficient for preventing a match with the corresponding entry in the blacklist database. The same result can be obtained by substituting a blacklisted expression or part of it with a slang idiom or an emoji. Tools exist which automatically modify a text by introducing small differences, in such a way that the modified text no longer matches the blacklist, but the original meaning can still be understood by humans.[84]

More advanced attack methods are needed to attack fingerprinting systems, as in the case of Shazam[85] a popular mobile phone app for identifying music. Shazam fingerprinting approach, based on hashes from spectrogram peaks, could be deceived by perturbing audio in order to remove some elements upon which the fingerprinting was based.[86]

With the complexity of filtering algorithm, also the complexity of attacks increases. Recently the adversarial attacks approach has been used to challenge automated filters.[87] In this approach two interacting systems are uses, a discriminator and a generator, each usually implemented through a neural network. [88] The generator's aim is to deceive the discriminator by producing messages that are misclassified by the discriminator, i.e., fake news or reviews that are classified as original ones, spam messages that are classified as non-spam, etc. The messages that succeed in deceiving the discriminator are then used to attack the filtering system being targeted. Spotting fake content generated by adversarial networks is very difficult, for both humans and automated systems. Adversarial attacks have succeeded in deceiving advanced filters for images, texts, and videos.[89]

[84] http://tools.seobook.com/spelling/keywords-typos.cgi.
[85] Wang et al. (2003).
[86] Saadatpanah et al. (2019).
[87] Saadatpanah et al. (2019).
[88] Goodfellow et al. (2014).
[89] Hao-Chen et al. (2020).

3.8. Availability and costs of filtering technologies

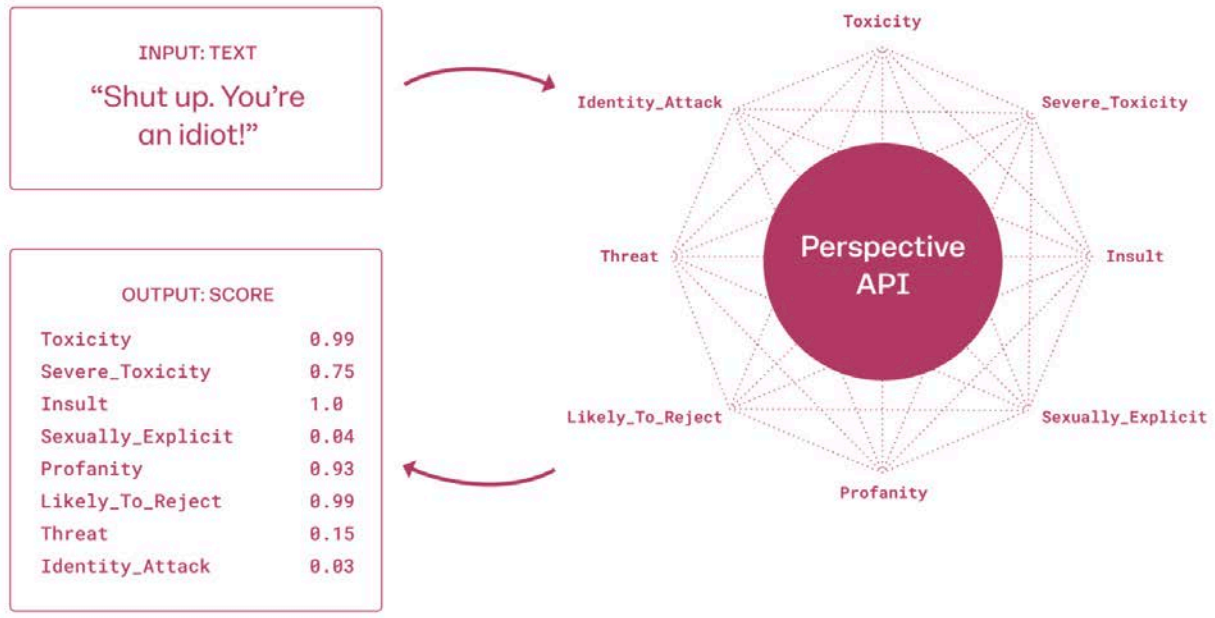
The growing use of content-platforms has expanded the need for technological solutions.

So far, big companies have been able to develop their own solution inhouse (see for instance Content ID from YouTube and other solutions cited in the previous sections), often investing large resources. For instance, Google until 2018 had invested about $100 million in the development of its Content ID system. [90]

However, the vast majority of content-platforms cannot afford inhouse development. This is due to the cost of designing, developing and maintaining an advanced filtering system as well as to lack of internal knowledge, competences and expertise. Moreover, developers and experts who are able to deploy and manage such solutions are difficult to recruit. Aside from these issues, it is also very difficult to find datasets for training system and test beds are not shared among companies.

Different third-party solutions are available on the market, including both proprietary and open source software. For instance, Audible Magic sells electronic media identification and copyright management solutions. Its products and services identify, monitor, track and manage copyrighted multimedia content in all of its forms, including analogue radio and TV broadcasts, Internet and satellite streams, stored digital files, and network file transfers. Its customers include Facebook, Sony, Disney, among many others. The cost of the service varies based on the size of the input stream to analyse, starting from $1,000 per month to analyse up to 1000 video to $28,000 per month for an input stream of up to 10 million videos. [91]

Figure 14. An input text is scored based on different attributes through an API



Source: https://support.perspectiveapi.com/s/about-the-api

Videntifier[92] is a visual search engine that is based on a patented large-scale local feature database. Around this solution, the company has built several services for law-enforcement

[90] Spoerri (2019).
[91] Spoerri (2019), Urban et al. (2017).
[92] https://www.videntifier.com/.

agencies, content owners, advertisers and others, to organise and create value from visual data. Essentially, this is a fingerprint-based image recognition tool which also works for the recognition of images within videos. The cost of the solution is based on the amount of hours of video or the number of images to be checked but also on the size of the reference database to be used as a validation set: from $490 for an input size of 15,000 hours of video or 3 million images to be checked against a control database of 5,000 hours and 1 million images, up to $9,990 for an input stream of 1.5 million hours of video and 300 million images to be checked against a control database of 50,000 hours of video and 10 million images[93].

