Silvia A. Carretta

Liability for Copyright Infringement and Algorithmic Content Moderation: A Matter of Proportion

1 Intro

Every day, a staggering amount of new online content is generated. Only a small amount of that online content is subjected to some form of editorial review before it is posted. Most of that online content is user-generated, from tweets to all kinds of different media uploads, which are often posted without any form of scrutiny from another human being checking the lawfulness of that content. Just to comprehend the sheer size of it, one set of estimates found that every minute, Facebook users upload 147.000 photos; users upload 500 hours of video on YouTube; Reddit sees 479.452 people engage with its content; 456.000 tweets are posted on Twitter.[1] In order to prevent the Internet from being flooded with unreliable misinformation, child pornography, copyright-infringing material and other harmful or unlawful content, an increasing amount of regulation is generated to cleanse the Internet. However, given the enormous amount of online content generated, the task of editorial overview over all of this content goes beyond human capacity. One would need another world population to monitor, review and censor the online content produced by our current world population. Where is the Heracles that can clean the Augean stables of our present-day Internet? And who should be held liable for the uploading of illegal content? In this contribution, I will explore this question in relation to the content

[1] For an infographic on these impressive data, see: Domo, Data never sleeps 8.0 (2020). <domo.com/learn/infographic/data-never-sleeps-8> accessed 7 August 2021.

moderation of copyright-infringing material and the controversial Article 17 of the Directive (EU) 2019/790 on Copyright in the Digital Single Market (DSM Directive)[2] which establishes a platform liability that, in practice, seems difficult to avoid without the platform taking recourse to automated upload-filters.

The concept of copyright protection is to protect creativity and give copyright holders the power to control reproduction and communication of their works to the public. The enormous amount of user-generated content, as mentioned above, can infringe on these rights when it includes copyright protected content (pictures, text, music, videos, etc.) that are shared in a way that makes them available for viewing, downloading or online distribution.[3]

Online content sharing has become the subject of extensive regulation, in order to protect copyright and limit the widespread issue of piracy in digital space.[4] As argued by Gorwa, "*copyright has historically been one of the first, if not the first, domain where strong economic interests demanded technologies to match and classify online content*"[5]. In copyright law, a distinction is made between primary liability for individual copyright infringers, that is, the users that upload copyright protected material, and secondary liability for third-party intermediaries that facilitate the users in their copyright infringements, for example, platforms like Facebook or Twitter. The EU legislative framework already contains instruments that establish the primary and secondary liability for copyright infringements. In particular, most jurisdictions have provisions whereby third parties can be held liable for contributing to copyright infringement by their

[2] 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.

[3] Brian Fitzgerald et al. Search Engine Liability for Copyright Infringement, in Amanda Spink, Michael Zimmer (eds), *Web Search: Multidisciplinary Perspectives* (Information Science and Knowledge Management, vol. 14 Springer 2008), 104.

[4] As per a EUIPO study from 2018, piracy remains a significant problem within the EU, where the average Internet user accesses pirated content 9.7 times per month in 2018. In fact, pirated video-material gets over 230 billion views a year, and more than 80% of global online piracy can be attributed to illegal streaming services. See: <dataprot.net/statistics/piracy-statistics/> and EUIPO, Online copyright infringement in the European Union music, films and tv (2017–2018). Trends and drivers (2019). <euipo.europa.eu/online_copyright_infringement_in_eu_en.pdf> accessed 7 August 2021.

[5] Robert Gorwa et al. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance* (Big Data & Society Vol. 7 2020).

users. The rationale behind such provisions is that these third parties are often in a better position to discourage infringement, by implementing mechanisms to monitor infringers' activity or, at least, making it more difficult to share copyright protected works.[6]

In particular, the often mentioned and debated Article 17 DSM Directive contains an obligation for certain type of intermediaries – i.e. Online Content Sharing Service Providers[7] – to ensure that copyright infringing uploads made by their users on their platforms are prevented and/or removed through content-filtering procedures, which are based either on information sent by the rights holders or on automatic preventive filtering. If these Intermediaries are not proactive in moderating content and removing infringing uploads, they could be directly liable for copyright infringement.

The main problem for these Intermediaries in complying with the law is the aforementioned immense number of uploads generated daily by their users. Furthermore, things are complicated by the fact that content may be created in one country and viewed in another, thus requiring that Intermediaries create and enforce different legal requirements and cultural policies for each country.

Such requirements for content moderation can only be manageable, thanks to automated filtering technologies based on artificial intelligence.[8] Artificial Intelligence (AI) can play a fundamental role in fighting online copyright infringements, thanks to its 'predict and prevent' approach: algorithms have the ability to rapidly analyse huge amounts

[6] On the economics of intermediaries' liability for copyright infringement, see: Douglas Lichtman and William Landes, *Indirect Liability for Copyright Infringement: An Economic Perspective* (Harvard Journal of Law & Technology Vol. 6 2003).

[7] Although there is a plethora of online intermediaries, each different definition and functions (e.g. Internet service providers, search engines, social media platforms, web hosting providers, etc.), this paper focuses only on Online Content Sharing Service Providers. These Intermediaries are not a new category of online providers in a technological sense. They are instead a new legal category regulated by a body of provisions from the E-Commerce Directive, the InfoSoc Directive, the Enforcement Directive and the DSM Directive. For the sake of readability, they will be referred to by the general term of 'Intermediaries'.

