ReCreating Europe

Impact of content moderation practices and technologies on access and diversity

D.6.3 Final Evaluation and Measuring Report

Authors

Sebastian Felix Schwemer, Christian Katzenbach, Daria Dergacheva, Thomas Riis, João Pedro Quintais

Deliverable Title	D.6.3 Final Evaluation and Measuring Report - impact of moderation practices and technologies on access and diversity
Deliverable Lead:	UvA (IViR), HIIG/UBremen, UCPH (CIIR)
Partner(s) involved:	UvA (IViR), HIIG/UBremen, UCPH (CIIR), SZG
Related Work Package:	WP 6 - Intermediaries: Copyright Content Moderation and Removal at Scale in the Digital Single Market: What Impact on Access to Culture?
Related Task/Subtask:	T6.3 - Evaluating Legal Frameworks on the Different Levels (EU vs. national, public vs. private) (Leaders: CPH, UvA. Contributors: USZ) T6.4 - Measuring the impact of moderation practices and technologies on access and diversity
Main Author(s):	Sebastian Felix Schwemer (UCPH, CIIR) Christian Katzenbach (UBremen) Daria Dergacheva (UBremen) Thomas Riis (UCPH, CIIR) João Pedro Quintais (UvA, IViR)
Other Author(s):	–
Dissemination Level:	Public
Due Delivery Date:	30.12.2022
Actual Delivery:	To peer-review: 18.11.2022; final: 31.01.2023

Project ID:	870626
Instrument:	H2020-SC6-GOVERNANCE-2019
Start Date of Project:	01.01.2020
Duration:	39 months

Version history table			
Version	Date	Modification reason	Modifier(s)
v.01	12/07/2022	First draft	Sebastian Schwemer (UCPH), Christian Katzenbach (UBremen)
v.02	18/11/2022	Draft version for internal peer review	Christian Katzenbach (UBremen), Sebastian Schwemer (UCPH)
v.03	31/01/2023	Final report (peer review comments received 18/01/2023)	Christian Katzenbach (UBremen), Sebastian Schwemer (UCPH)

TABLE OF CONTENTS

ABBREVIATIONS

AI	Artificial Intelligence
AI Act [Proposal]	Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM/2021/206 final
CDSM Directive	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
CFR	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
CMO	Collective Rights Management Organisation
DMCA	Digital Millennium Copyright Act
DSA Proposal	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
DSA	Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)
DSM	Digital Single Market
e-Commerce Directive	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 [2000] OJ L178/1
EC	European Commission
ECL	Extended Collective License(s)
ECS	European Copyright Society
E&Ls	Exception and /or limitations
Enforcement Directive	Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (OJ L 157, 30.4.2004)
EP	European Parliament

EU	European Union
GDPR	General Data Protection Regulation (GDPR)
IGF	Internet Governance Forum
InfoSoc Directive	Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society OJ L 167, 22.6.2001, p. 10–19
IP	Intellectual Property
ISSPs	Information society service providers
NTD	Notice-and-takedown
NSD	Notice-and-staydown
OCSSP	Online content-sharing service provider
P2B	Platform to Business
Technical Standards Directive	Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services (Text with EEA relevance) OJ L 241, 17.9.2015, p. 1–15
TEU	Treaty on the European Union
T&Cs	Terms and Conditions
TRIPS	Agreement on Trade Related Aspects in intellectual Property
UGC	User-generated content
UDHR	Universal Declaration of Human Rights
VLOP	Very Large Online Platform
VLOSE	Very Large Online Search Engine
WBM	Internet Archive's WayBack Machine
WIPO	World Intellectual Property Organisation
WCT	WIPO Copyright Treaty
WPPT	WIPO Performances and Phonograms Treaty

EXECUTIVE SUMMARY

This Report presents the results of research carried out as part of Work Package 6 "Intermediaries: Copyright Content Moderation and Removal at Scale in the Digital Single Market: What Impact on Access to Culture?" of the project "ReCreating Europe", particularly on Tasks 6.3 (Evaluating Legal Frameworks on the Different Levels (EU vs. national, public vs. private) and 6.4 (Measuring the impact of moderation practices and technologies on access and diversity). This work centers on a normative analysis of the existing public and private legal frameworks with regard to intermediaries and cultural diversity, and on the actual impact on intermediaries' content moderation on diversity.

Chapter 2 deals with the evaluation of legal frameworks on the different levels. First, the chapter expands on the assessment of the regulatory environment and revisits the starting point for access to culture and the creation of cultural value. It introduces a concept of "Rough Justice", which acknowledges the difficulties and differences vis-à-vis a full "fair trial" setup and proposes conceptualization in the context of procedural rules, substantive rules and competences. A second starting point for the legal evaluation is provided in analysing and evaluating the framework for quality of automated copyright content moderation as put forward in the CDSM Directive and the Digital Services Act in light of erroneous decisions. It is suggested that decision quality should be a decisive factor that is to be seen as a separate perspective from ex post mitigation mechanisms. It also analyses the benchmark put forward in the sector specific CDSM Directive and the horizontal Digital Services Act. A third perspective relates to the aspect that copyright content moderation increasingly requires an understanding of contextual use and the potential risk of "bias carry-over" from datasets to content moderation. It is suggested that the question bias mitigation and access to copyright data should increasingly be addressed in the regulatory framework.

Chapter 3 describes our efforts to measure the impact of copyright content moderation on access and diversity. We start this chapter by presenting existing research in the field and by discussing options to investigate these complex questions. On these grounds, we explain our research design consisting of three empirical sub-studies, and then present the results of this

work. In the first sub-study we investigate aggregated data on copyright and content moderation published by platforms themselves, often in the form of transparency reports; secondly, we analyse content level data with regard to the sustaining availability and the diversity of content on social media platforms; and thirdly we present results from in-depth interviews with cultural creators with regard to their experiences with copyright content moderation. Overall, the results indicate a strong impact of copyright regulation and content moderation on diversity, and potentially an impact that leads to a decrease in diversity of content. Yet, the research has also shown that these interpretations cannot be fully verified based on the limited data that is available to researchers and the public.

Chapter 4 presents joint conclusion based on the evaluation of the existing legal frameworks as well as existing practices and technologies. We particularly highlight the need for further research on issues of diversity and access on social media platforms, given its high relevance for European societies, and at the same time its complex nature, specifically in the context of contemporary fragmented media landscapes. We conclude with a strong call for robust mandatory data access clauses in future regulations.

1. INTRODUCTION

In the context of WP6 on **"Intermediaries: Copyright Content Moderation and Removal at Scale in the Digital Single Market: What Impact on Access to Culture?"**, we pursue two principal objectives, namely to:

- *Explain and evaluate the existing legal frameworks* (both public and private, existing and proposed) that shape the role of intermediaries in organising the circulation of culture and creative works in Europe, including content moderation and removal at scale.

- *Explain, critically examine and evaluate the existing practices and technologies* that intermediaries deploy to organise the circulation of culture and creative works in Europe, including content moderation and removal at scale.

This Report (D6.3 *Final Evaluation and Measuring Report - impact of moderation practices and technologies on access and diversity*) describes the results of the research carried out in the context of the normative analysis of both the legal framework (public and private) as well as the existing practices and technologies. It builds upon the earlier research on the mapping of the EU legal framework and intermediaries' practices on copyright content moderation and removal,[1] and on the research conducted during task T6.3 *Evaluating Legal Frameworks on the Different Levels (EU vs. national, public vs. private)* and T6.4 *Measuring the impact of moderation practices and technologies on access and diversity*.

This report is structured as follows: In lieu of a comprehensive report, we attach the draft articles[2] based on our research.

- Thomas Riis: "A theory of rough justice for internet intermediaries from the perspective of EU copyright law"

[1] Available here.
[2] All articles are or will be submitted to relevant peer-reviewed journals.

- Sebastian Felix Schwemer: “Quality of Automated Content Moderation: Regulatory Routes for Mitigating Errors”
- Thomas Margoni, João Pedro Quintais and Sebastian Felix Schwemer: “Algorithmic propagation: do property rights in data increase bias in content moderation?”
- Christian Katzenbach, Selim Basoglu and Dennis Redeker: “Finally Opening up? The evolution of transparency reporting practices of social media platforms”, submitted to ICA 2023.
- Daria Dergacheva, Christian Katzenbach: “Mandate to Overblock? Understanding the impact of EU’s Art. 17 on automated content moderation on YouTube”, submitted to ICA 2023.
- Daria Dergacheva, Christian Katzenbach and Paloma Viejo Otero: “Losing authenticity: social media creators’ perspective on copyright restrictions in the EU” submitted to ICA 2023.

These research articles are accompanied by brief descriptions in the following chapters.

Chapter 2 deals with the evaluation of legal frameworks on the different levels. First, the chapter expands on the assessment of regulatory environment and revisits the starting point for access to culture and the creation of cultural value. It introduces a concept of “Rough Justice”, which acknowledges the difficulties and differences vis-à-vis a full “fair trial” setup and examines this in the context of procedural rules, substantive rules and competences. A second starting point for the legal evaluation is provided in analysing and evaluating the framework for quality of automated copyright content moderation as put forward in the CDSM Directive and the Digital Services Act in light of error. A third perspective relates to the aspect that copyright content moderation increasingly requires an understanding of contextual use and whether the potential risk of “bias carry-over” from datasets to content moderation is sufficiently addressed in the current framework.

Chapter 3 describes our efforts to measure the impact of copyright content moderation on access and diversity. We start this chapter by presenting existing research in the field and by discussing options to investigate these complex questions. On these grounds, we explain our

research design consisting of three empirical sub-studies, and then present the results of this work. In the first sub-study we investigate aggregated data on copyright and content moderation published by platforms themselves, often in the form of transparency reports; secondly, we analyse content level data with regard to the sustaining availability and the diversity of content on social media platforms; and thirdly we present results from in-depth interviews with cultural creators with regard to their experiences with copyright content moderation.

Chapter 4 presents joint conclusion based on the evaluation of the existing legal frameworks as well as existing practices and technologies. We particularly highlight the need for further research on issues of diversity and access on social media platforms, given its high relevance for European societies, and at the same time its complex nature, specifically in the context of contemporary fragmented media landscapes. We conclude with a strong call for robust mandatory data access clauses in future regulations.

Further research partly related to or conducted within this work package and subtasks has been published in earlier writing:

- SF Schwemer (2022), "Digital Services Act: A Reform of the e-Commerce Directive and Much More" prepared for A Savin, *Research Handbook on EU Internet Law (Edward Elgar, 2023)*, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213014;
- SF Schwemer (2022), "Recommender Systems in the EU: from Responsibility to Regulation", *Morals & Machines* (2022), 60–69, https://doi.org/10.5771/2747-5174-2021-2-60;
- Quintais, J., & Schwemer, SF (2022). "The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?" *European Journal of Risk Regulation, 13*(2), 191-217. doi:10.1017/err.2022.1;

- Katzenbach, C. (2021). “AI will fix this” – The Technical, Discursive, and Political Turn to AI in Governing Communication. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211046182.

2. EVALUATING LEGAL FRAMEWORKS ON THE DIFFERENT LEVELS (T6.3)

The evaluation of the legal frameworks on different levels (T6.3) integrates the findings of T6.1 and T6.2[3] and carries out a normative assessment of how legal rules and contractual terms on the moderation and removal of copyright content on large-scale user-generated content (UGC) platforms affect digital access to culture and the creation of cultural value.

It assesses how such rules and terms shape the design of removal and moderation by UGC platforms, the activities of creators and users, and the role of fundamental rights and freedoms – namely the freedom of expression, the arts and to conduct a business – in shaping these rules and terms. It also evaluates how the state-enacted rules in the DSM shape the emergence of private models for content moderation and removal, examining how the production of law is shaped by the intrinsic characteristics and needs of the actors on the DSM within the legal framework conditions.

2.1 THE COMPLEXITY AND ELASTICITY OF THE LEGAL FRAMEWORK: OVERLAPS AND INTERPLAY

The Digital Services Act is the first European framework to provide a legal definition of "content moderation" (Article 3 lit. t DSA): it refers to the "the activities, automated or not, undertaken by providers of intermediary services aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or the recipients' ability to provide that information, such as the termination or suspension of a recipient's account."[4] In essence, content moderation can contain a large variety of activities that address content that is illegal or deemed incompatible with the private regulatory framework by that

[3] See Quintais, JP; Mezei, P; Harkai, I; Magalhães, J; Katzenbach, C; Schwemer, SF and Riis, T, Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis (ReCreating Europe, 2022). http://dx.doi.org/10.2139/ssrn.4210278.

[4] On further definitions and conceptualisations of content moderation, see, Quintais et al. (2022), p 33 ff.

respective platform in form of terms and conditions.[5] It is also distinct from content *recommendation*.[6]

As analysed in our mapping report, the legal framework for copyright content moderation consists of several parts: The relevant **substantive copyright rules** are contained in national copyright legislation, partly based on harmonising instruments such as the InfoSoc Directive. The relevant rules regarding **intermediary** or **platform regulation**, are contained in Article 17 of the CDSM Directive (and its national implementations), the e-Commerce Directive's framework for intermediary liability exemptions in Articles 12-15, which will be replaced and amended by the DSA.

In order to understand the regulatory, i.e., both law and self-regulatory, environment round the moderation of online content, it is necessary to recall that Article 14 e-Commerce Directive sets forth the horizontal basic rules for an intermediary's mandated response to illegal content, including copyright- infringing works. These rules will be replaced by the corresponding provision in the DSA on 17 February 2024.[7]

Notably, the e-Commerce Directive refrained from further specifying the notice-and-action regime. In this void (or more positively: freedom of operation) industry-practices have merged. These, in, turn, now appear to at least partly codified in the CDSM Directive with

[5] Note also the broad definition of terms and conditions in the Art. 2 lit. u DSA, which covers "all clauses, irrespective of their name or form, which govern the contractual relationship between the provider of intermediary services and the recipients of the service".
[6] Cf. Art. 2(o) DSA. See also Quintais et al. (2022), p. 35f.
[7] See for an in-depth comparison SF Schwemer, "Digital Services Act: A Reform of the e-Commerce Directive and Much More", prepared for A Savin, *Research Handbook on EU Internet Law* (Edward Elgar, 2023), available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213014.

regards to OCSSPs, and in the DSA with regards to other online platforms that fall outside the scope of Article 17 CDSM Directive.