BasicAI[94] is another platform for AI/ML data collection, data labelling, and iterative model training. The company offers customers a full suite of services and tools which leverage on expertise in machine learning, artificial intelligence and data annotation.

Some major players make available application programming interfaces (APIs) enabling third parties to access machine learning systems trained on big data[95]. One of them is Perspective[96], an API developed by Jigsaw and Google's Counter Abuse Technology team. The API makes it possible to interact with an AI system trained on millions of examples gathered from several online platforms and reviewed by human annotators. It uses machine learning models to score the perceived impact a comment might have on a conversation. This score can be used as real-time feedback which can help commenters or moderators. Presently the tool is free to use.

Facebook allows users to interact with its Rights Manager tool through an API (application programming interface).[97] The API enables publishers to claim copyright ownership for videos and manage copyright matching rules. It can be applied only to videos on pages, and all pages need to go through an enrolment process in order to be eligible to use the API.

It is worth noting that some state-of-the-art solutions are open source and available online[98], this makes it possible for small provider to develop in house solutions. Cloud platforms like Azure[99] or AWS[100], can make computational power and machine learning tools accessible to small and medium enterprises as well. Indeed, they provide the whole infrastructure needed to store and process big data, but also to train deep models.

.

[93] Japiot (2017).
[94] https://www.basic.ai/.
[95] Gorwa et al. (2020).
[96] https://www.perspectiveapi.com/.
[97] https://developers.facebook.com/docs/graph-api/rights-manager-api?locale=en_US.
[98] https://paperswithcode.com/sota.
[99] https://azure.microsoft.com/.
[100] https://aws.amazon.com/.

4. THE REGULATION OF FILTERING

As noted in Section 2.1.1, digital platforms enable users to upload content and make it available to online communities. The uploaded items of content strongly differ in their legal and social merit. Some items are unlawful and severely infringe on individual rights and social values. Other items are lawful but inappropriate: they are not forbidden by the law but are detrimental to the interaction taking place in the online community in which they are posted, being objectionable or inconvenient, or at least perceived as such by other users. Finally, most items are both lawful and appropriate, pertaining to the exercise or individual rights, or contributing to legitimate individual interests or to valuable social goals.

In Section 2, it was observed that moderation is needed to prevent online communities from becoming dysfunctional, relative to the interests of individual community members, to the rights of third parties, and to social values. Preserving productive, lawful and pro-social interactions may require filtering out content that is unlawful or inappropriate.

In Section 3, filtering technologies were considered. We noted that a number of these technologies have been employed to address the increasing amount of online materials. In particular, effective AI technologies are today available to detect unwanted content through a hybrid approach that also relies on human input.

In the present section we shall develop an analysis of the EU regulation of filtering, and we provide some policy guidelines.

We shall first present some premises that may be relevant in defining policies and then we shall make some policy recommendations.

4.1. Some premises for policies

In the present section some basic premises for the regulation of online filtering will be summarised.

4.1.1. Filtering objectionable material should not be discouraged

Moderation is required in online communities, to limit abuse and maintain friendly environments (see Section 2.1). Moreover, given the scale of online communities, a moderation architecture that includes automated analysis and filtering is needed, if harm has to be prevented and mitigated (Section 2.2). Thus, the law should favour good-faith initiatives meant to provide moderation, including automated filtering. An implication of this premise is that no preferential legal treatment should be provided for platforms owners who abstain from moderation, taking a hands-off approach, relative to those who actively engage in content moderation (see Section 2.4.4).

4.1.2. Moderation, and in particular filtering is fallible, even when well-intended

A key aspect to be considered in regulating filtering is that filtering is fallible (see Section 3.5). Even when filtering is exercised not only in good faith, but also with appropriate technological means and human oversight, there is always the possibility that some unlawful or abusive messages pass through the filter (false negatives), becoming available to the public. In parallel, there is the possibility that some innocent messages are rejected (false positives).

The fallibility of filtering must be taken into account, for designing a regulation that provides optimal results. In fact, as noted in Section 3.5.2, when appropriate technical means are

adopted, the reduction of misclassified unwanted messages can only be obtained by increasing the amount of misclassified innocent messages. Thus, a regulation that would have platform owners face legal liabilities for any unlawful information delivered on their websites, independently or whether this is due to failures in taking due care, would induce providers to adopt excessively strict screening tools and procedures, that would result in excluding much lawful and beneficial content. The stiffer such liabilities are, the more they could lead to the exclusion of valuable content.

4.1.3. Admissible filtering is not limited to unlawful content

Moderation, and in particular legitimate filtering may concern not only content that is unlawful, but also content that is "objectionable" (to use the terminology of the US Communication Decency act), in the sense, that it is unsuited to the community where it is posted, being aggressive, unpleasant or otherwise inappropriate to that community (see Section 2.1.1). In fact, aggressive or improper content, and uncivil interactions, can negatively affect the functioning of online communities and the participation of their members (especially those who are more easily intimidated, because of their character, or weaker social position) even when the expression of such content is legally permissible.