[8] For a detailed overview of the various filtering technologies, see the study requested by the JURI Committee to Giovanni Sartor, Andrea Loreggia: Policy Department for Citizens' Rights and Constitutional Affairs, *The Impact of algorithms for online content. "Upload Filters"* (2020), 35 <europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)657101> accessed 7 August 2021.

of data, identify patterns and proactively make predictions to evaluate if a work contains infringing content. When an algorithm identifies a content as potential copyright infringement, it can automatically proceed with an algorithmic assessment to decide on the best actions to take: these actions are preventive means to either filter the content or to remove the infringing works. The expanding liability of Intermediaries for copyright infringing content posted by their users is pushing the latter to foster the development of AI algorithms (alongside human reviewers) to optimise a speedy detection and removal system.[9]

Nevertheless, AI is not a panacea for all online infringements. The use of AI algorithms by Intermediaries to automatically moderate online content in order to limit their liability has been criticised due to the serious danger of AI likely leading to over-blocking of lawful content, as a collateral effect to automated decision-making. In fact, AI is still incapable of properly interpreting context-related uses, and distinguishing between lawful and unlawful uses, in particular, for cases that might fall under one of the exceptions or limitations provided for by national copyright legislation (such as parody or criticism). Over-blocking based on automated upload filters could thus result in potential limitations to and infringements on fundamental rights (for instance, the right to freedom of expression and of the arts) and basic principles of EU law (such as proportionality and legal certainty).

After having briefly presented the topic of discussion in this introductory Section 1, I will proceed to present in Section 2 the characteristics of AI-based mechanisms for automated algorithmic content moderation and then to introduce technical limitations of AI when used by Intermediaries for the private governance of their platforms. Subsequently, in Section 3, I will illustrate the EU legal framework for Intermediaries' liability, focusing in particular on the application of the aforementioned Article 17 DSM Directive, its safe harbour exceptions – which limit Intermediaries' liability – and the mandatory exceptions and safeguards created to strengthen the rights of copyright holders. Furthermore, I will present a critical analysis of how privately operated algorithms for content moderation might fail to appropriately balance the protection of copyright and fundamental rights, due to inherent limits and flaws of the technology. At last, I will draw conclusions in Section 4, together with

[9] Kirsten Gollatz et al., *The turn to artificial intelligence in governing communication online* (SocArXiv, 2018). <osf.io/preprints/socarxiv/vwpcz> accessed 7 August 2021.

speculations over further development needed for AI-based mechanisms to obtain a suitable balance between copyright protection and legal certainty on Intermediaries' liability.

2 Turning to AI for algorithmic content moderation at scale

As governments, advertisers, and users' pressure on major Intermediaries is growing, both companies and legislators are searching for technical solutions to the difficult puzzle of Intermediaries' governance and online content moderation against copyright infringement. Intermediaries have to handle an enormous volume of data due to the "*stratospheric*" quantity, velocity, and variety of content consumed online,[10] which makes it impossible for them to only rely on prompting human review.

In recent years, AI has been deployed by Intermediaries to reduce the reliance on users to flag content for review, to automatically being able to remove allegedly infringing content, or even filter it out before it is uploaded.[11] As accurately described by Gillespie, "*this link between platforms, moderation, and AI is quickly becoming self-fulfilling: platforms have reached a scale where only AI solutions seem viable; AI solutions allow platforms to grow further.*"[12]

2.1 The promise of AI

AI plays an important role in shaping online content moderation and in helping Intermediaries enforce content moderation with legal certainty, thus determining and proving more clearly where their liability originates or ends. In fact, to stay on the safe side, it might be easier for Intermediaries to hold a 'block-first, verify-later' approach, algorithmically blocking all content that could, even remotely, be infringing copyright. Nevertheless, this would lead to a serious risk of over-blocking lawful contents (as presented in section 3.1 below), which goes against the pur-

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

[11] Op. cit. Kirsten Gollatz (2018).

[12] Tarleton Gillespie, *Content moderation, AI, and the question of scale* (Big Data & Society Vol. 7 2020) 2.

pose of balancing copyright of the rights holders with the rights of users to consume and share lawful content online, without unduly restricting freedoms as collateral effect. This is where algorithmic content moderation and AI come into play.

Algorithmic content moderation can be defined as a governance mechanism enforced by Intermediaries, which use classification of user generated content to implement appropriate choices on how members of a community engage with each other and how content is shared, exploited or removed (e.g. through governance outcomes of removal, geo-blocking, account takedown)[13]. It requires the collection of massive amounts of data from uploaded content and the application of data analytics techniques to identify patterns and make predictions on the best actions to take, to achieve the given governance goals. In the case of online copyright, these goals are to proactively detect, or automatically evaluate whether it contains infringing content, then proceeding with preventive filtering or removal.[14]

Digital and computational methods can be combined usefully with statistical and rich qualitative methods to obtain successful large-scale algorithmic content moderation. These methods span from simpler technical approaches, including keyword filtering (i.e. scanning of a text to identify blacklisted words or phrases stored in a database) and hash matching (i.e. generation of a unique digital fingerprint for previously detected harmful images and videos, to which every new upload is compared to verify its harmfulness),[15] to more sophisticated machine learning-based systems, such as natural language processing (i.e. field of study aiming to enable algorithms to comprehend texts in a more extended

[13] Robert Gorwa et al. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance* (Big Data & Society Vol. 7 2020).