One issue related to the regulatory framework regards its complexity and potential overlaps and interplay. This is specifically relevant in the context of online platforms and copyright, where both Article 17 CDSM Directive and the Digital Services Act specify and adjust platforms' room of operation for content moderation and which we have previously explored.[8] Further complexity is added with the specific national implementations of Article 17 CDSM Directive as previously analysed.[9]

Besides this overlap, there are notable other areas where rules interact. Since content moderation often also involves the processing of personal data, for example, future research should look into the interplay between the GDPR and the sector specific CDSM Directive framework as well as the horizontal rules in the DSA. Since content moderation is –as explored earlier– regularly performed or supported by algorithmic means, furthermore, also the potential intersection with the Artificial Intelligence Act (AIA), a Regulation which was proposed on 21 April 2021, is of interest.[10] It introduces "rules regulating the placing on the market and putting into service of certain AI systems" (recital 4 AIA) and focusses on the regulation of the provider as well as the user of such AI system. In our context of copyright content moderation, the AIA is of interest given the broad and generic definition of AI system in Art. 3(1) AIA, which means "software that is developed with one or more of the techniques

[8] Quintais, J., & Schwemer, S. (2022). The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright? *European Journal of Risk Regulation, 13*(2), 191-217. doi:10.1017/err.2022.1

[9] http://copyrightblog.kluweriplaw.com/2021/11/29/whats-the-buzz-tell-me-whats-a-happening-around-article-17-tales-from-hungary-germany-italy-and-sweden/

[10] See Thomas Margoni, João Pedro Quintais and Sebastian Felix Schwemer: "Algorithmic propagation: do property rights in data increase bias in content moderation? Part II" (*Kluwer Copyright Blog*, 9.6.2022) http://copyrightblog.kluweriplaw.com/2022/06/09/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-ii/

and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with".[11] Suffice it here to note that content moderation technology likely falls within the scope of this definition. Furthermore, the scope the proposed Regulation focuses on risks inter alia to the protection of fundamental rights of natural persons concerned (see, e.g., recitals 1, 13, 27, 32, Arts. 7(1)(b), 65 AIA). Copyright content moderation might come with risks for inter alia freedom of expression or the arts. The AIA differentiates between four types of risk: AI systems that come with unacceptable risks are prohibited; AI systems with high-risk are permitted but subject to specific obligations; AI systems with limited risk are subject to certain transparency obligations. Neither, however, seems to encompass copyright content moderation at this stage.

2.2 THE QUEST FOR BENCHMARKS FOR NORMATIVE ASSESSMENTS

2.2.1 ACCESS TO CULTURE & CREATION OF CULTURAL VALUE

The main research theme of the ReCreating Europe project relates to the concept of "access to culture" within the context of a "culturally diverse, accessible, and creative Europe".[12] We have already at the mapping stage attempted to locate our research of WP6 in the context of these concepts.

Cultural diversity, accessibility and creativity are cornerstones of the EU. Art. 3(3) of the Treaty on European Union (TEU), for example, sets out that the Union "shall respect its rich cultural and linguistic diversity, and shall ensure that Europe's cultural heritage is safeguarded and enhanced." Instead of providing an in-depth analysis of the concepts and their deep and rich

[11] This definition, is complemented by Annex I, which contains a detailed list of approaches and techniques for the development of AI.

[12] reCreating Europe, 'ReCreating Europe - The Project: Discover the 4 Pillars of ReCreating Europe' <https://www.recreating.eu/the-project/> accessed 15 November 2022.

history, however, we focus on the dimensions most relevant for the analysis of online platforms engaging in copyright content moderation.

The further analysis is based on our hypothesis that *cultural diversity* is both a property of as well as in an interdependent relationship with *access to culture*. Similarly, possibilities for unfolding *creativity* are in an interdependent relationship with access to culture. In the following, we focus on our analysis on access to culture.

At an international level, it is possible to identify a basis for a concept of access to culture as it relates to copyright in art. 27 of the UN Universal Declaration of Human Rights (UDHR). Art. 27 states that

> *(1) Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits'.* [13]
>
> *(2) Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.'*

In simple terms, from a legal technical perspective, copyright law predominantly *excludes access* to protected works for the purpose appropriating economic value from those works. Since many expressions of cultural phenomena and artifacts are protected by copyright, there is an inherent conflict of interests between copyright and access to culture. In EU law, as interpreted by the CJEU, this is complicated by a relatively low threshold of originality for the copyright protection of works, broad exclusive rights and enforcement measures recognised to rights holders, and relatively narrow exceptions and limitations to the benefit of users.[14]

Online platforms constitute an important gateway for accessing protected content. The stress field of copyright-protected content and online platforms also concerns several fundamental

[13] Cf. art. 15 of the International Covenant on Economic, Social and Cultural Rights.
[14] There is ample scholarship describing these aspects. For recent overviews, see e.g. Eleonora Rosati, *Copyright and the Court of Justice of the European Union* (Oxford University Press 2019); Tito Rendas, *Exceptions in EU Copyright Law: In Search of a Balance Between Flexibility and Legal Certainty | Wolters Kluwer Legal & Regulatory* (Kluwer Law International 2021).

rights in the Charter of Fundamental Rights of the European Union (CFREU)[15], such as the right to property (art. 17(2) CFREU), the rights to privacy and data protection (arts 7 and 8 CFREU), the rights to freedom of expression (art. 11 CFREU), freedom of arts (art. 13 CFR), and the freedom to conduct a business (art. 16 CFREU).

Notably in the context of content moderation, the Digital Services Act now integrates fundamental rights deeper into platforms' operations. In the context of intermediary service providers' terms and conditions, Article 14(4) DSA requires them to "act in a diligent, objective and proportionate manner in applying and enforcing the restrictions (…) [in form of content moderation], with *due regard to the rights and legitimate interests of all parties* involved, including the *fundamental rights* of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter."[16] This is notable also against the background that Article 17 CDSM Directive not directly invokes fundamental rights but mere introduces them through the balancing mechanism of (mandatory) limitations and exceptions in Article 17(7) CDSM Directive. Thus, online platforms may in the future have to a larger degree consider perspectives such as access to culture.

For the purpose of our analysis, we previously proposed to distinguish between two dimensions of the concept of "access to culture" in relation to copyright content moderation by online platforms: a descriptive dimension; and a normative dimension of the concept. The former is addressed in a discussion on the quality of content moderation and error in decision making (section 2.2.4). The latter is addressed in a discussion on a theory of "rough justice" (section 2.2.3) also reflecting upon the fundamental right of fair trial. Both represent approaches and attempts from a legal perspective against to which assess the impact of copyright content moderation on access to culture.

[15] See, e.g., recital 84 CDSMD: "This Directive respects the fundamental rights and observes the principles recognised in particular by the Charter. Accordingly, this Directive should be interpreted and applied in accordance with those rights and principles."

[16] Emphasis added. See, e.g., N Appelman, JP Quintais and R Fahy, "Using Terms and Conditions to apply Fundamental Rights to Content Moderation" (*Verfassungsblog*, 1.9.2021) https://verfassungsblog.de/power-dsa-dma-06/

2.2.3 A CONCEPT OF "ROUGH JUSTICE"

This section is based on the draft article by Thomas Riis "A theory of rough justice for internet intermediaries from the perspective of EU copyright law", attached to this report.

The purpose of this report is to develop a model that can be used to say something meaningful about the quality of the legal framework that shapes the actual content moderation practices. It tries to evaluate the legal framework for the purpose of posing normative statements on how to improve the legal framework. In order to that is needed a value-based measuring scale. Common values in rights-enforcement and human rights can be used in such a measuring scale. One place to look for common values, is in the traditional legal perception of fair trial that includes values such as predictability, contradiction, production and presentation of evidence etc.. However, in relation to platforms' content moderation practices, for all practical purposes, it is not possible to ensure the relatively high level of due process known from traditional civil procedure.

The level of justice in traditional civil procedure cannot, just like that, be integrated into platforms' content moderation practices because it will simply be too burdensome for all practical reasons and require too many resources. Therefore, there is a need to modify the traditional conception of justice in the context of internet platforms and in this article such a modification is called 'rough justice'. A model of rough justice does not presume to provide full justice but is significantly better than no justice.[17]

The concept of fair trial is anchored in human rights law. Article 47 of the EU Charter consists of three parts. The first part establishes a right to an effective remedy.[18] The third part ensures legal aid to those who lack sufficient resources. The second part that is the most pertinent for the purpose of this article and the concept of justice reads:

[17] Peter Linzer, Rough Justice: A Theory of Restitution and Reliance, Contracts and Torts, Wis. L. Rev. 695-775 (2001), p. 766.
[18] Cf. art. 13 of European Convention on Human Rights.

'Everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal previously established by law. Everyone shall have the possibility of being advised, defended and represented.'[19]

Among the constitutive elements of fair trial, from a human rights perspective, the most important ones are the right to a fair hearing and information, the equality of arms principle and the right to a reasoned judgment. The right to a fair hearing is an essential element in the rights of the defence and, in particular, it implies that a party to a dispute shall have the opportunity to examine and comment on the facts and documents that a judicial decision is based on.[20] The principle of equality of arms basically means that there should be ensured a 'fair balance' between the parties in a dispute resolution process and that both parties are treated equally.[21] The right to a reasoned judgment requires that all judgments be reasoned to enable the defendant to see why judgment has been pronounced against him and to bring an appropriate and effective appeal against it.[22] Another rationale behind the right to a reasoned judgment is to ensure publicity of the legal reasoning that enables the public to predict valid law.[23]

A conception of rough justice on internet platforms must address two major general problems. The first one relates to the accuracy of the moderation practices. There are two types of content that are subject to content moderation: 1) illegal content and 2) incompatible content which is legal content that is deemed incompatible with the platform's terms and conditions.

As the point of departure, the optimal content moderation scheme can perfectly identify illegal and incompatible content and moderate it accordingly which means that it neither

[19] See Angela Ward, 'Article 47 – Right to and Effective Remedy and to a Fair Trial', in Steve Peers, Tamara Hervey, Jeff Kenner and Angela Ward (Eds.), 'The EU Charter of Fundamental Rights: A Commentary' 2nd ed. (Bloomsbury Publishing 2022), at 47.19.
[20] Case C-199/99, *P Corus UK Ltd,* paras. 19 and 41. See also C-348/16, *Sacko*, EU:C:2017:591, para. 37.
[21] Judgment of 29 May 1986, *Feldbrugge v. The Netherlands*, Application no. 8562/79, para. 44.
[22] Case C-619/10, *Trade Agency Ltd*, ECLI:EU:C:2012:531, para. 53.
[23] Monique Hazelhorst, 'The Right to a Fair Trial in Civil Cases' (Springer – Asser Press 2017), p. 150 f.

under-enforce nor over-enforce substantive law (illegal) and furthermore neither under-enforce nor over-enforce terms and conditions (incompatible content).

The second major problem concerns the inherent privatization of justice. Privatization of justice results when enforcement of rights is left to a private party and it can imply a distortion of the balancing of interests in substantive law. That happens when platforms' policies stipulate that otherwise legal content is deemed incompatible with the terms and conditions and for that reason is not available to the users.

All rights-enforcement systems shall take three general objectives into consideration. The first one is 'efficacy'. That objective is for instance incorporated into art. 47(1) of the EU Charter on Fundamental Rights (and it is part of Title VI of the Charter which has the headline 'Justice'. It implies firstly that there is access to justice in the sense that mechanisms for rights-enforcement shall be easily available and not overly costly. Secondly, effective remedies shall be available to redress wrongs.

The second objective is 'fair trial' which is found in art. 47(2) of the EU Charter.[24] 'Fair trial' is a more complex concept than efficacy and, in addition to the constitutive elements mentioned above, it includes such things as consistency and predictability in rights-enforcement, proportionality. Finally, the third objective is a balanced use of resources in rights-enforcement. That refers to the fact that law enforcement requires resources and the 'price' of enforcement may be too high. The price of enforcement shall be balanced against the costs of being mistaken. Where the "the cost of being mistaken" equals the harm of non-enforcement. And if the costs of being mistaken are too high, more resources must be allocated to rights-enforcement. A balanced use of resources in rights-enforcement does not have the same solid legal foundation in the principles of fundamental rights as efficacy and due process but it follows from a quite obvious common-sense notion that at a certain point the required resources for full enforcement and a high level of fair trial can be too many.

[24] Art. 6(1) of the European Convention on Human Rights.

A number of attempts to establish codes for fair trial on the internet have been presented. The article addresses three such attempts: 1) The Santa Clara Principles 2.0, 2) The Aequitas Principles on Online Due Process and 3) The Council of Europe's recommendation on the roles and responsibilities of internet intermediaries. Paragraph 2.3. of the Recommendation is dedicated to Content moderation. The three codes are quite diverse in their scope and substance. In addition to the codes, the provisions on content moderation in the recently adopted Digital Services Act is examined. These provisions deal with some of the same issues as the codes. The codes and the Digital Services Act are analysed from a critical perspective on the basis of the human rights approached to justice with a specific view to.

1. A substantial human rights norm to prevent over-enforcement
2. Transparency
3. Fair trial

The criticism of the codes and Digital Services Act is used a steppingstone for recommendation for improvements. Recommendations are also made on how to solve the problem associated with privatization of justice. The reason why privatization of justice is problematic is that private parties substitute public rules with private rules public and public rules pursue societal objectives and values whereas private rules must be assumed to pursue private objectives and values. The recommendations are the constitutive parts of the model of rough justice combined bits and pieces from the codes and the Digital Services Act.