Thus, inappropriate content can legitimately be excluded from online community even when such content is not unlawful. The constitutional right to freedom of expression does not entail that there is an obligation for private individuals to listen to any permissible expression or for private organisations to distribute it. The idea that all legally permissible expression should be allowed in every online community would indeed be dysfunctional to the good functioning of such communities, which would be without any protection from content that, while legally permissible under general laws, would still be inappropriate, offensive or intimidating in the context of the community in question.

The decision on what should count as objectionable relative to a platform or online community, when taken in good faith, is a determination that goes beyond the law. It may indeed have an ethical dimension, which involves providers as well as their users, while going beyond them, i.e., a vision of what the online information should be, to enable the realisation of individual and social values, i.e., of "the flourishing of the [information] environment as a function of its plurality and diversity".[101]

4.1.4. Arbitrary power to filter out may be limited

The possibility that legally permissible expressions are legitimately filtered out raises the issues of whether there are any limits to platform's powers to decide what content to exclude.

Two distinct issues may be raised. First, it may be asked whether the exercise of filtering for reasons other than the removal of objectionable or cacophonic material would amount to editorial control, i.e., to selecting what content is to be made available online on policy grounds. The exercise of such a power appears to be incompatible with the role of a content intermediary and should lead to full editorial responsibility for whatever content is published.

Second, it may be asked whether granting owners of an online platforms a power to discretionarily filter out or remove any content they dislike—for any reason or for no reason at all—may involve an excessive restriction on freedom of expression.

[101] Floridi and Taddeo (2017, 34).

In the case law of the European Court of Human Rights, we can find contrasting indications in this regard. As the convention applies to States, and not directly to private parties, the issues would be whether States, in not intervening when an online platform unjustifiably removes content, may violate their obligation to protect users' freedom of speech from interferences from third parties. There are indeed some cases in which the Court affirmed that a failure of States to address a private media refusal to allow certain instance of speech would involve a violation of freedom of expression under the convention.[102] In other cases, the Court took the opposite view, stating that the Convention did not grant a right to express political views in a private space, such as a supermarket.[103]

The determination of the permissibility of such exclusion has to be assessed by considering on multiple factors such as:

- the extent to which the content at issue is important to the individuals concerned or to society;
- the extent to which the exclusion of content is arbitrary, not being grounded in sound and proportionate rationales;
- the extent to which alternative equivalent ways to communicate the content are available.

As an example in which the balance of reasons would be against blocking or removing, consider the case of non-violent, non-aggressive and non-abusive political speech being posted to or accessed through a platform which—such as Facebook for social networks or Google for search engines— enjoys a quasi-monopoly position and therefore is the only outlet through which certain kind of communication can reach a broad audience.

4.1.5. Contractual clauses on filtering have dubious legal relevance

Users participating in online platforms usually sign contracts (so-called terms of service) which grant extensive powers to platforms with regard to blocking or removing content. Users usually do not read such contracts. Moreover, users would probably accept the contractual terms even if they took the time to read them, given the lack of alternatives to the leading online services.

We may wonder whether the clauses granting providers an arbitrary removal power may be deemed unfair under the Unfair Contact terms Directive. According to Article 3 of that directive a clause is unfair when "contrary to the requirement of good faith" it causes a significant imbalance in the parties' rights and obligations arising under the contract, to the detriment of the consumer." Arguably, this may be the case for clauses that give providers the unrestricted power to filter out or remove any content posted by a user without specifying conditions for the exercise of this power.

4.1.6. No set of clear and mechanical rules is sufficient to identify unlawful or inappropriate content

No mechanically applicable rule can definitively specify once and for all what content is unlawful or inappropriate. There has been a trend among leading online platforms to move

[102] In case Verein gegen Tierfabriken v. Switzerland, (app. no. 24699/94, ECHR 28 June 2001), this concerned the refusal to accept commercial on animal rights, by a private broadcaster, having a dominant position on the market.
[103] In the Appleby case, activists were barred from protesting in a publicly accessible, yet privately owned shopping centre. The Court found that the right to free speech did not override the owner's property rights.

from broadly and vaguely framed standards —such the requirement to remove child pornography, or gratuitous violence— aimed at moderators into detailed rules, specifying what would count as child pornography or unacceptable gratuitous violence.[104] Unfortunately, neither standards nor rules provide a complete solution: while standards may fail to provide adequate guidance to decision -makers and lead to capricious or unfair applications, precise rules may fail for being under or over inclusive; their application can lead to block harmful content, all the while excluding harmless or beneficial communications.

The existence of automated filtering tools does not exclude the need for appropriate high-level standards and rules: both are needed to assess the merit of particular decisions and evaluate the working of human and automated moderation according to shared legal or other standards and rules. Controversial cases may require balancing multiple valuable goals, such as freedom of speech and of information vs protection from abuse, non-discrimination and civility, or protection of intellectual property vs the right to access and create culture.

On the other hand, the complexity of such assessments does not exclude the use of machine learning based systems: through their training, such systems may have absorbed decision patterns that reflect the balance of such values, as understood by the human decision-makers that have provided the systems' training sets. However, the parties involved should have the option to challenge the automated decisions —which reflect the biases of the decisions in the training set, and are unable to take into account situations requiring a departure from past patterns— requiring a fresh human assessment of their case. The availability of a framework of public standards and rules is needed to provide users with criteria for assessing the automated decisions.