[14] The concepts of proactive detection and automated evaluation are mentioned in a variety of policies and legislations: e.g. "monitoring obligations" (article 15 Directive 2000/31/EC), "notice and stay-down", "upload filtering" (European Digital Rights, Copyright directive: Upload filters strike back. Protecting Digital Freedom (2019), "automatic detection and removal of content" (Conclusions. EUCO 8/17, par. 2, European Council meeting (22 and 23 July 2017). See Emma J. Llansò, *No amount of "AI" in content moderation will solve filtering's prior restraint problem* (Big Data & Society Vol. 7 2020).

[15] See e.g. Microsoft's PhotoDNA tool. <microsoft.com/en-us/photodna> accessed 7 August 2021.

way, closer to the way humans understand text and its context),[16] and optical character recognition (i.e. identification of text in an image and conversion of it into machine-readable format).

Moreover, AI approaches to image and video analysis can be used to detect the presence of pre-identified objects, scenes or elements, such as symbols or logos (i.e. object recognition is the identification of specific pre-defined object classes within an image), for semantic segmentation (i.e. detection and identification of harmful objects and their location by pixels analysis) and scene understanding (i.e. identification of scenes within images, by comparing their dimensional representation to other objects in the image). Other deep-learning methods enable techniques for audio channel separation (i.e. separation of audio sources for deeper analysis) and hash-matching (i.e. identification of audio by comparison to previously categorised audio tracks within a database)[17]. The use of these AI-based technologies allows us to limit the circumvention of filtering by slight alteration of the content in video, images and text (e.g. cropping an image, adding a filter, modifying the lighting conditions or resolutions, rotating/skewing of an element, or modifying the caption could defeat the filter's ability to identify an infringing content).

AI-based algorithms have been deployed in a variety of contexts to protect intellectual property rights. The first example to mention is the 'BookID' system used by Scribd, the subscription-based digital library of e-books and audiobooks. It is described as a system that "*algorithmically analyses computer-readable text for semantic data (such as word counts, letter frequency, phrase comparisons and so on) that it then encodes into a digital "fingerprint". It scans every document uploaded to Scribd and removes those that have the same, or a substantially similar, fingerprint. BookID's approach*

[16] For instance, Google & Jigsaw's Perspective API is an open-source toolkit that allows Intermediaries and users alike to use its machine learning models to evaluate the "toxicity" of a post or comment. <perspectiveapi.com/>accessed 7 August 2021.

[17] For a broader analysis, see: Cambridge Consultants for UK OfCom, Use of AI in Content Moderation, (2019), <ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-content-moderation> accessed 7 August 2021; Emma Llansò et al., Artificial intelligence, content moderation, and freedom of expression, In: *Transatlantic Working Group on Content Moderation Online and Freedom of Expression* (IViR 2019) <ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 7 August 2021.

reduces misidentifications and enables the detection of infringing works even if they have been altered to some degree."[18]

Similarly, Amazon's 'Project Zero', powered by a machine learning algorithm, continuously scans product listing updates to proactively remove suspected counterfeits, based on logos, trademarks, and key data provided by its partnering brands.[19]

A final example: YouTube has experimented since 2006 with a voluntary, automated content monitoring system called 'ContentID', which is formally and procedurally independent from its notice-and-take-down-process (legally necessary to satisfy its obligations under the safe harbour regimes and limit its liability for copyright infringing content uploaded by its users). This algorithm operates on a 'predict and prevent' approach, using a digital fingerprint system: it detects the matching between a newly uploaded video and a protected work, going as far as to monitoring live chats and video meta-data to predict whether an audio or video is a copyright-infringing live stream of sports games. Once a match has been found, rights holders are notified, and they have the ability to either block or take down the content[20]; a third option is to receive a portion of the advertising revenue from the uploaded content.

2.2 Technical limitations and conflicts of private governance of platforms

In an interesting analysis, Elkin-Koren presents how "*overall, content moderation by AI reflects the rise of unchecked private power, which may escape traditional checks and balances intended to ensure that power is exercised in the interest of society at large*"[21]. As automated, privatised, algorith-

[18] Scribd, About the BookID™ Copyright Protection System, 2021 <support.scribd.com/hc/en-us/articles/360037497152-About-the-BookID-Copyright-Protection-System> accessed 7 August 2021.

[19] Amazon, Project Zero leverages the combined strengths of Amazon and brands to drive counterfeits to zero, 2021 <brandservices.amazon.se/projectzero> accessed 7 August 2021.

[20] According to YouTube "*over 98% of copyright issues are handled through Content ID, rather than the notice-and-takedown process. […] as it automatically identified the work and applied the copyright owner's preferred action*". See Google, How Google Fights Piracy, 2018 <storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Piracy_2018.pdf> accessed 7 August 2021.