The model on rough justice is divided into three different parts: 1) Procedural rules, 2) Substantive rules and 3) Competences.

In respect of procedural rules, first of all, there is a need for more transparency into how content moderation works. Transparency enables a person whose content has been moderated to obtain an explanation for the reasons behind the moderation.[25] Furthermore,

[25] James Grimmelmann, "Regulation by Software," Yale Law Journal (2005), p. 1737. See e.g., Nicolas P. Suzor, Sarah Myers West, Andrew Quodling and Jillian York, 'What Do We Mean When We Talk About Transparency?

it provides an opportunity to detect and address errors and biases and enhance accuracy. Finally, it enables persons to assess the quality of the moderation processes and thus the legitimacy of the result of the moderation processes. Transparency should cover the functioning of algorithms and the logic behind and working conditions of humans in cases where humans are involved. At the moment nothing suggests that platforms voluntarily are willing to ensure transparency in respect of their algorithms. In most – perhaps all - cases platforms are able to protect their algorithms as trade secrets under the EU Trade Secrets Directive. Transparency thus requires legislative intervention that exempts algorithms for content moderation from trade secrets protection.

It must be acknowledged that the circumstances surrounding the appeal process in content moderation is not comparable to the traditional perception of fair trial and significant limitations in the procedure must be accepted. That relates for instance to the types of evidence admissible, the extent of evidence, and the number of pleadings. It does not follow explicitly from the various code etc. how many stages in the moderation and appeal process are needed in order to comply with the codes. The various instruments seem to presume two stages: a complaint/automatic removal and a subsequent counter notice which works as an appeal which seems to be a reasonable and manageable solution. The most problematic in such a process, as it is assumed in the DSA, is that in the first stage the user's content will be moderated before the user is heard which fits uneasily with the equality of arms principle. The hearing of users before content is moderated require sharing platforms to invest more resources in content moderation and, hence, it is unlikely that the platforms will do this voluntarily which suggests legislative intervention.[26]

The purpose of substantive rules on what to moderate is twofold. Firstly, it shall create a counter-weight to platforms' tendency to over-enforce. Secondly, it shall reduce moderation of incompatible but legal content. Substantial rules based on human rights would be an important means to align the platforms' terms of services to societal objectives and value and

Toward Meaningful Transparency in Commercial Content Moderation', International Journal of Communication 13(2019), 1526–1543.

[26] Cf. Rory Van Loo, Federal Rules of Platform Procedure, p. 849

thus counteract the adverse effect of privatization of justice. International human rights law is binding on states only, not on individuals or companies. Therefore, it is recommended that an obligation to fully respect human right are imposed on platforms, for instance by making international human rights directly applicable to platforms that moderate content.

The last part of the model on rough justice concerns the competences of humans who are involved in content moderation. Human competences impact the quality in the content moderation system. The codes and the Digital Services Act require human review in the appeal process. Automated content moderation involves a risk of biases – both original biases and developed biases and errors. To reduce biases and errors and thereby ensure accuracy in the first stage automatic moderation, there must be a certain level of human involvement. Hence, should be obligated to institute random test of accuracy by human intervention.

Finally, human competences must be ensured by adequate training and working conditions. When it comes to working conditions, the most essential issue is how much time is allocated to each human decision. The codes and the Digital Services Act are not very precise on requirements to human competences but simply state that appropriate training and working conditions shall be provided. Clearly it would be very difficult to establish appropriate and precise criteria for professional qualifications and working conditions. More important than setting up precise standards for qualifications and working conditions, is to impose an obligation on the platforms to inform on the internal criteria for appropriate qualifications and working conditions (transparency), so the users of the platform themselves are able to assess the legitimacy of the content moderation process.

2.2.4 QUALITY OF (AUTOMATED) CONTENT MODERATION: ERROR

This part is based on the draft article by Sebastian Schwemer "Quality of Automated Content Moderation: Regulatory Routes for Mitigating Errors", attached to this report.

In an overly simplistic worldview, in any (copyright and beyond) content moderation scenario there should be a "right" and a "wrong" outcome. This is because such decision basically

answers the question "is this illegal" and there should be only one answer.[27] Sometimes such decision might be straightforward, for example, in instances where content –in the words of the European Commission– is "manifestly illegal".[28] Sometimes such decision might require a detailed assessment by domain experts or even call for the involvement of the courts. In any case, however, we should be able to assess the "quality" of such content moderation decision.

The question then is what the benchmark for decision quality in copyright content moderation is. In the context of content moderated by online platforms, there are in principle several starting points:[29]

- the existing substantive legal rules which are "operationalized" by (automated or manual) content moderation.
- the existing private rules *inter partes* based on contract, be it in the form of Terms and Conditions or Community Guidelines. Related to this is compliance with those rules (i.e., their enforcement by the platform) including any secondary legislation regarding this.
- A final route, namely users' normative perception of either legal rules or policies by online platforms is outside the scope of this paper.

For the sake of argument of this approach, it sets out the assumption that the European copyright framework regarding substantive rights represent the starting point of "optimal"

[27] This is disregarding the fact that, empirically, the perception by creative content producers may not correspond to the legal reality.

[28] See, e.g. in the context of See European Commission, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, Brussels, 4.6.2021 COM(2021) 288 final, pp. 20-23. See also AG Øe, who refers to "information the unlawfulness of which *is obvious from the outset*, that is to say, it is *manifest*, without, inter alia, the need for contextualization", Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 198 and also para. 201.

[29] A third perspective is omitted in this article: namely, instead of taking existing rules as starting point, requiring content moderation to consider how rules ought to be in a *de lege ferenda* or political perspective.

regulation.[30] The intermediary liability (exemption) framework then is a second layer that can be adjusted for changing intermediaries'/platforms' behaviour.

Transferred to the context of copyright content moderation[31] by online platforms and the impact on access to culture this means the following. The "quality" of copyright content moderation is correlated to access to culture, because access to culture (as per the definition above) is considered embedded in the existing copyright framework. Since the existing framework is assumed to strike the appropriate balance between exclusivity in copyright protection and access to culture, any variation in that balance – beyond the margin of interpretation allowed by law – will impact on access to culture. Consequently, both excessive and insufficient content moderation will have a negative impact on access to culture. Simply put, excessive content moderation by platforms restricts access to culture. Conversely, insufficient content moderation increases access to culture, but in a harmful way because it encroaches on the legitimate interest of copyright holders and thus distorts the optimal balance. In other words: the smaller the difference between actual content moderation performed by intermediaries and the correct application of the legal framework, the smaller the negative impact on access to culture.

The consequence of this assumption is that the "quality" of content moderation can in simple terms be described in terms of correct and false results. The first set of outcomes that relates to correct result of content moderation (i.e., the absence of error). The second set of outcomes relates to false results of content moderation (i.e., the presence of error). The principal question, however, relates to the question what error rate is acceptable under the legislative framework.

[30] See also Quintais et al. (2022), p. 55. By "optimal" regulation we mean in this context that the framework strikes the *appropriate* balance between conflicting interests and fundamental rights, namely by recognizing time-restricted (exclusive) rights and corresponding exceptions and limitations. By "appropriate" we mean the balance that was struck as a result of the normal operation of a democratic legislative process. In other words, we do not mean to pass a value judgment on the desirability of such balance from the perspective of any normative theory or viewpoint about copyright law.

[31] The use of copyright moderation for non-copyright purposes and the use of non-copyright moderation (e.g. privacy) for copyright purposes is not addressed in the following.

- The **Digital Services Act**, a horizontal (i.e., not-copyright-specific) framework that will apply to all intermediary service providers, for example, addresses content moderation error rates in the context of several provisions. In relation to voluntary measures by online platforms[32] to ensure the unavailability of illegal (in our context: copyright-infringing[33]) content, recital 26 DSA states that automation technology must be "sufficiently reliable to limit to the maximum extent possible the rate of errors". In yet another context in relation to transparency reporting, Article 15(1)(e) DSA obliges intermediary service providers [34] to include in their transparency reporting information on "any use made of automated means for the purpose of content moderation, a qualitative description, a specification of the precise purposes, indicators of the *accuracy and the possible rate of error* of the automated means used in fulfilling those purposes, and any safeguards applied".[35] Both examples underline the crucial role of error in content moderation. They, however, also imply, that error rates are not (and need not be) equal zero. Importantly, the DSA does not differentiate between type-I (false-positive) or type-II errors (false-negative).
- The **copyright-specific framework for online content sharing service providers** (OCSSP), too, can be understood as addressing decision quality. Firstly, Article 17(4) lit. b CDSM Directive requires OCSSPs' best efforts to ensure the unavailability of specific works in accordance with high industry standards of professional diligence.[36] Article 17(7) CDSM Directive states that content moderation "shall not result in the

[32] Also referred to as good Samaritan actions, see Article 7 DSA.

[33] On the applicability of the horizontal DSA vis-à-vis the sector-specific regulation of online content sharing service providers in the CDSM Directive, see, JP Quintais and SF Schwemer, The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright? *European Journal of Risk Regulation, 13*(2), 191-217. doi:10.1017/err.2022.1; A Peukert et al, "European Copyright Society – Comment on Copyright and the Digital Services Act Proposal" *IIC - International Review of Intellectual Property and Competition Law*, 2022, 53(3), p. 358-376; E Rosati, "The Digital Services Act and Copyright Enforcement: The Case of Article 17 of the DSM Directive" in *Unravelling the Digital Services Act Package* (Strasbourg, European Audiovisual Observatory 2021).

[34] Which includes not only (very large) online platforms but notably also "regularly" hosting service providers and even mere conduit or caching providers; on content moderation outside the platform see, e.g., SF Schwemer, "Location, location, location! Copyright content moderation at non-content layers" in E Rosati, *The Routledge Handbook of EU Copyright Law* (Routledge, 2021), 378–395.

[35] Emphasis added.

[36] In other words, it sets the standard for false-negatives.

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." Read in conjunction with Article 17(9) para. 3 CDSM Directive it seems that the standard is stricter than that: it notes in a more straight-forward fashion that the Directive "shall in no way affect legitimate uses, such as uses under exceptions or limitations provided for in Union law (…)". Furthermore, the provision in Article 17(7) CDSM Directive also harmonises the mandatory limitations and exceptions for quotation, criticism, review and the use for the purpose of caricature, parody or pastiche. In this context, the **European Commission's Guidance on Article 17** notes that "to restore legitimate content ex post (…) once it has been removed or disabled" would "not be enough for the transposition and application of Article 17(7)".[37] Therefore, "automated blocking, i.e. preventing the upload by the use of technology, should in principle be limited to manifestly infringing uploads".[38]

- The question of copyright content moderation quality and error in the context of this provision was also touched upon by Advocate-General Øe in his opinion in **Case C-401/19**, *Poland v Parliament and Council*. Øe points out that Article 17(7) CDSM Directive "does not mean that the mechanisms which lead to a negligible number of cases of 'false positives' are automatically contrary to that provision".[39] Yet, the AG notes that error rates "should be as low as possible".[40] Therefore, AG Øe argues that in situations where the current technological state of the art for automatic filtering tools is not sufficiently advanced to prevent a significant false-positive rate, the use of such tool should be precluded.[41] In conclusion, Article 17 CDSM Directive contains both indicators as to the acceptable error rate for false-negatives and false-positives.

[37] See European Commission, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, Brussels, 4.6.2021 COM(2021) 288 final, p. 20.
[38] Ibid.
[39] Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 214.
[40] Ibid.
[41] Ibid.

It is, however, noteworthy that over-blocking –i.e. a higher false-positive rate– according to AG Øe may be justified in certain cases in in light of "effectiveness of the protection of the rights of rightholders" in light of the CJEU's case law.[42] Thus, the acceptable error rate for false-positives does not necessary correspond to that of false-negatives in copyright content moderation by OCSSPs.

In any case, however, the issue of error rates in all the above examples can only consist of a contextual analysis. A first factor should relate to the volume of content moderation decisions taken. The goal cannot only be to have a low percentage of error (error rate) but rather a low number of actual "wrong" content moderation decisions. A second factor should relate to the "harm" caused by the wrong decision (and whether such harm can be mitigated ex-post).

2.2.5 CONTEXT AND BIAS IN CONTENT MODERATION[43]

This part is based on the research conducted in a collaboration between WP6 and WP3 of the ReCreating Europe project. Preliminary results have been published in *Kluwer Copyright Blog* (attached to this report) and a journal article is under preparation.

In the context of copyright content moderation and OCSSPs, Article 17 CDSM Directive, in simple terms, incentivizes certain platforms to filter content uploaded by users to comply with their "best efforts" obligations to deploy preventive measures against infringing content. Prior to the introduction of this legal regime, some platforms already "voluntarily" relied on similar automated content moderation (e.g., YouTube's' ContentID). At the current state of technology, filtering appears to be done mainly through matching and fingerprinting. However, these tools are incapable of assessing **contextual uses**.[44] Therefore, they are not suitable to ensure the required protection of freedom of expression-based exceptions like parody, criticism and review, as required by Article 17(7) CDSM Directive. Accordingly, more

[42] Ibid, para. 183.
[43] This research is based on a collaboration between WP6 and WP3 of the ReCreating Europe project.
[44] See, e.g., https://communia-association.org/2019/12/03/article-17-stakeholder-dialogue-day-3-filters-not-meet-requirements-directive/

sophisticated tools seem necessary to enable preventive measures while respecting users' rights and freedoms, as recently confirmed by the CJEU in case C-401/19.[45] This suggests that machine learning algorithms may increasingly be employed for copyright content moderation given their alleged superiority in identifying (understanding?) contextual uses.

Against this background, a crucial question emerges for the future of (copyright) online content moderation and fundamental rights in the EU: what happens when these tools are based on "biased" datasets? More specifically, if it is plausible that any bias, errors or inaccuracies present in the original datasets be carried over in some form onto the filtering tools developed on those data: (1) How do property rights in data influence this "bias carry-over effect"? and (2) what measure (transparency, verifiability, replicability, etc.) can and should be adopted to mitigate this undesirable effect in copyright content moderation in order to ensure an effective protection of fundamental rights?