4.1.7. Due care/reasonableness standards need to be used to assess providers behaviour

As more and more advanced technologies become available that enable providers to detect unlawful content, the issue arises of whether and to what extent failing to use such technologies could ground legal liabilities. In particular it may be asked whether providers that fail to deploy accessible and effective technologies could still rely on rules granting immunity for user-generated content. Could such immunities still cover the harm that could have been avoided through the use of the omitted technologies? The finding that a provider unduly omitted certain precautionary measures should be based on several factors, such as:

- the gravity of the risk of unlawful user behaviour that the omission of the measures would entail, this risk including both the probability of such behaviour and the seriousness of the harm it may cause;
- the technologies that are available for implementing such measures;
- the economic sustainability of the measure, given the (socially beneficial) business model adopted by the intermediary;
- the way in which such measures may affect the rights and interests of the users or the intermediary.

Let us consider some examples of ways in which these criteria may operate. The availability of effective content identification technologies, and their economic sustainability, may lead to the conclusion that there is an obligation to remove or to pre-emptively block the re-posting of the same work on the same platform. Similarly, the prohibition on providing

[104] Klonick (2018).

a link to certain materials in response to queries based on a person's name may extend to the future posting of copies of such materials, if such copies can be automatically identified. In the same way availability of effective ways to control the distribution of malicious software may lead to liability for the intermediary that has hosted or transmitted that software.

On the other hand, a request to filter out all unauthorised copyrighted works transmitted over a communication network will go beyond the required reasonable care, by excessively burdening the provider and disproportionately affecting users' rights, as argued by the EU Court of Justice in the Promusicae case.[105]

4.1.8. Inappropriate regulation can induce excessive or insufficient filtering

Any regulation meant to influence the way and the extent to which platforms filter user-generated content must be precisely targeted in order to avoid both under-detection and overkill.

Making platforms (or moderators) punishable or even liable to pay compensation for any unlawful harm caused by their users would lead to excessive interference in users' behaviour, leading to blocking or removing useful content, and to the banning of useful contributor: given the fallibility of moderation, and the costs involved in a detailed analysis of each single case, the only way out for platform owners may consist in adopting very strict and rigid criteria. This would not only negatively affect the satisfaction users' preferences, but may also interfere with users' rights to expression and information, and with the social value of open debate and shared information.

Limiting legal obligation and liabilities to platform owners that choose to actively engage in moderation, would on the contrary induce many platforms owners to refrain from any moderation, which would lead to the proliferation of unlawful and abusive content.

The legal quality and the social value of a regulation of filtering should therefore be judged not only by the extent to which it successfully deters unlawful communications, but also by the extent to which it does not deter lawful communication. Indeed, filtering should ideally maximise the difference between its good outcome, namely the importance of the unlawful communications it deters, and its bad side effect, i.e., the importance of the set of lawful communications it deters. In other words, liabilities should not extend in such a way that the benefit of preventing additional unlawful communication is outweighed by the damage of preventing additional lawful communication.

4.1.9. Online diversity has a positive value

As noted above, while the law itself determines what online content is unlawful (though the application of legal standards may require difficult and controversial assessments, that may involve balancing competing rights and values), each platform or community may develop different standards for determining what legally permissible content should be rejected or downgraded as being inopportune, objectionable, uncivil, misleading or simply unsuitable or cacophonic. Such standards may be used for human moderation, but they may also be implemented in automated filters. Platforms may also adopt less intrusive alternatives to the removal of inappropriate content, by downgrading such content, so that it does not appear at the top of general searches, and proving users with the ability to exclude unwanted content

[105] Productores de Música de España v Telefónica de España SAU, Case C-275/06.

from their view. Diversity among online communities is valuable, with regard to what content is viewed as objectionable, and therefore excluded or downgraded, enables different people, with different preferences, to look for the environments most suited to their interests and their normative expectations.

This does not mean that the law should be completely neutral relative to all online communities and every kind of legally permissible content. Public institutions may stimulate and support valuable initiatives that address non-coercively permissible antisocial behaviour, such as certain instances of fake news, aggressive language, and disparaging comments on individuals and groups. This may be done by stimulating the adoption of codes of conduct, providing incentives for beneficial private initiatives, educating the public, etc.

4.1.10. What obligations to monitor content are prohibited under EU law is uncertain

Article 15 of the eCommerce directive prohibits Member States from imposing any "general obligation on providers […] to monitor the information which they transmit or store" or any general obligation to "actively to seek facts or circumstances indicating illegal activity". A noted in Section 2.4, this important provision today has to be applied in in the context of an increased pressure toward a proactive role of intermediaries, as also emerges from subsequent EU rules. In particular, the 2018 Audiovisual Media Service Directive and the 2019 Copyright Directive require members state to impose on provider obligations to address unlawful online content. While both directives do not challenge Article 15, they express a tendency toward the law imposing a more active role for providers in monitoring online content, and blocking and removing unlawful materials.

According to Article 15, the issue of determining what obligation to monitor are prohibited under EU law is translated into the issue of establishing what obligations are general rather than specific. As the object of an obligation to remove or block content is defined in less specific terms, the scope of the obligation increases progressively, and it may be wondered where the threshold distinguishing specificity from specificity is to be located. For instance, an obligation to remove or block content may concern

- a single copy of unlawful file (e.g., a downloadable video or a web page) being identified by ab univocal identifier (e.g., its URL, the universal locator for resources available online);
- all copies of an individually identified file that are present in on an online repository (e.g., all copies of a certain copyrighted video);
- not only the copies of a file that currently made accessible by the provider, but all subsequent reposting of further copies;
- not only the copies of a given unlawful file, but also all variations obtained from that file, while keeping its unlawful meaning;
- all content that is relevantly similar top, and equally unlawful as, a specified unlawful file, even when such content is not obtained by changing that document (e.g., all messages conveying a certain kind of defamatory or hateful content relatively to certain individuals);
- all documents that are unlawful in virtue of certain common features (e.g., unauthorised reproduction of copyrighted works contained in a given database, paedophilia, incitation to hatred, or terrorism, etc.)