[21] Niva Elkin-Koren, *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence* (Big Data & Society Vol. 7 2020), p. 2.

mic content moderation systems are more frequently used by Intermediaries for online governance, scholars have raised the concern that these algorithms might take over decision-making power normally assigned to courts and administrative agents.[22] In fact, as described by Heldt in a recent work, algorithmic content moderation systems are endowed with censorial power, bypassing traditional checks and balances secured by the law.[23]

Anticipating what will be discussed more in depth in the following section, it is noteworthy to highlight how these AI-based moderation systems are controversial because of the risk of unduly restricting freedom of expression and of the arts, which are bestowed over users to access, experience and share creative content online (e.g. through scientific publications, cultural assets and news reports)[24].

Consequently, another issue arises from the potential negative outcomes of using algorithms for copyright enforcement, due to prioritization of efficiency over accuracy,[25] which might lead to potential misidentification and over-blocking. As seen further on, algorithmic content moderation faces extensive challenges when context needs to be considered to interpret the meaning of different formats (such as text, images, video or audio). In fact, these heavily automated systems lack contextual sensitivity, having difficulty in identifying subtlety, sarcasm and subcultural meaning,[26] as well as in detecting context and exceptions or limitations provided by the law.

First, limitations arise from the opaqueness of AI algorithms, which makes it harder to ensure that users' rights are adequately protected. There are multiple sources of opacity: to begin with, algorithms and data are often protected as trade secrets, preventing public access in order to

[22] Adam Bridy, Copyright's digital deputies: DMCA-plus enforcement by internet intermediaries. In: John Rothchild (eds.) *Research Handbook on Electronic Commerce Law* (2016), 185–208.

[23] Amélie Pia Heldt, *Upload-filters: Bypassing classical concepts of censorship* (JIPITEC 10 (1) 2019).

[24] Op. cit. Emma Llansò (2019).

[25] Joanne E. Gray, *Google Rules: The History and Future of Copyright under the Influence of Google* (Oxford University Press 2020).

[26] Natasha Duarte et al., *Mixed messages? The limits of automated social media content analysis*, Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR 81:106–106, 2018) <cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf> accessed 7 August 2021.

determine why one piece of content, rather than another, has been subjected to removal.

Secondly, automated decision-making frequently occurs by means of so-called 'black box' algorithms whose predictions are often difficult to interpret – even for the data scientists who designed the system. Moreover, there are additional concerns about transparency, accountability and protection of fundamental rights since the data and training models are often kept confidential by Intermediaries who seek to avoid public scrutiny. In fact, machine learning algorithms are only as good as the datasets they are trained on: if training data does not include a representative number of examples of different languages and different groups or minorities, there will be significant risks of bias and erroneous classifications of underrepresented groups.[27]

Finally, the accuracy of these algorithms is oftentimes embellished: while some Intermediaries indeed use AI for preventive content moderation, most only use a sophisticated version of hash/pattern matching, which could hardly be included under the definition of AI, except under the broadest possible one[28].

3 Principles of secondary liability of online Intermediaries

As governments and users alike continue to strengthen their pressure over Intermediaries to take a more active role in moderating online content, it becomes increasingly important to deploy proper mechanisms to hold Intermediaries to account for copyright infringements originating on their platforms. Throughout the years, various EU and national copyright legislations[29] introduced provisions to encourage Intermediaries to

[27] Op. cit. Tarleton Gillespie (2018).

[28] Researcher Julian Togelius addresses this question well in his blog post: "*There is no such thing as an artificial intelligence. AI is a collection of methods and ideas for building software that can do some of the things that humans can do with their brains. Researchers and developers develop new AI methods (and use existing AI methods) to build software (and sometimes also hardware) that can do something impressive, such as playing a game or drawing pictures of cats*". See Julian Togelius, Some advice for journalists writing about artificial intelligence (2019) <togelius.blogspot.com/2017/07/some-advice-for-journalists-writing.html> accessed 7 August 2021.

[29] In the EU, safe harbour provisions were initially introduced with the 'E-Commerce' Directive 2000/31/EC.

develop automated mechanisms to moderate online content in exchange for exemption from liability for content posted by their users. Recently, this Intermediaries' liability regime has been revised by Article 17 of the DSM Directive. This provision is indeed one of its most controversial provisions, and its new liability regime can be summarised as follows.

First and foremost, what is special about Article 17 is not the characterisation that certain Intermediaries perform acts restricted by copyright,[30] rather how it treats the liability of these Intermediaries.[31] Article 17(1)

[30] This is already provided for by Article 3 of the 'InfoSoc' Directive 2000/29/EC, and CJEU case law.

[31] Article 17(1) to (4) DSM Directive on the use of protected content by online content-sharing service providers states that:

"*1. Member States shall provide that an online content-sharing service provider performs an act of communication to the public or an act of making available to the public for the purposes of this Directive when it gives the public access to copyright-protected works or other protected subject matter uploaded by its users.*

An online content-sharing service provider shall therefore obtain an authorisation from the rightholders referred to in Article 3(1) and (2) of Directive 2001/29/EC, for instance by concluding a licensing agreement, in order to communicate to the public or make available to the public works or other subject matter.

2. Member States shall provide that, where an online content-sharing service provider obtains an authorisation, for instance by concluding a licensing agreement, that authorisation shall also cover acts carried out by users of the services falling within the scope of Article 3 of Directive 2001/29/EC when they are not acting on a commercial basis or where their activity does not generate significant revenues.