Based on this, we explore the possible links between conditional data access regimes and content moderation performed through data-intensive technologies such as fingerprinting and, within the realm of "artificial intelligence" or rather machine learning algorithms. More specifically, we look at whether current EU copyright rules may have the effect of favoring the propagation of bias present in input data to the algorithmic tools employed for content moderation and what kind of measures could be adopted to mitigate this effect. Algorithmic content moderation is a powerful tool that may contribute to a fairer use of copyright material online. However, it may also embed most of the bias, errors and inaccuracies that characterize the information it has been trained on. Therefore, if the user rights contained in Article 17(7) CDSM Directive are to be given an effective protection, simply indicating the expected results omitting *how* to reach them, may not be sufficient. The problem of over-blocking is not simply a technical or technological issue. It is a cultural, social and economic issue, as well and, perhaps more than anything, it is a power dynamic issue. Recognizing parody, criticisms and review as "user rights", as the CJEU does in C-401/19, may be a first step towards the strengthening of users' prerogatives. But the road to reach a situation of power symmetry

[45] See, e.g., https://verfassungsblog.de/filters-poland/

with platforms and right holders seems a long one. Ensuring that bias and errors concealed in technological opacity do not circumvent such recognition and render Article 17(7) ineffective in practice would be a logical second step.

2.3 CONCLUSIONS

The research conducted in the context of this chapter has shown that the existing legal framework has increasingly focus on how it shapes the role of intermediaries in organising the circulation of culture and creative works in Europe, including content moderation and removal at scale. The *assessment of the existing legal frameworks* that shape the role of online platforms in organising the circulation of culture and creative works in Europe through content moderation has shown the complex landscape of interacting rules that meet and address the reality of content moderation at scale. It suggests that rather than envisioning a private-regulatory copy of a "full trial" setup, a different conceptual approach of "rough justice" is necessary to catch the developments and in this context suggests recommendations for improvements regarding procedural rules, substantive rules and competences. Furthermore, it shows that with regards to access to culture and cultural diversity, decision quality should be emphasised as a separate factor from ex post mitigation mechanisms. Both the DSA and the CDSM Directive (including case law) provide starting points for this. The analysis also points to the fact that content moderation increasingly requires an understanding of contextual use, but further work is needed on the potential risk of "bias carry-over" from datasets to content moderation. In this context, it is also worthwhile to point out that content moderation technology appears to be only stepmotherly treated in currently ongoing negotiations of the AI Act.

A central issue relevant in the context of copyright content moderation relates to the possibility to study content moderation. Whereas transparency arguably does solve the issue of decision quality (see above) and is only indirect connected to the rules that regard the content moderation and removal at scale by online platforms *as such*, it is a necessary condition for its study. The German implementation of Article 17 CDSM Directive for example contains a clause for researchers' access. Questions remain, however, how to operationalise

this in order to make access for researchers available, while maintaining online platforms' trade secrets and protecting users' fundamental rights, notably with regards to privacy and the protection of personal data. Also the horizontal framework of the DSA will enable data access to VLOPs and VLOSEs for "vetted researchers" under certain conditions.[46] Under Article 40(4) in conjunction with Article 40(8) DSA on data access and scrutiny, researchers can be granted the status of "vetted researchers" for the "sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union (…) and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures (…)" put in place for VLOPs and VLOSEs.

Finally, whereas this research has focussed on issues of content *moderation*, we note the – related but separate issue of content *recommendation*. Whereas the actual phenomena are somewhat related, however, they relate to a different set of issues and perspectives. We note that more research is needed in the field of copyright content recommendation as well as copyright's role in content recommendation with a view to access and diversity.

[46] See Article 40(8) lit. a-g DSA.

3. MEASURING THE IMPACT OF MODERATION PRACTICES AND TECHNOLOGIES ON ACCESS AND DIVERSITY (T6.4)

The move to pass the European Copyright in the Digital Single Market (CDSM) Directive (2019/790) has been happening while social media platforms have clearly become key players in contemporary societies (van Dijk, 2013), and AI technologies are increasingly presented as solutions to the major societal problems (Katzenbach, 2021). Under increasing public and political pressure, social media platforms have massively expanded their efforts to monitor and moderate content on their sites. Platforms have invested strongly in both quickly growing their teams of content moderators (Roberts, 2019) as well as in algorithmic systems to automatically govern contested content (Gorwa et al., 2020),

What is the impact of these increasing content moderation practices, policies, and technologies, including for copyright, and of the CSDM Directive on access to culture and diversity? Legal scholarship and social science work has clearly identified that they are highly relevant for the future role of platforms as intermediaries and their impact on cultural diversity and access to culture. The directive specifically raises the level of platforms' liability for the content that they host in cases of copyright infringement. For large platforms now "automated content filtering is required to comply with the best-efforts obligations in Article 17(4) CDSMD" (D6.2 Quintais et al. 2022). Critics' concern of mandatory upload filters and potential structural overblocking thus seemed to be real. At the same time, the issue of private platforms algorithmic moderation systems remains opaque if not completely intransparent, including institutional and legal issues (Perel & Elkin-Koren, 2017; Gray & Suzor 2020). All this raises the stakes for understanding better, how platforms and copyright content moderation impacts diversity and access to culture.

In this part of the project, we have investigated first of all the aggregated data on copyright moderation provided by the platforms themselves. Secondly, we have analysed content level data of platforms with regard to changes and factors of cultural diversity on social media and streaming platforms, specifically YouTube. And finally, we have explored creators'

understanding and experiences of copyright moderation in relation to their creative work and the labor of media production on social media platforms.

3.1 EXISTING RESEARCH ON DIVERSITY, CONTENT MODERATION AND ALGORITHMS

This research builds on existing research on access to content and cultural diversity in media economics, sociology and communication sciences. Cultural diversity in the movie industry (Moreau & Peltier, 2004; Lévy-Hartmann, 2011), TV networks Hellman, 2001; McDonnald & Lynn, 2004), recording companies (Ranaivoson, 2010), publishing (Benhamou and Peltier, 2007), and broadcasting (Farchy and Ranaivoson, 2011). Also more recent studies have explored exposure diversity in terms of algorithmic news curation on social media platforms (Wojcieszak et al, 2022; Jurgers & Stark, 2022).

Scholarship on copyright takedowns is multidisciplinary as well, mobilizing a variety of methodologies and relying on data from different platforms. Scholars have studied notice and takedown procedures in Google search (Bar-Ziv & & Elkin-Koren,2017; Urban et al: 2017), takedown practices with regard to YouTube's Content ID system (Tushnet: 2014; Edwards, 2018; Erickson and Kretschmer, 2018), as well as with regard to Amazon's Kindle (Tushnet: 2014). Gray and Suzor (2020) have conducted the largest study on YouTube to date, aiming to understanding copyright moderation and over-blocking by the platform. They were able to collect a random sample of metadata text for 76.7 million YouTube videos.

Broader studies of mapping YouTube and its content include a massive data collection and analysis study by Rieder et al. (2020). This was the first, and due to further inaccessibility of platform's data, perhaps the last large-scale description of YouTube's content universe. More qualitatively, Burgess and Green (2009 and 2018) have characterized YouTube as one of the world's most powerful digital platforms, which combines the logic of community and commerciality. YouTube's core structure and network was also examined by Paolillo (2008) through tags attached to the published videos. Bärtl (2018) attempted to present an overall

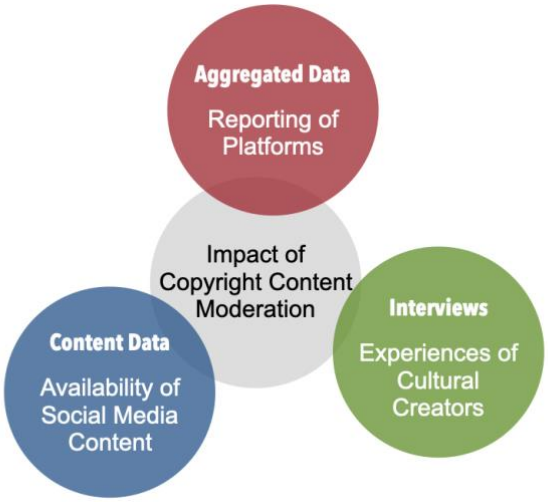
characterization of YouTube over the course of the past 10 years, based on a random sample of channel and video data.

As clearly shown, algorithmic systems play a critical role in curating and moderating content on platforms, scholarship on 'critical algorithm studies' (Gillespie and Seaver, 2015) is also relevant for this study. Critical assessment of algorithmic governance in platform studies (Gorwa et al. 2021) has shown that the process is not transparent and easily understood by creators or understood at all (Eslami et al., 2015; Poell et al., 2021). Caplan & Boyd (2018) highlight companies' institutional dependency on algorithmic intermediaries. Willson (2017) and Hallinan and Striphas (2016) explore the place of algorithmic content distribution in contemporary 'everyday'. Bucher (2017) identifies affective dimensions of algorithms, and shows how algorithms contribute to the establishment of participatory norms through validation and punishment (Bucher, 2012). Duffy et al., 2021 map out methods of studying platformization of the cultural industries.

3.2 RESEARCH DESIGN OF THE EMPIRICAL STUDY

Building on this existing research we have assessed different options to assess the impact of copyright regulation and content moderation on diversity and access to culture. Unfortunately, the most the adequate option was no longer viable when starting the empirical work. Gray and Suzor (2020) had assesses the life-circle of content on YouTube by tapping into YouTube's API, collecting a random sample of content directly at the moment of upload. In defined periods later, they checked for the availability of those content items. YouTube's APIv2 at that same provided information about reasons if content was no longer available. On these grounds, researchers could in fact evaluate the actual scope and effects of copyright content moderation. Unfortunately, due to restrictions of access to data in current YouTube's API v.3, this option no longer exists.

Against this background we have developed a research design that technically circles around the key question at hand, taking three different approaches investigating data on different levels:



We have investigated the *aggregated data* on copyright and content moderation that platforms themselves publish; we have analysed *content level* data with regard to the sustaining availability and the diversity of content on social media platforms; and we have *interviewed cultural creators* with regard to their experiences with copyright content moderation.

In the first step, we have compared *aggregated data from transparency reports* published by major platforms present in the European Union. In this sub-study we have analysed both the kinds of data platforms have started to disclose in the recent years, as well as the substantial numbers on copyright content moderation.

In the second step, we have analysed on the *content level* availability and diversity of content on a selected platform, YouTube. For a timeframe from 2019 (before CDSM) and 2022, we have analyses both the scale of copyright-based content deletion and blocking, as well as measured differences in the diversity of content available on the platform across time and selected countries.

Further on we collected samples of channels from all the four countries, and their descriptions, in order to compare the changes in diversity supply that happened from 2019 till 2022.

Finally, we conducted semi-structured interviews with creators on various platforms: the sample was derived from those taking part in the survey on digitalization of creative work from the same project ReCreating Europe (Poort & Pervaist, 2022).

3.3 ASSESSING TRANSPARENCY REPORTS

This section represents a short summary of the paper "Finally Opening Up? The Evolution Of Transparency Reporting Practices Of Social Media Platforms" attached in annex of this report.

3.3.1 INTRODUCTION

In this first study we investigate the historical evolution and current situation of transparency reporting with a focus on copyright-based content moderation and examines the convergence and divergence in social media platforms' content moderation practices along with the transparency habits in a broader sense also by elaborating on substantial numbers of content moderation data. To that end, we discuss the general trend toward greater accountability of non-state actors and its relevance to legitimacy, besides the demands of civic society for more transparency. Thereafter, we outline our empirical approach, first focusing on our research design and then on definitions of social media platform reporting criteria, including specifically those for copyright-based content moderation.

3.3.2 MAKING CONTENT REMOVALS TRANSPARENT: COPYRIGHT ENFORCEMENT AND BEYOND

Recent years have seen a rise in so-called transparency reports by platforms, reporting on different aspect of their activities including content moderation. This development needs to be considered in the context of broader debate on accountability of platforms. Accountability of non-state actors is an important topic in the governance literature. Transnational non-state

actors have not had the attention as public organizations had until the late twentieth century (Redeker & Martens, 2018). Due to rising power of Internet-based corporations, the discussion of transparency has further increased, though. The goal of platforms when engaging transparency-increasing measures is to gain legitimacy relating to the "right to govern" in the eyes of the users and politically powerful stakeholders. Civil society organizations have put out several documents over recent years to compel platforms for implementing them. A prominent example is the Santa Clara Principles (2021) which is agreed by big platforms such as Facebook, Youtube, Twitter, etc. Such calls for action are not limited with civil society. Regulators also prescribe how platforms are required to report about their content moderation practices around the globe recently from national states such as India and Germany, and from supranational institutions such as the EU (Tewari, 2022; European Union, 2019). Specifically with regard to copyright-related notice-and-takedowns, additional voluntary transparency initiatives such as the Lumen Project allow researchers to gain understanding about the overarching trends from companies such as Google and Twitter (Lumen, 2022).

This sub-study provides a new way of looking into platform transparency reporting through comparison of what is being reported on between platforms and over time. In addition to this meta-perspective, we investigate the reporting of the same set of platforms with regard to actual numbers of copyright-related content moderation, in order to understand both trends and the quality of information that is available through this reporting.