In the case law of Court of justice, we cannot find a precise criterion to distinguish permissible specificity from prohibited generality. When addressing this distinction, the Court has rather decided particular cases based on proportionality assessments. In other term, the criterion used to determine whether an obligation to remove and block content is general or specific is not generality properly understood, namely the amplitude of the scope of that obligation. It is rather the extent to which an obligation having a certain scope would have an impact on individual rights and social values, i.e., whether the advantages that imposing that obligation would produce by reducing unlawfulness and harm would outweigh the disadvantages it would cause.

This approach was adopted in various copyright cases, such as the 2008 Promusicae case[106], the 2011 Scarlet Extended case[107] or the 2012 SABAM case[108]. In SABAM, the Court argued that the injunction to filter out all copyrighted content managed by SABAM would impose prohibited general obligation to monitor. This conclusion was reached by arguing that

> "the injunction to install the contested filtering system is to be regarded as not respecting the requirement that a fair balance be struck between, on the one hand, the protection of the intellectual-property right enjoyed by copyright holders, and, on the other hand, that of the freedom to conduct business enjoyed by operators such as hosting service providers" as well as "the fundamental rights of that hosting service provider's service users, namely their right to protection of their personal data and their freedom to receive or impart information, which are rights safeguarded by Articles 8 and 11 of the Charter respectively."

In the 2019 Glawischnig-Piesczek case on the other hand, the ECJ ruled that an obligation to "to remove information, the content of which is equivalent to the content of information which was previously declared to be unlawful, or to block access to that information" is not a prohibited general obligation, provided that

> the monitoring of and search for the information concerned by such an injunction are limited to information conveying a message the content of which remains essentially unchanged compared with the content which gave rise to the finding of illegality and containing the elements specified in the injunction, and provided that the differences in the wording of that equivalent content, compared with the wording characterising the information which was previously declared to be illegal, are not such as to require the host provider to carry out an independent assessment of that content (thus, the host provider may have recourse to automated search tools and technologies).

This conclusion is argued considering the need to "strike a balance between the different interests at stake", i.e., to "effectively protecting a person's reputation and honour" without "imposing an excessive obligation on the host provider." Following the reasoning of the Court, an obligation to monitor defamatory content would not be excessive

> in so far as the monitoring of and search for information which it requires are limited to information containing the elements specified in the injunction, and its defamatory content of an equivalent nature does not require the host provider to carry out an independent assessment, since the latter has recourse to automated search tools and technologies.

[106] Productores de Música de España v Telefónica de España SAU, Case C-275/06, ECLI:EU:C:2011:771.
[107] Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM) Case C-70/10, ECLI:EU:C:2011:771.
[108] SABAM v. Netlog, Case C-360/10, ECLI:EU:C:2012:85.

5. POLICY OPTIONS

In this section, some policy options will be discussed.

5.1.1. General principles on providers' immunity

The uncertainties on the regulation of filtering reflect broader issues concerning the regulation of secondary liability of online providers under EU law.[109] As noted in Section 26, the immunities originally granted by the eCommerce directive are challenged today, in particular to the extent that also cover providers that choose to remain passive, rather that engage with harmful user behaviour. A revision of the provisions on provider's immunity in the eCommerce Directive may be considered.

As a model for a possible revision we might consider a recent proposal for revising Section 230 (1) of the US Communication Decency, by adding the bolded language to the 1999 text:[110]

> No provider or user of an interactive computer service **that takes reasonable steps to address unlawful uses of its service that clearly create serious harm to others** shall be treated as the publisher or speaker of any information provided by another information content provider.

According to this proposal, providers would lose their immunity from liability for the content published by their users, if they fail to adopt reasonable measure to prevent or mitigate grave harm, and their failure to adopt such measures results in harm. A rule of this kind could coexist with a "Good Samaritan" clause according to which providers should not be considered liable when acting in good faith to restrict access to or availability of unlawful of objectionable material.

We could indeed consider the provision of the eCommerce Directive along these lines would possibly provide some relevant advantages over the current version of the text:

- It would overcome the limitation of providers immunity to the three kinds or services specified in the eCommerce directive (mere conduit, cashing, and hosting), a typology that no longer fits the reality of today's digital services.
- It would clarify that providers hosting user-generated content or making it accessible are not protected if they fail to take reasonable measures to prevent unlawful harm to third parties. This failure, rather than the fact of playing an active role in making user-generated content accessible, grounds providers' liability.

Determining the reasonableness of a measure would require a complex assessment involving the considerations such as the following:

- the extent to which the measure would effectively contribute to prevent unlawful content and the gravity of the harm that distributing such content might cause;
- the extent to which the measure would also filter out permissible content and the gravity of the harm that non distributing such content might cause;
- the sustainability of the costs that the measure may impose on providers;

[109] For a recent account, see De Streel and Buiten (2019).
[110] Citron and Wittes (2020),

- the extent to which it may affect the usability of a platform or user engagement in a community;
- the size of a platforms and its business model (measures that would be unsustainable for small or no-profit providers, may be appropriate to larger commercial ones).

The difficulty of such an assessment does not mean that it is unmanageable. In certain cases, it would be clear that reasonable measures have not been adopted, as in the case of websites devoted to the distribution of unauthorised copyrighted content, or to making available revenge porn. In other cases, specific normative rules, judicial decisions or decisions by public could can provide sufficient guidance to providers.

Appropriate legislative language must be used to specify that the reasonable measures the omission of which would lead to liability of the provider only include those measures that can be viewed as a "minimally necessary" to provide an adequate protection to third parties. They do not include further measures that could be voluntarily and legitimately adopted by willing providers, to address to a larger extent unlawful and objectionable content (e.g., using filtering techniques, in combination with moderation, to preventively block or deprioritise hate speech or terrorist propaganda).