3. When an online content-sharing service provider performs an act of communication to the public or an act of making available to the public under the conditions laid down in this Directive, the limitation of liability established in Article 14(1) of Directive 2000/31/EC shall not apply to the situations covered by this Article.

4. If no authorisation is granted, online content-sharing service providers shall be liable for unauthorised acts of communication to the public, including making available to the public, of copyright-protected works and other subject matter, unless the service 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 or other subject matter, and made best efforts to prevent their future uploads in accordance with point (b)."

establishes primary liability for acts of communication, or making available to the public jointly committed by the Intermediaries and its users, which morphs into secondary liability under par. (4) where Intermediaries are liable for infringing content uploaded by their users when failing to obtain the necessary authorisation from rights holders. However, Article 17(4) provides three conditions for Intermediaries to escape liability, by demonstrating to have: i) undertaken 'best efforts' to obtain authorisation; or ii) 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 rights holders have provided the service providers with the relevant and necessary information*" (par. (4)b); iii) acted expeditiously, subsequent to notice from rights holders, to take down infringing content and made best efforts to prevent its future upload[32].

Having initially favoured licensing agreements and preventive authorisations to limit liability, par. (7) of Article 17[33] introduces new mandatory exceptions and limitations applicable to user uploads so that the latter can safely share content online by relying on a general exception for quotation, criticism, review or use for the purpose of caricature, parody or pastiche.[34] Furthermore, a clarification in par. (8) specifies that Article 17 does not entail general monitoring obligations.

[32] This condition seems to introduce a notice-and-takedown mechanism, similar to that of Article 14 E-Commerce Directive and a notice-and-stay-down (or re-upload filtering) obligation for Intermediaries.

[33] Article 17(7) DSM Directive states: "*The cooperation between online content-sharing service providers and rightholders shall not result in the prevention of the availability of works or other subject matter uploaded by users, which do not infringe copyright and related rights, including where such works or other subject matter are covered by an exception or limitation. Member States shall ensure that users in each Member State are able to rely on any of the following existing exceptions or limitations when uploading and making available content generated by users on online content-sharing services: (a) quotation, criticism, review; (b) use for the purpose of caricature, parody or pastiche*".

[34] These new mandatory exceptions and limitations operate alongside the one provided for by Article 5(3) of 'InfoSoc' Directive. In situations of conflict (i.e. an exception is explicitly mentioned in Article 17(7) but unavailable at the national level ex InfoSoc Directive), the former creates an obligation under EU law to transpose under national legislation these exceptions and limitations.

Lastly, par. (9)[35] introduces three safeguards to protect users and to minimise the risks of broad filtering and over-blocking.[36] First, any request by rights holders for the removal of specific content must be justified; second, it requires Member States (while transposing the Directive) to ensure that Intermediaries put in place "*effective and expeditious complaint and redress mechanisms*" which users can avail themselves of in case of disputes over contentious *"decisions to disable access to or remove uploaded content*" (which should be subject to human review); third, Member States should create out-of-court dispute settlement mechanisms, which are independent of the judicial redress. This is in order to guarantee that Intermediaries optimise algorithmic mechanisms for the uniform protection of fundamental rights and freedoms across the EU.[37]

After having presented the novelties of Article 17, I now introduce two technical issues arising from practical application of this provision: i) the risk of over-blocking due to automated preventive filtering and ii) the

[35] As written in Article 17(9) DSM Directive: "*Member States shall provide that online content-sharing service providers put in place an effective and expeditious complaint and redress mechanism that is available to users of their services in the event of disputes over the disabling of access to, or the removal of, works or other subject matter uploaded by them.*

Where rightholders request to have access to their specific works or other subject matter disabled or to have those works or other subject matter removed, they shall duly justify the reasons for their requests. Complaints submitted under the mechanism provided for in the first subparagraph shall be processed without undue delay, and decisions to disable access to or remove uploaded content shall be subject to human review. Member States shall also ensure that out-of-court redress mechanisms are available for the settlement of disputes. Such mechanisms shall enable disputes to be settled impartially and shall not deprive the user of the legal protection afforded by national law, without prejudice to the rights of users to have recourse to efficient judicial remedies. In particular, Member States shall ensure that users have access to a court or another relevant judicial authority to assert the use of an exception or limitation to copyright and related rights.

This Directive shall in no way affect legitimate uses, such as uses under exceptions or limitations provided for in Union law, and shall not lead to any identification of individual users nor to the processing of personal data, except in accordance with Directive 2002/58/EC and Regulation (EU) 2016/679.

Online content-sharing service providers shall inform their users in their terms and conditions that they can use works and other subject matter under exceptions or limitations to copyright and related rights provided for in Union law".

[36] João Pedro Quintais et al., *Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive* (JIPITEC 10(3), 2019) 277–282.

[37] Krzysztof Garstka, Guiding the Blind Bloodhonds: How to mitigate the risks Article 17 of Directive 2019/970 poses to the freedom of expression in: in: Paul Torremans (eds.) *Intellectual Property and Human Rights* (Wolters Kluwer Law & Business 2020) 335.

risk of broad limitations to lawful content due to AI's lack of contextual sensitivity to detect exceptions and limitations.