3.3.3 METHODS

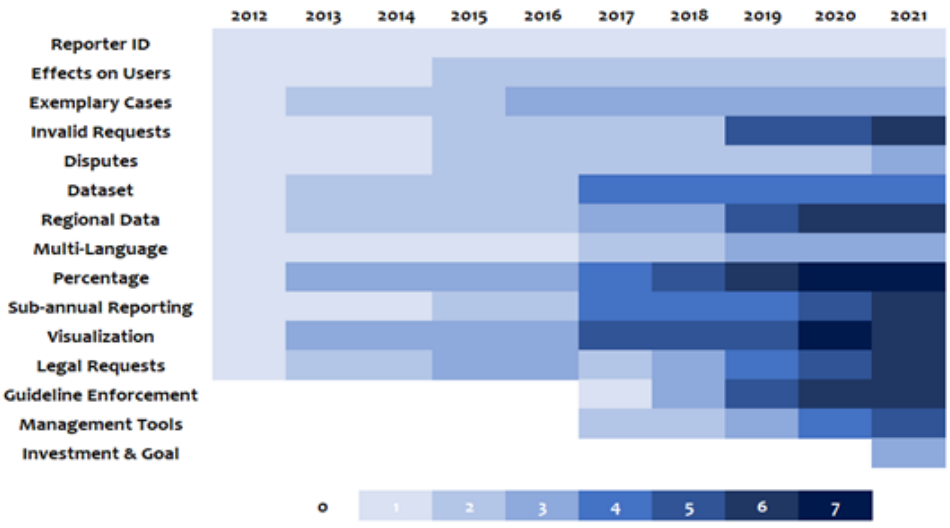
The research design for this sub-study is a longitudinal comparative approach, operationalized with a qualitative content analysis of published transparency reports of major platforms and the information given in their transparency centers. We compare reporting practices and substantial copyright-related decisions of seven of the largest social media platforms: Facebook, Instagram, Pornhub, TikTok, Tumblr, Twitter and YouTube. These were selected based on content production, user traffic and economic impact. From a broader sample including 20 platforms, a set of generic operational definitions of transparency

categories and copyright-related transparency categories were created. The first analysis is to examine whether a certain platform report on a transparency category, in the second we performed a substantive coding in line with the categories set for substantive copyright-based moderation reporting. An important aspect of the study is its focus on longitudinal changes. The chosen platforms' transparency reports were investigated since the year 2012, which is the year the earliest transparency data are available. The data collection extends until 2021. If information in the transparency categories can be accessed through any of the abovementioned sources (reports, transparency centers) even indirectly, we take this as disclosed data. For the substantive comparison of copyright-related content moderation, we have used the same set of sources. Where platforms release numbers bi-annually or quarterly, these numbers were aggregated and annualized.

3.3.4 RESULTS

Transparency Reporting on Copyright Content Moderation by Individual Platforms

Figure 1: Inclusion of reporting criteria in platform transparency reports over time (cumulative)



We start with portraying the inclusion and chronological occurrence of transparency categories in transparency reports of the seven major social media platforms.

Twitter has published detailed transparency data since 2012, including the identity of the top reporters. Twitter consistently sustained the relatively more transparent disclosure of data over the years since the first transparency report. **Tumblr** has published data regarding copyright content moderation since 2015 but it does not produce a dataset nor does it reveal information about the tools and investments for algorithmic detection of copyright. Additionally, Tumblr only revealed data about the total number of disputed claims and valid counter-notices by the uploaders until 2018.

In 2017, **Facebook** and **Instagram** started to disclose data about copyright-based moderation, first ones among the investigated platforms. Since then, they reveal the number of removed content due to copyright infringement. Facebook started to calculate the copyright removal rate differently in 2019, taking the actual number of items which are subject to a copyright report. Facebook and Instagram also have never disclosed any data concerning disputed claims. In 2019, **TikTok** started to publish data about copyright content moderation, albeit in a very limited way. The platform reveals only the number of notices and the removal rate. Thus, not much information is available about TikTok's moderation of copyright content. In 2020, **Pornhub** began to reveal data about copyright removal. For the year 2020, there is only the total number of removed content. For 2021, additionally they revealed the data of the number of received notices. Lastly, **YouTube** published numbers for copyright content moderation for the first time with 2021 a quite comprehensive transparency report. The number of removed content, the number of copyright notices, the removal rate, the number of disputed claims and the results of the disputes were revealed. Besides, Youtube gave information about their tools, investments, and goals for further development of algorithmic systems.

Comparative Analysis of Transparency Reporting on Copyright Content Moderation

With regard to content removal based on conflict with platform content policies (also community guidelines) there is isomorphism across platforms, only Tumblr is not disclosing data on this. The evolution of this common practice is interesting, though. It is actually one of the latest categories that was introduced into transparency reporting.

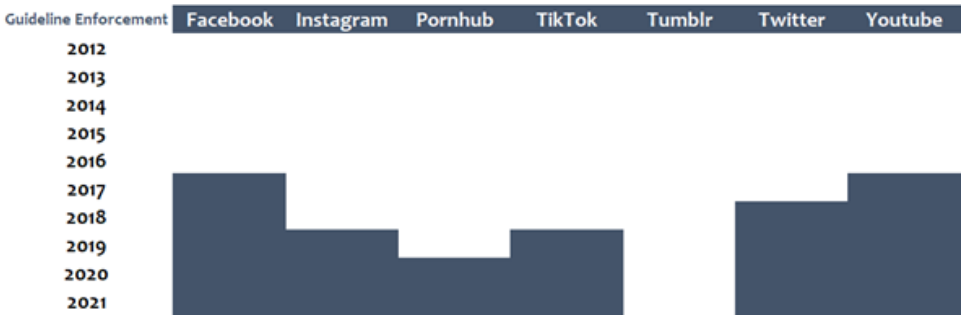


Figure 2: Comparative chronological inclusion of policy-based removal data in platform reporting

When it comes to copyright content moderation, the most essential data relates to the scale of removed content. Twitter leads the other platforms in this category, too, by starting to reveal the data for the number of removed content items already in 2012.

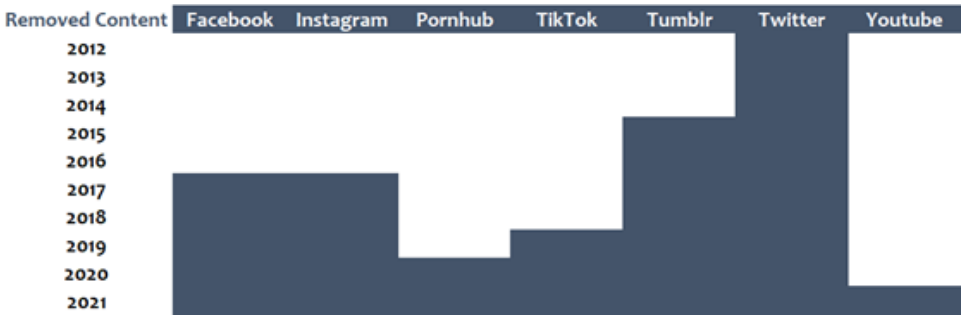


Figure 3: Comparative chronological inclusion of copyright-based removal data in platform reporting

3.3.5 COMPARATIVE ANALYSIS OF COPYRIGHT CONTENT MODERATION NUMBERS

Next, we analyse the data described above across the platforms under study and years for a selected set of categories. Figure 14 shows the *amount of content removed for copyright reasons*. Over the years, it is possible to observe a general increase in the amount of removed content overall on the social media platforms, driven by the growth of the user base and impact of Facebook, Instagram and Twitter, specifically. A similar pattern can also be observed concerning the number of *copyright notices* reported to the platforms. Except for

Tumblr, almost every year platforms receive more such takedown notices compared to the previous year. For Twitter the increase is relatively mild whereas for Instagram, numbers increase significantly. As can be seen in Figure 15, there has been a considerable acceleration in the number of notices received by Facebook, particularly: from 2019 to 2021, the number of notices has almost doubled.

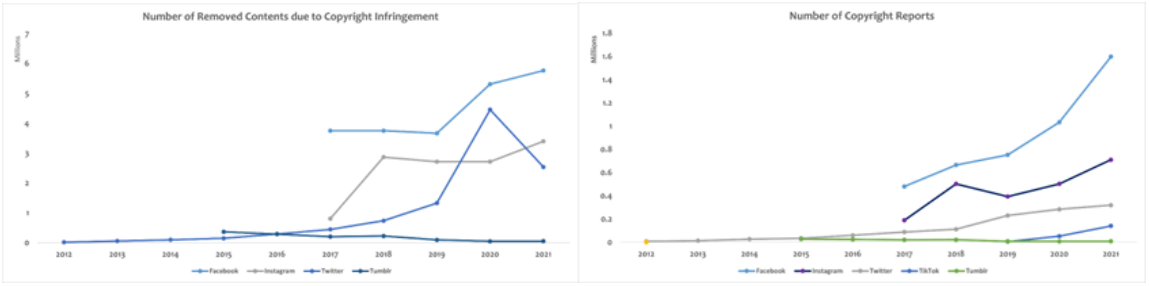


Figure 4: Amount of content removed due to copyright claims per platform & Figure 5: Number of takedown notices received per platform

3.3.6 DISCUSSIONS AND CONCLUSION

Any transparency reporting, and this report subsequently, have a number of important limitations that potentially jeopardize platforms' perceived accountability and positive effects of the reporting on their legitimacy in the eyes of external stakeholders. First of all, "aggregated data in transparency reports only shows the platforms' own assessments, and not the merits of the underlying cases (and) researchers cannot evaluate the accuracy of takedown decisions or spot any trends of inconsistent enforcement" (Keller & Leerssen, 2020, p. 228). Additional limitations of transparency reports in their current form are that they largely focus on the removal of content (and accounts) rather than other (often called "softer") forms of moderation. More recently, practices described as "shadow banning" have taken hold on platforms; users' content is not outright deleted but instead merely not shown to wider audiences, effectively stymying free expression (Savolainen, 2022). Due to the lack of notice of users and their resulting inability to dispute such a moderation measure, shadow banning or the related downranking of content are controversial issues. Even the extent of such "softer" practices is still relatively opaque as "platforms like Instagram, Twitter and

TikTok vehemently deny the existence of the practice" (Savolainen, 2022, p. 1092). Shadow banning is likely less relevant for copyright-based moderation, since there are more categorical issues when intellectual property is being reproduced without permission. In general, the lack of information on how moderation algorithms work is a shortcoming for platform transparency; platforms often engage in "black box gaslighting" to deflect critique (Cotter, 2021). All in all, there is still room for improvement of platform transparency practices, as there is for their moderation practices.

3.4 MEASURING CONTENT BLOCKING AND DELETION ON PLATFORMS, AND ITS IMPACT ON DIVERSITY

This section represents a short summary of the paper "Mandate To Overblock? Understanding The Impact Of EU's Art. 17 On Automated Copyright Content Moderation On Youtube" attached in annex of this report.

In addition to the screening of aggregated data in transparency reports, this second part of the empirical assessment has sought to find evidence about the impact of copyright content moderation on the *content level of social media platforms*. How does copyright content moderation impact on the availability of content and its diversity? While a systematic study on the diversity of content circulating of social media platforms is already challenging, pinning down and isolating the impact of copyright regulation and content moderation is overtly ambitious.

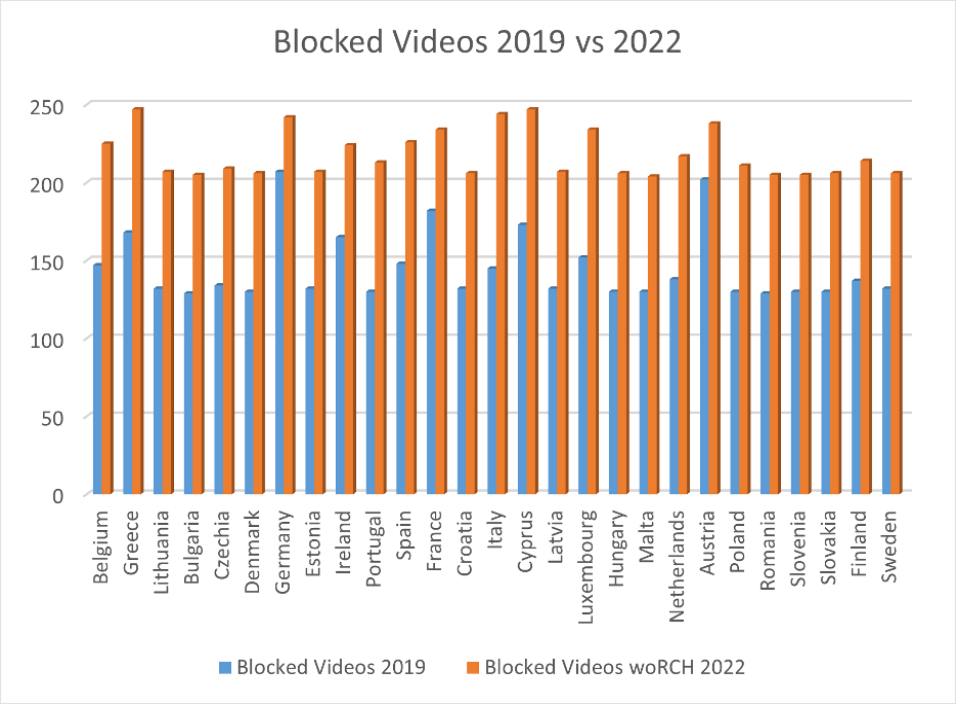
Against this background, this empirical study investigates the changes and influences in access and cultural diversity on social media and streaming platforms, specifically YouTube, in the timeframe 2019 to 2022, exactly focusing on the period between the closure of CDSM Directive negotiations and today, where many national implementations are in effect only shortly or still not existing.

Our results consist of two parts. The first part presents general findings on the copyright takedowns on YouTube in the EU after 2019. The second part measures the diversity of content available on the platform in selected four countries of the EU in 2019 vis-à-vis 2022. For measuring diversity we use the diversity index developed by Stirling and adapted by the UN. Countries were selected depending on specifics of their national copyright regime and their CDSM directive, thus they function as proxies for the impact of copyright regulation. In this data, we investigate if there were any changes in content supply diversity during that time and whether it varies by the countries in the sample.