In fact, each provider has to strike a balance between the need to shield users and third parties from unlawful or objectionable content, and the need to enable free expression and interaction. There is no single legally permissible criterion to be used in striking such a balance: it is up to providers to determine where to set the line and what measure consequently to adopt, taking into account their own interests and values as well as the interest and values of their users and of society. However, with regard to unlawful content, there is a threshold, a "minimum" of reasonable measures, the omission of which puts the provider outside of the liability exemption.

5.1.2. Clarifying the scope of liability exemption.

A most difficult issue to be considered is to determine in what cases a host provider deserves immunity for hosted unlawful content, an issue that has been often addressed by using the misleading distinction between active and passive providers. This rather an issue that has to be addressed teleologically, considering what activities by host providers deserve being shielded from liabilities and what activities, would should rather be subject to ordinary liability rules.

We have argued that certain activities deserve being shielded from ordinary liabilities:

- engaging in indexing, prioritising, or linking the materials in such a way that users can find what is most relevant to them, according to their judgement, but also according to socially shared reasonable principles (e.g., prioritising most reliable sources)
- engaging in removing, filtering out, or deprioritising unlawful or objectionable content

To the extent that harm to third parties results from activities that fit with this description, providers should continue to enjoy exemptions for liability, or rather a reduced liability regime that protects them as long as they adopt minimal reasonable measures to prevent unlawful behaviour and mitigate its effects.

Given the current level of technology, as described in section, it does not seem that upload filters can be included in the minimal reasonable measures that should be mandatorily applied, unless such filters concern copies of univocally identified items of content, or items of

content that can automatically detected as resulting from a modification of variation of precisely identified items. Filtering out works that are only characterised as pertaining to an abstract description (e.g., terrorist propaganda, violations of IP, etc.) does not fit within mandatory reasonable measures, at the stage of today's technologies.

A similar analysis also applied to the imposition of obligations to filter out or remove unlawful content, as indicated in the next paragraph.

5.1.3. Specifying permissions and obligations to filter

A noted in Section 4.1.10, the key provision in the current EU framework is Article 15 in the eCommerce regulation, which prohibits Member States from imposing on providers any "general obligation to monitor the information which they transmit or store" or to "actively to seek facts or circumstances indicating illegal activity." We have also noted that the interpretation of this provision is difficult. As the pressures toward a stronger proactive involvement in monitoring unlawful content have increased, the case law has reacted by making "generality" depend on the balancing of interests involved, rather than on any descriptive criteria, such an the amplitude of the scope of an obligation to monitor, or on the openness of the language through which it is specified. Moreover, as technologies to detect unwanted content become more and more effective, and thus automated monitoring becomes more accurate and cost-effective, the judicial interpretation of what is general is likely to be further restricted.

Maybe this evolution should be reflected in the legislation, through provisions clearly distinguishes the default prohibition to establish general obligations to monitor, from the possibility that less specific obligations can be introduced under particular conditions.

For instance, EU law could state that an obligation is specific whenever it does concern items of content that are individually identified (e.g., through their URL), and that non-specific obligations to monitor content are prohibited, unless all of the following conditions obtain: the obligations concern a set of unlawful items, identified in clear non-ambiguous language; automated tools exist that enable accurate and cost effective detection of such items; such tools are accessible to the concerned providers having regard to their resources and business models; and the application of such tools to the items concerned is acceptable according to the balance of the interests at stake.

An noted above optional filtering may extend much beyond what is legally compulsory: it may cover unlawful content that there is no obligation to remove or block in the absence of a specific complaint, and it may also cover non-unlawful content that is "objectionable" or cacophonic relative to the concerned audience. Under such conditions also optional filtering should not entail the loss of liability exemptions.

5.1.4. Introducing procedural remedies against online removals

The automated decision to remove content posted by certain individuals can strongly affect the freedom of expression of such individuals, while the decision to leave it online can affect the interests of those who may be harmed by that content. All parties should have the opportunity to challenge automated decisions, expressing their points of view. In fact the parties involved —on the one hand the issuer of the message and on the other hand the alleged victim of it— may have opposing views; the first may view the message as a way

to express a permissible opinion, advancing a legitimate individual interest or even a valuable social cause; the second may view the same message as being illegal or harmful.

The positions of such parties, as addressees of automated decisions that affect their interests and rights can possibly be likened to the positions of the individuals that are the subject to automated decisions based on their personal data. According to Article 22 of GDPR[111], at least with regard to important automated decisions, measures should be adopted that "safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."

Thus, it could be argued the law should require providers to grant similar procedural rights to the individuals which are subject to automated decisions relative to the content that is posted online. It has indeed been affirmed that:

> "expansive rules on transparency concerning content removals, their processing, mistakes, actors and notifications could be introduced with personalized explanations for affected users and audits for authorities or researchers.[112]"

The generation of explanation for the decision of AI-based systems raises difficult issues, but this should not prevent the enforcement of such requirement in the filtering domain. In fact the explanations to be provided to uploaders on the reason why an item of content has been filtered out or demoted in searches can be relatively simple, just pointing to the category of abuse (violation of IP of a certain protected work, nudity, hate, etc.) which is at stake. An important complement to the management of complaints can consists in technological solutions to distinguish non-repetitive human complaints, which deserve a human answer, from artificially generated spam or fake complains that do not deserve such attention.