3.1 Automated preventive filtering and risk of over-blocking

There is an internal conflict within the systematic structure of Article 17. Specifically, par. (7) provides that the cooperation between rights holders and Intermediaries – presented in par. (4) – shall not prevent *ex ante* the availability of content uploaded by users which does not infringe copyright including, especially if it is covered by an exception or limitation.[38] At the same time, par. (4)(b) encourages Intermediaries to make preventive "*best efforts*" to ensure the unavailability of specific works, in order to avoid secondary liability. Here originates the issue from the use of AI-based algorithms for preventive filtering: the obligation to ensure that users can upload lawful content, while preventing copyright infringing uploads, is extremely difficult to realise with automated means, especially in cases of context-contingent uses under copyright exceptions or limitations. Things are more complicated when trying to program into an AI system the hierarchy between Article 17(7), formulated as an absolute standard (*"shall not result in the prevention of the availability of works or other subject matter uploaded by users"*), and 17(4), which is based on a relative criterion such as the "*best efforts*" obligation to obtain authorisation or make the content unavailable expeditiously.

As a consequence of regulatory and stakeholders' pressure on Intermediaries to provide an easier mechanism towards infringing content removal, and since the coming into force of the DMS Directive in 2019, which affected the monitoring obligations of Intermediaries, the Intermediaries have largely implemented AI-based automated preventive filtering and blocking of content at the point of upload, before it is even made available to the public.[39] This general, algorithmic filtering (that leverages machine learning to restrict upload *ex ante*) is difficult to justify as it might result in over-blocking of lawful uses of content. As seen fur-

[38] See very clearly in this sense, with references to the protection of the fundamental rights of users, Recital 70 DSM Directive.

[39] Martin Senftleben, *Institutionalized Algorithmic Enforcement – The Pros and Cons of the EU Approach to UGC Platform Liability* (Florida International University Law Review 14 2020) 299–328.

ther, it might even account as a form of censorship since it could cause disproportionate consequences and detrimental effects on users' freedoms in comparison with the protection of copyright holders required by Article 17.

Similarly, in its recent Guidance on Article 17, the EU Commission has shifted from a position that rejected *ex ante* blocking of content to a more permissive take towards *ex ante* blocking beyond manifestly illegal content.[40] By allowing rights holders to 'earmark' content "*unauthorised online availability of which could cause significant economic harm to them*"[41] they can circumvent the principle that automatic blocking should be limited only to manifestly infringing uses. Consequently, uploads that include 'earmarked' protected content do not benefit from the *ex ante* protections for likely legitimate uses, allowing Intermediaries to use AI-based filters to block its upload from the beginning. Thus, the Guidance promotes a switch to a system based on privately governed mechanisms of preventive monitoring and enforcement of automated filtering, which undermines the principle that automated filtering cannot limit lawful upload and overcome the use of exceptions and limitations.

This approach of the EU legislator that sees automated algorithmic filtering as a necessary consequence for Intermediaries to discharge their monitoring obligations, even if in combination with other non-automated mechanisms,[42] is based on the misconception that AI might be able to solve all copyright enforcement problems. Instead, in the current state of the technology, algorithmic content moderation is not as sophisticated as believed. It is best to keep in mind how extremely difficult it is to programme into AI-based automated systems all contextual factors needed to be assessed to avoid overenforcement by filtering, as shown by

[40] Communication from the Commission to the European Parliament and the Council: Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market, COM/2021/288 Final ('Guidance').

[41] "*When providing the relevant and necessary information to the service providers, rightholders may choose to identify specific content which is protected by copyright and related rights, the unauthorised online availability of which could cause significant economic harm to them. The prior earmarking by rightholders of such content may be a factor to be taken into account when assessing whether online content-sharing service providers have made their best efforts to ensure the unavailability of this specific content and whether they have done so in compliance with the safeguards for legitimate uses under Article 17(7), as explained in part VI below*" Guidance, section V.2. p.14.

[42] Gerald Spindler, *The Liability system of Art. 17 DSMD and national implementation – contravening prohibition of general monitoring duties?* (JIPITEC 10 2020) 356.

recent empirical studies of automated copyright enforcement that report substantial over-blocking of content on video sharing platforms.[43] In any case, although Article 17(8) is very clear in stating that the fulfilment of the Intermediaries' obligations shall not lead to a general monitoring obligation. Intermediaries should not presume the infringing nature of contents. Thus, the availability of uploaded content for the public should be limited by fully automated filtering only for cases of manifestly infringing uploads (i.e. material that is identical or equivalent to the 'earmarked' content, previously requested by the rights holders).

3.2 Automated preventive filtering and lack of contextual sensitivity to detect exceptions and limitations

When fulfilling their monitoring obligations, Intermediaries must be careful not to allow algorithms to restrict users' rights to lawfully share and access information. Recital 70 DSM Directive explicitly recognises the importance of striking a balance between the right to intellectual property (Article 17(2)) and the fundamental freedom of expression and freedom of the arts, respectively, under Articles 11 and 13 of the Charter of fundamental rights of the EU[44]. By introducing this balance, the EU legislator has decided to award special status to these new mandatory exceptions and limitations, grounding their basis in fundamental rights.[45]

Achieving this balance between different fundamental freedoms and rights is largely dependent on the technological solutions that Intermediaries will employ to discharge their obligations. Without further repetition, it is worth underlining here the potential conflicts between the required monitoring obligations and the risk of misidentification and over-blocking when using algorithms for content moderation – due to their lack of contextual sensitivity in detecting specific context-related

[43] See e.g. Sharon Bar-Ziv, Niva Elkin-Koren, *Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown* (Connecticut Law Review, Vol. 50, 2017); or Kris Erickson and Martin Kretschmer, This video is unavailable (JIPITEC 9(1)2018).