3.4.1 BLOCKED AND DELETED VIDEOS ON YOUTUBE (2019 – 2022)

Investigating the level of content blocking and deletion on YouTube between 2019 and 2022, we have found that almost 3.8% of videos were blocked or deleted on YouTube in the EU member states between 2019 and 2022 in our sample of 91 000 videos. While lack of data does not allow us to verify exact reasons for content blocking and deletion, we do have excluded items whenever we could identify other reasons. For example, the banning of Russian channels in the context of the invasion to the Ukraine has accounted for substantive numbers of blocked or deleted videos. The resulting 3.8% is a much bigger number than the share for blocking and deletion due to copyright content moderation deletion found in previous studies (1% with Gray & Suzor, 2020).

Example: Videos blocked in the 27 member states of the EU (comparison of 2019 and 2022 samples)

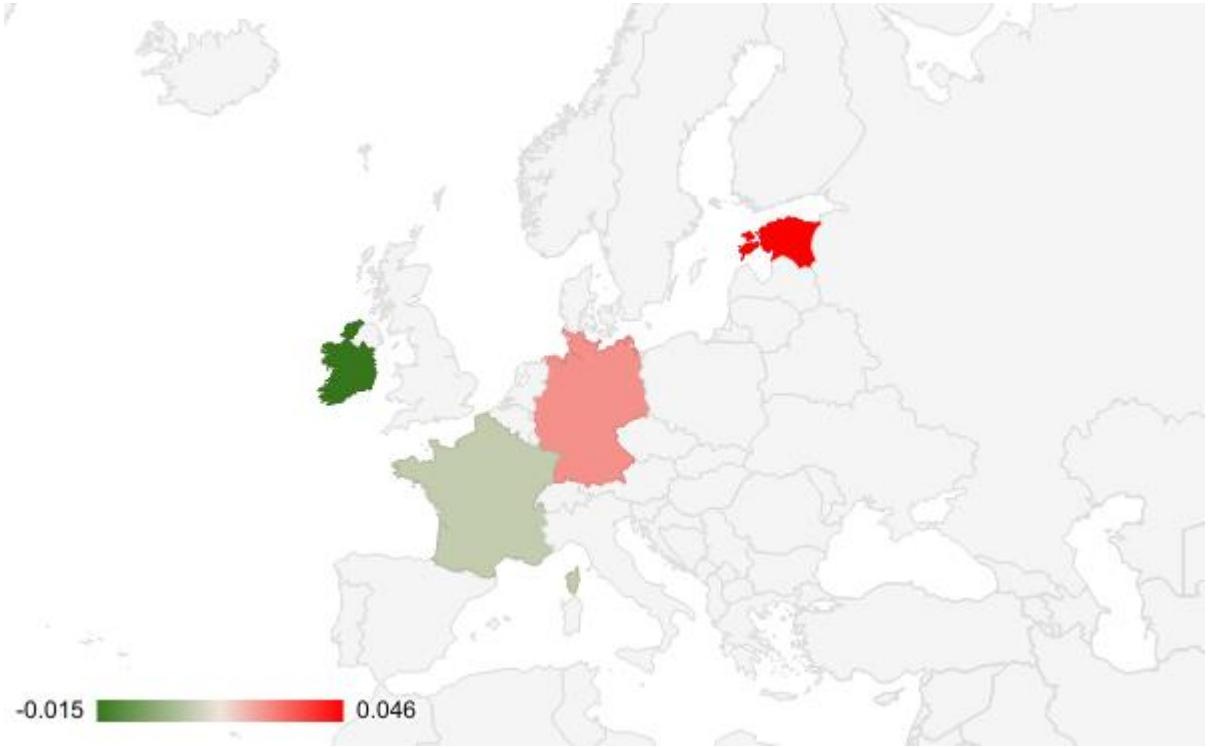


In order to better understand the specifics of blocked or deleted videos vis-à-vis available content, we have run statistical analyses (dominance and regression analysis) to determine key factors. As a result, we found that the main predictor of videos being blocked or deleted was their belonging to certain content categories, namely Film & Animation, People & Blogs, Entertainment, Music & Gaming. In other words, videos from these categories were significantly more likely to be blocked or deleted than videos from other categories. Interestingly, these predictors are consistent with previous research on copyright moderation on YouTube (Gray & Suzor, 2020; Erickson and Kretschmer, 2018). This finding supports the assumption that copyright moderation is indeed the reason for blocking and deletion in most of our cases.

3.4.2 CHANGING DIVERSITY OF YOUTUBE CULTURAL SUPPLY IN FOUR EU COUNTRIES (2019 – 2022)

In the second sub-study we have assessed whether the diversity of cultural goods supply has changed in four countries selected countries of the EU: Estonia, France, Germany, and Ireland. We have sampled these countries in order to represent both large and smaller member states of the European Union; with different traditions of copyright regimes, and different timelines in their national implementation of the CDSM (cf. D.6.2). For measuring diversity, we relied on the Stirling model of diversity (Stirling, 2011), adapted by the UN, namely two out of the three indicators: variety and balance. In our data, variety is represented by the number of categories the YouTube channels of each country belong to, and balance is represented by a number of channels each of the categories has.

Graph 5: differences in diversity 2019 – 2022: Ireland, France, Germany, Estonia



Country	Difference in diversity index
Estonia	0.04597307
Germany	0.02668197
France	0.00824886
Ireland	-0.01518726

The results show that there a decline in diversity regarding the supply of cultural goods in YouTube between 2019 and 2022 in three out of four countries of our sample. In Estonia, the HHI index displayed the greatest change, and saw a decline of 459 points; in Germany, the index decreased by 266 points; in France by 82 points. Only in Ireland we saw a slight increase of diversity in the supply of YouTube channels (by 151 points). These changes do not correlate directly with the implementation of Article 17 of the CDSM Directive and its timing, so we could differentiate specific influence factors. The questions if the CDSM has contributed to the general decrease of diversity on YouTube remains to be discussed and warrants future research.

3.4.3 CONCLUSIONS

Summing up this data-driven investigation of content blocking and content availability on YouTube with a focus on content diversity, we can conclude three things: Firstly, we found a *high share of blocked and deleted content* in our sample. While previous research has identified a share of roughly 1%, our sample identified a share of 3.8%. Due to restricted access to data, though, it is hard to really pin down and isolated the exact reasons for content deletion and take-down. These 3.8% might include other types of content deletion and blocking, although we have applied the measures available to clean the data.

Secondly, we have found a general *decrease of diversity* with regard to available content. Within the four countries under study, three countries display a noticeable decrease in the

diversity index, with Ireland representing a contrary development with a light increase. The country differences do not correlate, though, with national differences in copyright regulation and specifically with the variation in substance and timing of the national implementation of the CDSM. This makes it hard to assess and isolate the actual impact of copyright content moderation and the implementation of the CDSM on content diversity. Is the general decrease of diversity a result of the CDSM? Or rather the product of changing monetization strategies of media companies, shifting media usage routines, or YouTube's algorithmic systems? Some of these research limitations concern the timeline of the study: actual implementation of Article 17 of the CDSM Directive is not yet fully in place in the countries under study, and it is possible that we could not yet see its full-scale influence. Future and continuing research is needed to assess these questions, when the policy implementations become effective and visible at full scale.

But more importantly, thirdly, we have been confronted with the limitations of research in this space due to lack of data access. In the current landscape, it results close to impossible to systematically study the questions posed in this project. What is the impact of copyright regulation and content moderation on content diversity? In fact, this research is not only highly limited, but also dependent on internal decisions of platforms on giving access to (different types of) data. Hence, there is urgent need for more robust rules on data access for researchers. Mandatory data access clauses such as those included in the German NetzDG, the German CDSM implementation as well as in the Digital Services Act pave an important avenue in this regard. Yet, it remains to be seen how robust and effective these clauses are, since they demand highest levels of data security and infrastructure facilities on the side of researchers and their institutions. Finding practical and fair solutions as well as best practices for data access that are not only accessible to researchers at elite and perfectly-equipped institutions is a key challenge for policy and research in the next decade.

3.5 SOCIAL MEDIA CREATORS' PERSPECTIVE ON COPYRIGHT CONTENT MODERATION IN THE EU

This section represents a short summary of the paper "Losing Authenticity: Social Media Creators' Perspective on Copyright Restrictions In The EU" attached in annex of this report.

In the third sub-study, we have taken another angle at understanding copyright content moderation – understanding the experiences of cultural creators who share their work primarily on social media platforms. As social media creators and users in the EU may see a rise in algorithmic copyright moderation after implementation of Article 17 of the CDSMD, we focus this sub-study on creators' understanding and experiences of copyright moderation in relation to their creative work and the labor of media production on social media platforms. To what extent does copyright moderation on the former influence the creations that are posted there? What about the changes to one's creative process? In order to answer these questions, we have interviewed creators with regard to their experiences and descriptions of their interaction with copyright moderation and algorithms. This allows us to better understand the changes and influences that automated copyright moderation brings to creative work.

Cultural creators mainly seeking audiences online are strongly dependent on social media platforms. They have to constantly be involved in pursue of algorithmic visibility as measured by quantified metrics such as likes, views, and shares (Duffy & Meisner, 2022; Bucher, 2017). At the same time, the way platforms curate and govern content and interactions on their sites and its dynamic and intransparent character evokes the threat of 'invisibility' to creators, and this was described before as being 'dangerous' for creators (Cunningham and Craig, 2019).

3.5.1 METHODS

Our study draws upon semi-structured interviews with 14 artists from various EU countries. The sample was drawn from those artists who participated in the survey on digitization and digital access to cultural content, done by (Poort & Pervaist, 2022) in the context of the ReCreating Europe Project. The artists interviewed used a wide range of social media platforms: Instagram, Facebook, TikTok, YouTube, Behance, Etsy, LinkedIn, Vimeo, Pinterest and Dailymotion.

The research draws on a multimodal framework to analyze the copyright governance of creative practices and products, focusing on regulative dimension (in our case – adaptation of the Article 17(4) CDSMD)), normative dimension (prevalent assumptions about legitimate and illegitimate behavior in a specific community) and the influence of technological affordances relevant to creative work. (Katzenbach, 2018).

For many interviewees, anticipation of platform punishments directly influenced the cultural products that they produced. Indeed, most of them used self-censorship, avoidance and concerted efforts to circumvent algorithmic intervention in their creative work before posting it on social media platforms. Our research has also confirmed previous findings on user folk theories regarding algorithms (De Vito et al., 2017) and algorithmic gossip, searching for the shared meaning of algorithmic moderation practices (Natale, 2019; Bishop, 2020). We have found evidence of exploitation, instability, and overwork culture (Duffy & Meisner, 2022) among those artists whose work and income relates to social media platforms but not depends on it fully.

Interviews followed a semi-structured interview protocol; and lasted between 30 and 90 minutes. Participants received a gift card ($50) in exchange for their time and insight, and interviews were conducted on Zoom. All the interviews were recorded after acquiring consent for this. After the completion of the interviews the audio was transcribed and edited for any discrepancies. From the transcripts, the study's authors developed the coding categories and applied focused codes to the dataset.

3.5.2 RESULTS

A basic finding is that creative content producers have only scarce understanding of the regulatory mechanisms functioning on social media platforms. Assumptions on 'right' and 'wrong' practices differ, and usually do not correspond to legal realities, thus questioning the actual impact of the regulative dimension of the provision and enforcement of formal rules on everyday cultural practices. Consistent with studies on understanding algorithms (Bishop, 2019), creatives on social media platforms often gain knowledge on issues of copyright through gossip. "I don't know much actually. What I know, it's from personal stories never an actual study about it or an actual thing explaining what is what and how we can do stuff."

Creators, even if they themselves have not experienced moderation yet, are anticipating copyright moderation, and thus adjust the content beforehand. Some interviewees confessed to quitting posting certain products to platforms all together, since they were not sure about copyright moderation. Nine out of fourteen interview participants either had their creations taken down for copyright infringement or knew someone who had. When asked directly, most of the participants did not think that moderation due to copyright issues increased during the last year. However, when they remembered instances of such moderation, often their cases were from not-so-distant past, such as "Last Christmas" (Lip Sync video, creator from Bulgaria), "last year" (reel, creator from Croatia), "during the past three or four years" (video, creator from Romania). Another detail is that some interviewees had their old videos or posts taken down in the recent two years, although they had online for a long time before. Our participants in general have found the appeal, report and complain processes on platforms not very helpful. Sometimes they had to use networks of friends and followers in order to solve the issue, so not directly complaining to a platform but using other mechanisms.

The European Copyright in the Digital Single Market (CDSM) Directive (2019/790) has been adapted and came into force in June 2019. However, some creators do not think that the legislation works. Even those from the countries which have adapted the Directive, are either not aware or have not seen the laws put into action when it comes to protection of their own work.

3.5.3 CONCLUSION

The main takeaway from our study is that users of social media platforms which do creative work on those, are influenced by algorithmic content moderation. Perhaps our most important finding which extends understanding on how algorithmic content moderation influence creative work on platforms, is that creators engage in self-censorship, avoiding posting certain content or adjusting it in advance. For many artists, anticipation of platform punishments directly influenced the cultural products that they produced. In addition, because the regulative dimension of algorithmic copyright moderation is opaque for creators, they engage in algorithmic gossip (Bishop, 2019) and use user folk theories (De Vito et. Al, 2019) trying to guess which practices are accepted and which are not. These are important policy implications from this research, such as that more transparency in platform governance is needed, both from policy makers and platforms themselves, so that the automated content moderation does not add to the uncertainty and insecurity of the creators' media production work on social media platforms.

3.6 CONCLUSIONS

These empirical assessments on the impact of copyright regulation and content moderation on diversity have generated key findings. At the same time, they have shown the clear limitations of existing research options given the notorious scarce access to data held by platforms. With regard to substantive findings, the analysis of transparency report has identified an almost uniform trend towards reporting certain aggregated data. There is obviously strong co-orientation between platforms with a clear upward trend to disclose more data. With a view to substantial data on copyright content moderation, we see between 2012 and 2021 a strong increase of notice, takedown, and conflicting cases, but mostly reflecting the general growth of the platforms. With regards to removal rates, the picture is much more complex, not yielding a specific trend.