5.1.5. Establishing authorities dealing with online content

As noted above, neither general principles (standards) not precise rules are capable of providing sufficient guidance for determining what content can be lawfully distributed online, in what cases. In this regard too, a parallel can be established between the online distribution of content and data protection. In both cases multiple rights, values, and legitimate interests are at stake, and optimal solutions with regard to the balance of such interests require taking into account the specificities of rapidly evolving contexts and technologies. Therefore, a clear line distinguishing admissible from prohibited behaviour cannot be established through comprehensive legislative rules

In this regard too, data protection law may provide a useful paradigm. The GDPR does not provide precise rules for distinguishing permissible from impermissible processing of personal data; it rather states high-level standards, such as the requirements of fairness and transparency, the necessity of a legal basis, the necessity of risk prevention and mitigation measures, etc. The application of such standards is not mechanical, but rather involves taking account the importance of the interests at stake and the extent to which they can be affected, as well the technological and organisational resources that are available. The specification of such provisions is assumed to place thanks to the activity of data protection authorities, and

[111] General Data Protection Regulation -Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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 (O.J. L 119(2016), p. 1–88.
[112] De Streel and Husovec (2020, 48).

to their interactions with controllers and data subject. In this domain soft law also plays important role, though codes of practices, certification and the highly influential opinions and guidelines by the Article 29 Working Party, and now the European Data Protection Board.

We may consider whether a similar framework could also suit the regulation of online moderation, including automated filtering. The Commission could be delegated the task of enacting specific regulations on this matter. National media or telecommunication authorities could be entrusted with powers to enjoin providers to remove or reinstate content or adopt specific measures to prevent unlawful harm, and also to specify what such reasonable measures should be implemented for different kinds of content and in different contexts. The decisions of such an authority could be challenged in front of national judiciaries. Alternative Dispute Resolution mechanisms could also be established, and its use may be supported by protecting from liability those providers that comply with the outcome of such mechanisms.[113]

European coordination could be ensured at the European level through the activity of European bodies, such as the Body of European Regulators for Electronic Communications (BEREC) or the European Regulators Group for Audiovisual Media Services (ERGA). The creation of a new authority specifically devoted to the coordination of national authorities in regulating online platforms could also be considered. It has been noted that "an EU regulator could ensure effectiveness and internalise the cross-countries externalities but in close partnership with the national regulatory authorities to meet the principle of subsidiarity.[114]" In that regard, an interesting example could come from the enforcement of financial regulation over banks through the Single Supervisory Mechanism, within the European Central Bank.

Coordination between the judicial decisions on the matter should obviously be provided by the European Court of Justice, through its interpretation of the relevant provisions of EU law, in particular those concerning provider's liability and their obligations.

5.1.6. Supporting small enterprises

There is the risk that the imposition of moderation, and particularly of filtering requirement may be too costly for small enterprises, making it more difficult for them to compete with big internet companies. The latter have vast technological and economical resources and moreover can spread the cost of their filtering technologies across their larger user base. To ensure competitiveness in the digital market and survival of SMEs and micro enterprises two parallel directions could be taken.

On the one hand the "reasonable measures" that content platforms should adopt to counter unlawful content should be to be tailored to the size of such companies, the extent of the potential harm to their users, and their ability to access the needed technology and sustain the corresponding costs.

Public initiatives should be taken to support the development and accessibility of technically sound and affordable solution for automated filtering, which may enable operators to comply with legal requirements. In particular, the provision open source tools and data sets could be encouraged through financial and other incentives.

[113] De Streel and Husovec (2020, 42).
[114] De Streel and Husovec (2020, 51).

5.1.7. Providing an EU-wide approach

There is a strong need for a consistent EU-wide approach to the regulation of online filtering. It is true the legal standards for the lawfulness of only content are mostly provided in national laws, particularly when it comes to issues such as defamation, pornography, and hate crimes. However, national diversity should not hide the convergence that exist among the Member States, based on their cultural and legal affinities, but also of the shared legal framework that is provided by EU law, and by the European Convention on Human Rights, and the Charter of Fundamental Rights of the EU, together with a number of relevant EU legal instruments.

The judgement to be entrusted to EU law, with regard to filtering, does not concern the national standards for unlawfulness, but rather the lawfulness and proportionality of the national measure that are adopted towards providers, to induce them to address unlawful content. If this matter were not regulated by EU law, it would fully revert to the diversity of national approaches to civil and criminal liability, with a great increase in uncertainty and fragmentation. This would, today even more that at the time in which the eCommerce directive was enacted, "prevent the smooth functioning of the internal market, in particular by impairing the development of cross-border services and producing distortions of competition" (eCommerce Directive, Recital 40).

The exchange of information and approaches between national authorities, and the development of shared approaches, opinions, and guidelines at the EU level, may contribute not only to the convergence solutions at the national level but also to developing more appropriate solutions that better fit the Digital Single Market.