[44] Article 11 of the Charter on freedom of expression and information states: "1. *Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. 2. The freedom and pluralism of the media shall be respected.*"

Article 13 of the Charter on Freedom of the arts and sciences affirms: "*The arts and scientific research shall be free of constraint. Academic freedom shall be respected.*"

[45] See e.g. op. cit. Emma Llansò et al. (2019).

elements. In fact, automated, AI-based filters are unable to recognise contextual nuances, which are necessary to distinguish *prima facie* infringements from uses that fall within the scope of exceptions or limitations provided by the law (e.g. reproduction of a part of a work for parodic use or permitted quotation).

A similar approach is portrayed in the EU Advocate General's opinion in the recent case C-401/19, through which the Polish Government has filed an action for annulment of Article 17 due to violation of freedom of expression under Article 11 of the EU Charter. According to previous CJEU case law which rejected general monitoring obligations that would monitor all the transmissions within a network[46], the AG describes how the 'generality' of an obligation will not have to be determined by the amount of information processed, but by the specific content that is being surveyed. He then illustrates how this provision actually imposes a 'specific' monitoring obligation to 'ensure the unavailability of specific works and other subject matter' previously earmarked by the rights holders ex Article 17(4). Any other conclusion (such as considering this a 'general' monitoring obligation) "*de facto obliges an intermediary provider to filter, using software tools, all of the information uploaded by the users of its service, even if it is a matter of searching for specific infringements, (and it) would regrettably amount to ignoring the technological developments which make such filtering possible and to depriving the EU legislature of a useful means of combating certain types of illegal content*"[47].

In light of the above, whether these concerns can be mitigated with effective and appropriate technological measures will be decisive in combatting unduly restrictions of fundamental freedoms.[48] Quintais et al. note that the application of preventive algorithmic content moderation is only possible as long as a proper filtering technology is available on the market and meets the legal requirements set forth in Article 17. In essence, preventive algorithmic filtering should only be allowed if it: (i)

[46] *Scarlet Extended SA v. Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM),* C-70/10, ECLI:EU:C:2011:771; *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV*, C-360/10, ECLI:EU:C:2012:85; *Tobias Mc Fadden v, Sony Music Entertainment Germany GmbH*, C-484/14, ECLI:EU:C:2016:689.

[47] Advocate General's Opinion in Case C-401/19, Poland v Parliament and Council, 15 July 2021, point 113.

[48] Sere e.g. Christophe Geiger and Bernd Justin Jütte, *Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match* (GRUR International, Vol 70, 2021).

meets the proportionality requirements in paragraph 17(5); (ii) enables the recognition of the mandatory exceptions and limitations in paragraph 17(7), including their contextual and dynamic aspects; and (iii) in no way affects legitimate uses, as mandated in paragraph 17(7) and (9).[49]

3.3 The future legal framework on Intermediaries' liability

As seen above, Article 17 is an extremely complex legal provision. As Dusollier notes, it is the "*monster provision of the Directive, both by its size and hazardousness*"[50]. The difficulties in interpreting the provision, and for Intermediaries to apply it correctly, in order to exclude liability is shown by the significant legal scholarship already existing, most of which was written even before the national implementation deadline[51].

To complicate things, on 15 December 2020, the EU Commission proposed two legislative initiatives to upgrade rules governing digital services in the EU to create a safer and more open digital space "*in which the fundamental rights of all users of digital services are protected*"[52]. The first of these proposals introduces Regulation on a Single Market for Digital Services (Digital Service Act), which provides a new regulatory approach to online Intermediaries through horizontal rules interlacing with a variety of EU legislations. For the purpose of this paper, it is relevant to highlight how the overlap between the Digital Service Act and the DSM Directive will shape the future legal framework around Intermediaries' liability and will impact how the latter shall programme or update their algorithms for automated filtering and content moderation, in order to enjoy the relevant liability exemptions.

[49] Op. cit. João Pedro Quintais, et al. (2019).

[50] Séverine Dusollier, *The 2019 Directive on copyright in the digital single market: some progress, a few bad choices, and an overall failed ambition* (Common Market Law Review, Vol. 57 2020).

[51] For a compilation of interventions and publications see: <create.ac.uk/cdsm-implementation-resource-page/#consultations-transpositions> accessed 7 August 2021. At the time of writing, the implementation of the DSM Directive at national level has been quite slow with only two Member States having fully completed the transposition into national law, partially also due to the discussions and uncertainty involving this provision.

[52] Proposal for a regulation of the European Parliament and of the Council on a Single Market For Digital Services ('Digital Services Act') and amending Directive 2000/31/EC, COM/2020/825 final.

First of all, both the DSM Directive and the proposed Digital Services Act establish obligations on various online intermediaries (including those that are central to the argument of this paper, i.e. online content sharing service providers) on how to handle illegal information. The DSM Directive targets copyright infringing content and the Digital Services Act targets illegal content in general (including content which infringes on copyright). Despite the fact that the two instruments have a different legal nature (the DSM Directive will have to be transposed in Member State law, whereas the Digital Services Act is a directly applicable Regulation), and that they seem to operate at different, complementary levels, it is worth mentioning potential overlaps between the two legislations and present how the Intermediaries' liability framework might look like in the future, if and when the Digital Service Act might enter into force.