The analysis of content level data at YouTube has found a high share of blocked and deleted content, with 3.8% higher than reported in earlier studies. Lack of rich and robust data,

though, makes it hard to draw conclusions on the impact of copyright regulation and content moderation. Is that number higher than in previous studies from 2016-19 because in 2022 CDSMD has already been in, so that we actually see more takedowns across the board, or is this an artefact of different data collection and analysis methods? Even after best-efforts have been done, this is hard to disentangle. Similarly with regard to diversity of available content: We have found a general decrease of diversity with regard to available content. Yet, the country differences do not correlate with national differences in copyright regulation, and specifically not with the variation in substance and timing of the national implementation of the CDSM. So, while the general decrease of diversity might be interpreted as a confirmation of major concerns that critics of Art 17 had raised in the political process, the country-specific result do not support this interpretation.

In consequence, the empirical assessment has produced findings that indicate a strong impact of copyright regulation and content moderation on diversity, and potentially even an impact that leads to a decrease in diversity of content. Yet, the research has also shown that these interpretations cannot be fully verified based on the data that is available. Researchers are strongly dependent on options that platforms themselves provide. Our own research was massively constrained, for example, by YouTube's changed API provision that did not allow us to reproduce the study by Gray and Suzor (2020) and thus hindering us to come to clear conclusions on reasons for content deletion and removal.

As a consequence, robust and fair access for research to platforms' moderation data is a key challenge and regulatory issue for the near future. Several regulatory measures, including Art. 35 of the DSA, Art. 40 of the GDPR, national implementations of the CDSMD and other national initiatives, are already establishing mandatory access rights. Yet even there, implementation questions and practical challenges remain a heavy burden to researchers and research institutions. At EU level, EU institutions and in particular the Commission should prioritize issuing clear guidance and provide adequate resources to enable researchers, including from institutions with few resources, to pursue the research needed for shaping platforms and their regulation in democratic societies.

4. JOINT CONCLUDING REMARKS

The *assessment of the existing legal frameworks* that shape the role of online platforms in organising the circulation of culture and creative works in Europe through content moderation has shown the complex landscape of interacting rules that meet and address the reality of content moderation at scale. It suggests that rather than envisioning a private-regulatory copy of a "full trial" setup, a different conceptual approach of "rough justice" is necessary to catch the developments and in this context suggests recommendations for improvements regarding procedural rules, substantive rules and competences. Furthermore, it shows that with regards to access to culture and cultural diversity, decision quality should be emphasised as a separate factor from ex post mitigation mechanisms. Both the Digital Services Act and the CDSM Directive (including case law) provide starting points for this. The analysis also points to the fact that content moderation increasingly requires an understanding of contextual use but further work is needed on the potential risk of "bias carry-over" from datasets to content moderation. In this context, it is also worthwhile to point out that content moderation technology appears to be only stepmotherly treated in currently ongoing negotiations of the AI Act.

The *empirical assessment of the impact of copyright regulation and content moderation on diversity* has produced a set of key findings, but also highlighted critical loopholes and limitations for research on platforms in general, and their role in cultural diversity in particular. Overall, the results indicate a strong impact of copyright regulation and content moderation on diversity, and potentially an impact that leads to a decrease in diversity of content. Yet, the research has also shown that these interpretations cannot be fully verified based on the limited data that is available to researchers and the public. While platforms increasingly publish transparency reports with growing sets of aggregated data about content moderation and other relevant aspects, researchers increasingly are confronted with massive challenges to get access to data that is needed to pursue their research.

This report has also highlighted the need for further research on issues of diversity and access on social media platforms, given its high relevance for European societies, and at the same

time its complex nature, specifically in the context of contemporary fragmented media landscapes. Diversity is a key theme and important goal in European public and policy debates. In consequence, both understanding as well as addressing issues of diversity and access to culture in the context of digitalization and platformization must come much more to the forefront of policy and research agendas. Evaluating diversity is already conceptually a major challenge in today's fragmented media landscape. But strong limitations in access to platform data sharply constrains systematic research into diversity issues in the context of platforms and social media.

For that reason, we conclude this report with a strong and clear call for robust mandatory data access clauses in all future platform regulations. The existing initiatives, including Art. 35 of the DSA, Art. 40 of the GDPR, national implementations of the CDSMD, pave the way here and do lay important and necessary foundations, yet implementation questions and practical challenges remain open. Thus, it is of high importance to clearly flag this as a key are for future policy-making. If the general clauses do not receive robust implementation, there is high risk that requirements to make of these clauses will constitute heavy burdens to researchers and research institutions, in effect potentially limiting this privilege only to researchers and institutions with strong resources. At EU level, EU institutions and in particular the Commission should prioritize issuing clear guidance and provide adequate resources to enable researchers, including from institutions with few resources, to pursue the research needed for shaping platforms and their regulation in democratic societies.

5. BIBLIOGRAPHY

Appelman, N., Quintais, J.P. & Fahy, R. (2021). Using Terms and Conditions to apply Fundamental Rights to Content Moderation. *Verfassungsblog* (1.9.2021). https://verfassungsblog.de/power-dsa-dma-06/

Bärtl, M. (2018). YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence, 24(1), 16–32. https://doi.org/10.1177/1354856517736979

Bar-Ziv, S., & Elkin-Koren, N. (2018). Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown (SSRN Scholarly Paper No. 3214214). https://papers.ssrn.com/abstract=3214214

Benhamou, F., & Peltier, S. (2007). How should cultural diversity be measured? An application using the French publishing industry. Journal of Cultural Economics, 31(2), 85–107. https://doi.org/10.1007/s10824-007-9037-8

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. New Media & Society, 21(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Bishop, S. (2020). Algorithmic Experts: Selling Algorithmic Lore on YouTube. Social Media + Society, 6(1), 2056305119897323. https://doi.org/10.1177/2056305119897323

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. New Media & Society, 14(7), 1164–1180. https://doi.org/10.1177/1461444812440159

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. Information, Communication & Society, 20(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Burgess, J. & Green, J. (2018). YouTube: Online Video and Participatory Culture, 2nd Edition. Polity Press. ISBN: 978-0-745-66019-6

Burgess, J.& Green J. (2009). YouTube: Online Video and Participatory Culture, 1st Edition. Cambridge, UK: Polity Press, ISBN 978-0-7456-4478-3, 140 pages

Caplan, R., & Boyd, D. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. Big Data & Society, 5(1), 2053951718757253. https://doi.org/10.1177/2053951718757253

Cotter, K. (2021). "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society, 1-18.*

Cunningham, S., & Craig, D. (2019). Creator Governance in Social Media Entertainment. Social Media + Society, 5(4), 2056305119883428. https://doi.org/10.1177/2056305119883428

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017). "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 3163–3174. https://doi.org/10.1145/3025453.3025659

Duffy BE, Pinch A, Sannon S, et al. (2021) The nested precarities of creative labor on social media. Social Media + Society 7(2): 1–12.

Duffy, B. E., & Meisner, C. (2022). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. Media, Culture & Society, 01634437221111923. https://doi.org/10.1177/01634437221111923

Edwards, D. W. (2018). Circulation Gatekeepers: Unbundling the Platform Politics of YouTube's Content ID. Computers and Composition, 47, 61–74. https://doi.org/10.1016/j.compcom.2017.12.001

Erickson, K., & Kretschmer, M. (2018). "This Video is Unavailable": Analyzing Copyright Takedown of User-Generated Content on YouTube (SSRN Scholarly Paper No. 3144329). https://papers.ssrn.com/abstract=3144329

Eslami, M., Rickman, A., Vaccaro, K. et al (2015). I always assumed that I wasn't really that close to [her] : Reasoning about Invisible Algorithms in News Feeds. Conference: 33rd Annual ACM Conference on Human Factors in Computing Systems. 10.1145/2702123.2702556.

European Union (2019). *Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services* (Text with EEA relevance). Retrieved on October 28, 2022 from https://eur-lex.europa.eu/eli/reg/2019/1150/oj

Farchy, J., & Ranaivoson, H. (2011). Measuring the Diversity of Cultural Expressions: Applying the Stirling Model of Diversity in Culture. Undefined. https://www.semanticscholar.org/paper/Measuring-the-Diversity-of-Cultural-Expressions%3A-of-Farchy-Ranaivoson/7514314a6d462408b4be77a44621123c79b9d089

Gillespie T and Seaver N (2015) Critical algorithm studies: A reading list. Available at: https://socialmediacollective.org/reading-lists/critical-algorithm-studies/ (accessed 10.10.2022)

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gorwa, R. (2021). Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG. Telecommunications Policy, 45(6), 102145. https://doi.org/10.1016/j.telpol.2021.102145

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance: *Big Data & Society, 7*(1), 1–15. https://doi.org/10.1177/2053951719897945

Gray, J. E., & Suzor, N. P. (2020). Playing with machines: Using machine learning to understand automated copyright enforcement at scale. Big Data & Society, 7(1), 2053951720919963. https://doi.org/10.1177/2053951720919963

Grimmelmann, J. (2005). Regulation by Software. *Yale Law Journal*.

Hallinan, B., & Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. New Media & Society, 18(1), 117–137. https://doi.org/10.1177/1461444814538646

Haupt, J. (2021). Facebook futures: Mark Zuckerberg's discursive construction of a better world. *New Media & Society, 23*(2), 237–257. https://doi.org/10.1177/1461444820929315

Hazelhorst, M. (2017). *The Right to a Fair Trial in Civil Cases.* Springer – Asser Press.

Hellman, H. (2001). Diversity - An End in Itself?: Developing a Multi-Measure Methodology of Television Programme Variety Studies. European Journal of Communication, 16(2), 181–208. https://doi.org/10.1177/0267323101016002003

Jürgens, P., & Stark, B. (2022). Mapping Exposure Diversity: The Divergent Effects of Algorithmic Curation on News Consumption. Journal of Communication, 72(3), 322–344. https://doi.org/10.1093/joc/jqac009

Katzenbach, C. (2018). There Is Always More Than Law! From Low IP Regimes To A Governance Perspective In Copyright Research. Journal of Technology Law and Policy, 22(1).

Katzenbach, C. (2021). "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211046182

Katzenbach, C. (2021). "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society, 8*(2). https://doi.org/10.1177/20539517211046182

Katzenbach, C., Kopps, A., Magalhaes, J.C., Redeker, D., Sühr, T., & Wunderlich, L. (2022). The Platform Governance Archive. A longitudinal dataset to study the governance of communication and interactions by platforms. Unpublished Manuscript. Alexander von Humboldt Institute for Internet and Society.

Keller, D., & Leerssen, P. (2020). Facts and where to find them: empirical research on internet platforms and content moderation. *Social Media and Democracy: The State of the Field and Prospects for Reform, 220*, 224.

Keller, P. (2019). Article 17 stakeholder dialogue (day 3): Filters do not meet the requirements of the directive. Communia Association (3.12.2019). https://communia-association.org/2019/12/03/article-17-stakeholder-dialogue-day-3-filters-not-meet-requirements-directive/

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131*, 1598.

Lévy-Hartmann, F. (2011). Une mesure de la diversité des marchés du film en salles et en vidéogrammes en France et en Europe. Culture Méthodes, 1, 1. https://doi.org/10.3917/culm.111.0001

Linzer, P. (2001). Rough Justice: A Theory of Restitution and Reliance, Contracts and Torts, *Wis. L. Rev*. 695-775.

Margoni, T., Quintais, J.P. & Schwemer, S.F. (2022). Algorithmic propagation: do property rights in data increase bias in content moderation? Part II. *Kluwer Copyright Blog* (9.6.2022). http://copyrightblog.kluweriplaw.com/2022/06/09/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-ii/

McDonald, D. G., & Lin, S.-F. (2004). The Effect of New Networks on U.S. Television Diversity. Journal of Media Economics, 17(2), 105–121. https://doi.org/10.1207/s15327736me1702_3

Mezei, P., Jahromi, H.K., Priora, G. (2021). What's the buzz? Tell me what's a-happening (around Article 17)? Tales from Hungary, Germany, Italy and Sweden. Kluwer Copyright Blog

(29.11.2021). https://copyrightblog.kluweriplaw.com/2021/11/29/whats-the-buzz-tell-me-whats-a-happening-around-article-17-tales-from-hungary-germany-italy-and-sweden/

Moreau, F., & Peltier, S. (2004). Cultural Diversity in the Movie Industry: A Cross-National Study. Journal of Media Economics, 17(2), 123–143. https://doi.org/10.1207/s15327736me1702_4

Natale, S. (2019). If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. New Media & Society, 21(3), 712–728. https://doi.org/10.1177/1461444818804980

Paolillo, J. C. (2008). Structure and Network in the YouTube Core. Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008). https://doi.org/10.1109/HICSS.2008.415

Perel, M., & Elkin-Koren, N. (2018). Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. Florida Law Review, 69(1), 181.

Peukert A. et al. (2022). European Copyright Society – Comment on Copyright and the Digital Services Act Proposal. *IIC - International Review of Intellectual Property and Competition Law*, 358-376.

Poell T., Nieborg D., & Duffy B. (2021). Platforms and Cultural Production. Polity Press, 260 p. ISBN: 978-1-509-54052-5

Poort, J., & Pervaiz, A. (2022). Report(s) on the perspectives of authors and performers. Zenodo. https://doi.org/10.5281/zenodo.6779373

Quintais, J.P. (2022). Between Filters and Fundamental Rights: How the Court of Justice saved Article 17 in C-401/19 - Poland v. Parliament and Council. *VerfBlog* (16.5.2022). https://verfassungsblog.de/filters-poland/

Quintais, J., Mezei, P., Harkai, I., Magalhães, J., Katzenbach, C., Schwemer, S., & Riis, T. (2022). Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis. Zenodo. https://doi.org/10.5281/zenodo.7081626

Quintais, J. & Schwemer, S.F. (2022). The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?. *European Journal of Risk Regulation, 13*(2), 191-217. doi:10.1017/err.2022.1

Ranaivoson, H. (2010). The Determinants of the Diversity of Cultural Expressions—An International Quantitative Analysis of Diversity of Production in the Recording Industry.