6. REFERENCES

- Balkin, J.M. (2014). Old school / new school speech regulation. *Harvard Law Review*,2296–2342.
- Baraniuk, C. (2018). White noise video on YouTube hit by five copyright claims. *BBC News*.
- Barnett, T., S. Jain, U. Andra, and T. Khurana (2018). Cisco visual networking index (VNI), complete forecast update, 2017–2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*.
- Barrett, P. M. (2020). Who moderates the social media giants? *New York University Stern Center for Business and Human Rights*.
- Bolton, D. (2016). Facebook removes image of Copenhagen's Little Mermaid statue for breaking nudity rules. *The Independent, Wednesday*.
- Borisyuk, F., A. Gordo, and V. Sivakumar (2018). Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 71–79.
- Browne, M. (2017). YouTube removes videos showing atrocities in Syria. *New York Times 22*.
- Brynjolfsson, E. and A. McAfee (2014). *The Second Machine Age*. Norton.
- Cambridge Consultants (2019). Use of AI in online content moderation. 2019 Report produced on behalf of OFcom.
- Citron, D., and M. A. Franks (2020). The internet as a speech machine and other myths confounding section 230 reform. *Boston University School of Law Public Law and Legal Theory Paper No. 20-8.*
- Citron, D. and B. Wittes (2017). The internet will not break: Denying bad Samaritans § 230 immunity. *Fordham Law Review 86*, 401–23.
- Cohen, J. D. (2019). *Between Truth and Power. The Legal Constructions of Informational Capitalism*. Oxford University Press.
- de Streel, A., M. Buiten, and M. Peltz (2019). *Liability of Online Hosting Platforms: Should Exceptionalism End*. CERRE.
- de Streel, A. and M. Husovec (2020). The eCommerce directive as the cornerstone of the internal market. Assessment and options for reform. Study requested by the IMCO committee. Technical report, European Parliament.
- Deng, L. and Y. Liu (2018). *Deep learning in natural language processing*. Springer.
- Deriu, J. M. and M. Cieliebak (2016). Sentiment analysis using convolutional neural networks with multi-task training and distant supervision on Italian tweets. In *Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Napoli, Italy, 5-7 December 2016*. Italian Journal of Computational Linguistics.
- European Commission (2020). White paper on artificial intelligence - a European approach to excellence and trust. Technical report.

- Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
- Floridi, L. and M. Taddeo (2017). New civic responsibilities for online service providers. In *The responsibility of Online Service Providers*, pp. 13–42. Springer.
- Fortuna, P. and S. Nunes (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*.
- Gomez, R., J. Gibert, L. Gomez, and D. Karatzas (2020). Exploring hate speech detection in multimodal publications. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 1470–1478.
- Gorwa, R., R. Binns, and C. Katzenbach (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society 7* (1), 2053951719897945.
- Grimmelmann, J. (2015). The virtues of moderation. *Yale Law and Technology Journal*, 17 42–105.
- Havaei, M., A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-
- M. Jodoin, and H. Larochelle (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis 35*, 18–31.
- Helberger, N., J. Piersonb, and T. Poell (2018,). Governing online platforms: From contested to cooperative responsibility. *The Information society 34*, 1–14.
- Hinton, G. E., A. Krizhevsky, and S. D. Wang (2011). Transforming auto-encoders. In *International conference on artificial neural networks*, pp. 44–51. Springer.
- Japiot, O. (2017). Copyright protection on digital platforms: Existing tools, good practice and limitations.
- Jee, C. (2020). Facebook needs 30,000 of its own content moderators, says a new report. *MIT Technology Review, June*.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Klonick, K. (2018). The new governors: the people, rules, and processes governing online speech. *Harvard Law Review 131*, 1599–670.
- Kotu, V. and B. Deshpande (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.
- Lapowsky, I. (2019). Why tech didn't stop the New Zealand attack from going viral. *Wired.com*.
- Leerssen, P. (2020). Cut out by the middle man: The free speech implications of social network blocking and banning in the EU. *Jipitec*.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Orwell, G. ([1945] 1972). The freedom of the press. *New York Times*.

- Perrault, R., Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, and J. C. Niebles (2019). The ai index 2019 annual report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stan- ford, CA*.
- Quintarelli, S., F. Corea, F. Fossa, A. Loreggia, and S. Sapienza (2019). An ethical perspective on artificial intelligence: Principles, rights and recommendations. *BioLaw Journal 2019* (3), 159–177.
- Radford, A., J. Wu, D. Amodei, J. Clark, M. Brundage, and I. Sutskever (2019). Better language models and their implications. *OpenAI Blog https://openai. com/blog/better-language-models*.
- Russell, S. J. and P. Norvig (2016). *Artificial Intelligence. A Modern Approach* (3 ed.). Prentice Hall.
- Sartor, G. and F. Lagioia (2020). *Study: The impact of the General Data Protection Regulation on artificial intelligence*. European Parliament.
- Singh, S. (2019). Everything in moderation - an analysis of how Internet platforms are using artificial intelligence to moderate user-generated content. *New America Foundation, July*.
- Spoerri, T. (2019). On upload-filters and other competitive advantages for big tech companies under article 17 of the directive on copyright in the digital single market. *J. Intell. Prop. Info. Tech. & Elec. Com. L. 10*, 173.
- Takikawa, T., D. Acuna, V. Jampani, and S. Fidler (2019). Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5229–5238.
- Urban, J. M., J. Karaganis, and B. Schofield (2017). Notice and takedown in everyday practice. *UC Berkeley Public Law Research Paper* (2755628).
- Valcke, P., Kuczerawy, A., and Ombelet, P.-J. (2017). Did the romans get it right? what Delfi, Google, eBay, and UPC Telekabel Wien have in common. In Floridi, L. and Taddeo, M., editors, *The Responsibility of Online Service Providers*, 101–15. Springer.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.
- Walker, K. (2017, June 18). Four ways Google will help to tackle extremism. *Financial Times*.
- Yoo, C. (2012). *The Dynamic Internet: How Technology, Users, and Businesses Are Transforming the Network*. AEI Press.

This study, commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the JURI Committee, addresses automated filtering of online content. The report introduces automated filtering as an aspect of moderation of user-generated materials. It presents the filtering technologies that are currently deployed to address different kinds of media, such as text, images, or videos. It discusses the main critical issues under the present legal framework and makes proposal for regulation in the context of a future EU Digital Services Act.

PE 657.101

Print ISBN 978-92-846-7065-9|doi:10.2861/848927|QA-01-20-543-EN-C
PDF ISBN 978-92-846-7064-2|doi:10.2861/824506|QA-01-20-543-EN-N