At first sight, these regimes do not particularly overlap with each other since Article 17 is *lex specialis*, as per Recital 9 and Article 1(5)(c) Digital Service Act. The latter states that the Regulation is "*without prejudice to the rules laid down by (…) Union law on copyright and related rights*"[53]. However, in the view of certain scholars, this unaffected result "*can only relate to aspects which indeed are specifically covered by those* (copyright) *rules*"[54]. In fact, the intersection between the DSM Directive and the Digital Service Act is more complex than what Recital 9 Digital Service Act and Article 1(5)(c) Digital Service Act seems to suggest at first sight. The EU Commission provided to the Council's Working Party on Intellectual Property (Copyright) some internal insights on how the relationship between the two instruments can be interpreted.[55] The Commission

[53] Recital 9 Digital Service Act states that "*This Regulation should complement, yet not affect the application of rules resulting from other acts of Union law regulating certain aspects of the provision of intermediary services […]. Therefore, this Regulation leaves those other acts, which are to be considered lex specialist in relation to the generally applicable framework set out in this Regulation, unaffected. However, the rules of this Regulation apply in respect of issues that are not or not fully addressed by those other acts as well as issues on which those other acts leave Member States the possibility of adopting certain measures at national level".* Supporting Recital 11 Digital Service Act adds that the "*Regulation is without prejudice to the rules of Union law on copyright and related rights, which establish specific rules and procedures that should remain unaffected*".

[54] João Pedro Quintais et al., Interim report on mapping of EU legal framework and intermediaries' practices on copyright content moderation and removal, ReCreating Europe (2021), p. 43.

[55] Council of the European Union, Working Paper, N° Cion doc.: 14124/20, Digital Services Act and EU copyright legislation – Information from the Commission (2021).

has advised that the "*Digital Service Act is not an IPR enforcement tool*" given its general and horizontal nature. Nevertheless, it "*includes a full toolbox which can be very useful for the enforcement of IPR*" and should be applied "*without prejudice to existing IPR rules*". In short, it looks like the Commission upholds that Article 17 DSM Directive will remain "*unaffected*" by the rules on liability proposed in the Digital Service Act. Only time will tell if that will be the case. As of now, without digressing too much on this topic, the Digital Service Act is believed to apply to Intermediaries only insofar as it contains rules that regulate matters not covered by Article 17 DSM Directive, or in cases of specific matters which Article 17 leaves to the discretion of Member States.[56]

4 Conclusions

As seen above, due to the pressures from governments and users alike to take a more active role, more and more Intermediaries are turning towards AI to moderate online content on a large scale, since the deployment of AI-based algorithms can give rise to successful, automated detection, evaluation and removal of infringing content.

Accordingly, the EU legislator has increasingly become more accepting of the idea of imposing preventive filtering obligations on Intermediaries to screen out copyright infringing content and to hold Intermediaries accountable for copyright infringements originating from their users' activity. This, thanks to the assumption that algorithmic filtering technologies have become more sophisticated.

Having presented the legal framework and the interpretative issues around Article 17, I showed how AI algorithms play a fundamental role in helping fight online copyright infringements thanks to a 'predict and prevent' approach: considering the immense amount of data generated daily by users, these algorithms can quickly analyse huge amounts of data, proactively evaluate if a work contains infringing content, and then remove it.

Nevertheless, this approach, which sees automated algorithmic filtering as a necessary consequence for Intermediaries to discharge their mon-

[56] This includes e.g. rules from the Digital Service Act relating to the liability and to due diligence obligations for online Intermediaries of different sizes. For a speculative interpretation see: João Pedro Quintais, Sebastian, Felix Schwemer, *The interplay between the digital services act and sector regulation: how special is copyright*? (Forthcoming 2021).

itoring obligations, is based on the misconception that AI technology might be able to solve all copyright enforcement problems. Instead, in the current state, AI technology is not as sophisticated as believed and surely not a panacea for all issues related to online infringements. AI is still characterised by many technical limitations which conflict with the protection of users' fundamental rights and freedom. First, it might present serious danger of over-blocking lawful contents, unduly restricting freedoms as collateral effect of the 'block-first, verify-later' type of approach. Secondly, AI lacks contextual sensitivity, necessary to distinguish *prima facie* infringements from uses that fall within the scope of exceptions or limitations provided by the law. Whether these concerns can be mitigated with effective and appropriate technological measures will be decisive in combatting unduly restrictions of fundamental rights and freedoms by AI-based filtering algorithms.

In light of the above, the legislators (both at EU and national level) should therefore act with caution and avoid narratives about all-powerful algorithms, instead helping to shape users' online experience through provisions that combine the deployment of AI filtering algorithms together with safeguards of fundamental rights and freedoms of users. It is thus paramount in the future to introduce standards to enhance transparency and accountability of content moderation practices (e.g. introducing secondary human-review, due diligence processes and other risk assessment methodologies), and to ensure that users have access to complaint and redress mechanisms, to remedy wrongly executed automated decisions by AI.