Redeker, D., & Martens, K. (2018). *NGOs and accountability*. In: Aynsley Kellow, Hannah Murphy-Gregory (Eds.). Handbook of Research on NGOs, 303-324. Cheltenham: Edgar Elgar.

Rendas, T. (2021). *Exceptions in EU Copyright Law: In Search of a Balance Between Flexibility and Legal Certainty*. Kluwer Law International.

Rieder, B., Coromina, Ò., & Matamoros-Fernández, A. (2020). Mapping YouTube: A quantitative exploration of a platformed media system. First Monday. https://doi.org/10.5210/fm.v25i8.10667

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Rosati, E. (2019). *Copyright and the Court of Justice of the European Union*. Oxford University Press.

Rosati, E. (2021). The Digital Services Act and Copyright Enforcement: The Case of Article 17 of the DSM Directive in European Audiovisual Observatory *Unravelling the Digital Services Act Package*.

Santa Clara Principles. (2021). T*he Santa Clara Principles: On Transparency and Accountability in Content Moderation*. Retrieved October 28, 2022, from https://santaclaraprinciples.org/

Savolainen, L. (2022). The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society, 44*(6), 1091–1109. https://doi.org/10.1177/01634437221077174

Schwemer, S.F. (2022). Digital Services Act: A Reform of the e-Commerce Directive and Much More" prepared for A Savin, *Research Handbook on EU Internet Law (Edward Elgar, 2023)*, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213014

Schwemer, S.F. (2022). Recommender Systems in the EU: from Responsibility to Regulation. *Morals & Machines* (2022), 60–69. https://doi.org/10.5771/2747-5174-2021-2-60

Stirling, A. (2011). Measuring the Diversity of Cultural Expressions: Applying the Stirling Model of Diversity in Culture. 10.13140/RG.2.1.3228.2082.

Suzor, N., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette, 80*(4), 385–400. https://doi.org/10.1177/1748048518757142

Suzor, N., Myers West, S., Quodling, A. & York, J. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 1526–1543.

Tewari, S. (2022). *Transparency initiatives in the DSA: An Exciting Step Forward in Transparency Reporting*. Lumen Project. Retrieved on October 28, 2022 from: https://www.lumendatabase.org/blog_entries/transparency-initiatives-in-the-dsa-an-exciting-step-forward-in-transparency-reporting

Tushnet, R. (2014). All of This Has Happened before and All of This Will Happen Again: Innovation in Copyright Licensing. Georgetown Law Faculty Publications and Other Works. https://scholarship.law.georgetown.edu/facpub/1459

Urban, J. M., Karaganis, J., & Schofield, B. (2017). Notice and Takedown in Everyday Practice (SSRN Scholarly Paper No. 2755628). https://doi.org/10.2139/ssrn.2755628

van Dijck, J. (2020). Governing digital societies: Private platforms, public values. Computer Law & Security Review, 36, 105377. https://doi.org/10.1016/j.clsr.2019.105377

Ward, A. (2022). Article 47 – Right to and Effective Remedy and to a Fair Trial', in Steve Peers, Tamara Hervey, Jeff Kenner and Angela Ward (Eds.), *The EU Charter of Fundamental Rights: A Commentary*. 2nd ed. Bloomsbury Publishing.

Willson, M. (2017). Algorithms (and the) everyday. Information, Communication & Society, 20(1), 137–150. https://doi.org/10.1080/1369118X.2016.1200645

Wojcieszak, M., Menchen-Trevino, E., Goncalves, J. F. F., & Weeks, B. (2022). Avenues to News and Diverse News Exposure Online: Comparing Direct Navigation, Social Media, News Aggregators, Search Queries, and Article Hyperlinks. The International Journal of Press/Politics, 27(4), 860–886. https://doi.org/10.1177/19401612211009160

APPENDIX

PAPER 1: A THEORY OF ROUGH JUSTICE FOR INTERNET INTERMEDIARIES FROM THE PERSPECTIVE OF EU COPYRIGHT LAW

PAPER 2: QUALITY OF AUTOMATED CONTENT MODERATION: REGULATORY ROUTES FOR MITIGATING ERROR

PAPER 3: ALGORITHMIC PROPAGATION: DO PROPERTY RIGHTS IN DATA INCREASE BIAS IN CONTENT MODERATION?

PAPER 4: FINALLY OPENING UP? THE EVOLUTION OF TRANSPARENCY REPORTING PRACTICES OF SOCIAL MEDIA PLATFORMS

PAPER 5: MANDATE TO OVERBLOCK? UNDERSTANDING THE IMPACT OF EU'S ART. 17 ON AUTMATED CONTENT MODERATION ON YOUTUBE

PAPER 6: LOSING AUTHENTICITY: SOCIAL MEDIA CREATORS' PERSPECTIVE ON COPPYRIGHT RESTRICTIONS IN THE EU

The ReCreating Europe project aims at bringing a ground-breaking contribution to the understanding and management of copyright in the DSM, and at advancing the discussion on how IPRs can be best regulated to facilitate access to, consumption of and generation of cultural and creative products. The focus of such an exercise is on, inter alia, users' access to culture, barriers to accessibility, lending practices, content filtering performed by intermediaries, old and new business models in creative industries of different sizes, sectors and locations, experiences, perceptions and income developments of creators and performers, who are the beating heart of the EU cultural and copyright industries, and the emerging role of artificial intelligence (AI) in the creative process.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

A theory of rough justice for internet intermediaries from the perspective of EU copyright law[1]

1. Introduction

Internet intermediaries and in particular online platforms constitute an important gateway for accessing internet content including copyright-protected works. Internet content is distributed via internet intermediaries and such intermediaries engage in content moderation. Content moderation can be necessary for the intermediary to escape liability for copyright infringements that take place on the intermediary's infrastructure and it is also an essential means to ensure compliance with the intermediary's terms of service. In the proposed Digital Services Act 'Content moderation' is defined as

> *'the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account'.*[2]

As a means of rights-enforcement, content moderation raises two major problems. The first relates to the *accuracy* of the moderation practices. As the point of departure, the optimal content moderation scheme can perfectly identify illegal and incompatible content and moderate it accordingly which means that it neither under-enforce nor over-enforce substantive law (illegal content)[3] or the intermediary's terms of service (incompatible content). The second problem concerns the inherent *privatization of justice* that results when enforcement of rights is left to a private party who in addition can distort the balancing of interests in substantive law by stipulating what otherwise legal content that are deemed incompatible with the intermediary's terms and conditions and thus is not available to the users.

In the following, the problems will be examined within the field of copyright law. The balancing of interest in copyright law is reflected in Art. 27 of the UN Universal Declaration of Human Rights (UDHR). Art. 27 states that

> *1. Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits'.*[4]
> *2. Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.'*[5]

---

[1] Professor Thomas Riis, PhD, LLD, Centre for Information and Innovation Law, University of Copenhagen.
Acknowledgements: This research is conducted under the reCreating Europe project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870626.
[2] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), article 3(t), O.J. 2022, L 277/1.
[3] Cf. the definition in Commission Recommendations of Mar. 1, 2018, on Measures to Effectively Tackle Illegal Content Online, C(2018) 1177 final, 1, 1-2 (Mar. 1, 2018), Chapter I(4)(b).
[4] Cf. Art. 15, the International Covenant on Economic, Social and Cultural Rights.
[5] Cf. Art. 17(2) CFR: 'Intellectual property shall be protected'.

In simple terms, from a legal technical perspective, copyright law predominantly excludes access to protected works for the purpose of appropriating economic value from those works. Since many expressions of cultural phenomena and artifacts are protected by copyright, there is an inherent conflict of interests between copyright and access to culture. Copyright law's rationale is to exclude access to protected works but only to the point necessary to create economic incentives for the creation of new works. Beyond that point access to works ought to be free. Ideally, content moderation should ensure that copyright infringing works are inaccessible contrary to non-infringing works. However, intermediaries may institute further restrictions. For instance, according to Facebook's terms and conditions, photographs depicting most forms of nudity are moderated despite the photographs being legal. The additional restrictions on legal content creates a field of tension in relation to fundamental rights, namely the freedom of expression and information,[6] freedom of arts[7] and the freedom to conduct a business.[8]

2. Rights-enforcement and other elements of justice

In the field of copyright law, content moderation has been closely related to enforcement of rights. The ease of copying and distributing copyrighted content by every internet user has made it very difficult for rights holders to identify, litigate and remedy copyright infringements. Digital enforcement by online intermediaries addresses this difficulty by providing an opportunity of a swift removal of infringing content.

Over-enforcement of copyright has been reported in a number of times in certain types of cases. It can be difficult for the intermediary, particularly when using a machine learning system, to establish whether the use of a protected work is allowed under one of copyright law's exceptions and limitations, e.g. the exception on caricature, parody or pastiche.[9] Furthermore, in a survey on empirical studies *Bar-Ziv* and *Elkin-Koren* found large-scale use of notice and take down procedures to remove non-copyrighted materials.[10] Over-enforcement is related to arguably the most widespread concern about intermediaries' content moderation which is the lack of procedural safeguards and lack of transparency.[11]

---

[6] Art. 11 CFR.
[7] Art. 13 CFR.
[8] Art. 16 CFR.
[9] Cf. COM(2021)288, "Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market", para. VI, where the Commission states that cases where the user has significantly modified the work in a creative manner, for example by adding elements to a picture to create a 'meme', would generally not be manifestly infringing because the result may be covered by the parody exception, and since the use is not manifestly infringing it ought to be subject to human review. See also e.g. Guzel, "Directive on Copyright in the Digital Single Market and Freedom of Expression: The EU's Online Dilemma" in Synodinou, Jougleux, Markou & Prastitou Merdi (Eds.), *EU Internet Law in the Digital Single Market* (Springer International Publishing, 2021); and Erickson and Kretschmer, "This Video is Unavailable": analyzing copyright takedown of user-generated content on YouTube analyzing copyright takedown of user-generated content on YouTube", (2018) *JIPITEC*, 2190–3387.
[10] Bar-Ziv & Elkin-Koren, "Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown", (2018), *Connecticut Law Review*, 379.
[11] Ibid. 379; Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation", (2020) *Fordham International Law Journal*, 956 ff. with references; Black, Decentring regulation: Understanding the role of regulation and self-regulation in a 'post-regulatory' world", (2001) *Current Legal Problems,* 103; Perel & Elkin-Koren "Accountability in algorithmic copyright enforcement", (2016) Stanford Technology Law Review, 473–532; Keats Citron "Technological due process", (2008) *Washington University Law Review*, 1249; Elkin-Koren, "After twenty years: Revisiting copyright

The amount of content distributed via internet intermediaries is growing exponentially and it is widely held that it has become impossible to rely on human review for content moderation in all cases. As a consequence, intermediaries have developed automated content filters and high-speed removal systems to enforce rights efficiently.[12] It goes without saying that under such circumstances accuracy in the rights-enforcement as known from judgments in civil proceedings made possible by extensive production of evidence is impossible to preserve.[13] The same applies to the procedural safeguards of procedural law (formal justice). Accordingly, in the context of content moderation of intermediaries, the ordinary conception of justice in rights-enforcement must be substituted by a new and viable conception of 'rough justice'. In the following I will sketch out a model of rough justice in intermediaries' content moderation based on a human rights approach.

All rights-enforcement systems shall take three general objectives into consideration:

The first objective is efficacy. According to *e.g*. Art. 47(1) of the EU Charter on Fundamental Rights (CFR)[14] which is part of Title VI on 'Justice', everyone whose rights and freedoms are violated has the right to an effective remedy before a tribunal.[15] Efficacy implies firstly that there is access to justice in the sense that mechanisms for rights-enforcement shall be easily available and not overly costly. Secondly, effective remedies shall be available to redress wrongs.

The second objective is fair trial. Fair trial is a more complex concept than efficacy and it includes consistency and predictability in rights-enforcement, proportionality and a certain degree of symmetry between opposing parties (equality of arms).[16]

The third objective is a balanced use of resources in rights-enforcement. Law enforcement requires resources and the 'price' of enforcement may be too high. The scholarly literature on law enforcement and resources primarily focuses on criminal law and police expenditures and demonstrates that there is an upper limit to the amount of resources that can be used for enforcing criminal law even in cases

---

liability of online intermediaries" in: Frankel and Gervais (Eds), *The Evolution and Equilibrium of Copyright in the Digital Age*, (Cambridge University Press, 2014), pp. 29–51; Suzor, Myers West, Quodling and York, "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation', (2019) *International Journal of Communication*, 1137, Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, 15 (Apr. 6, 2018), paras. 71-72; Douek, "Facebook's oversight board: move fast with stable infrastructure and humility", (2019) *North Carolina Journal of Law & Technology*, 5.

[12] Elkin-Koren, "Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence", (2020) *Big Data & Society July–December*, 1–13, Ofcom, *The use of AI in content moderation. Report produced by Cambridge Consultants on behalf of Ofcom* (2019). Available at: <www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf> (last visited 27 Janyary 2023), Guzel, "Directive on Copyright in the Digital Single Market and Freedom of Expression: The EU's Online Dilemma" in Synodinou, Jougleux, Markou & Prastitou Merdi (Eds.), *EU Internet Law in the Digital Single Market* (Springer International Publishing, 2021) and Douek, "Facebook's oversight board: move fast with stable infrastructure and humility", (2019) *North Carolina Journal of Law & Technology*, 5-6 and 10-12.

[13] E.g. Land, "Against Privatized Censorship: Proposals for Responsible Delegation," (Winter 2020) *Virginia Journal of International Law*, 428-429.

[14] O.J. 2012, C 326/02.

[15] Cf. Art. 8 UDHR.

[16] Art. 6(1), European Convention on Human Rights; Arts. 47(2)-(3) and 48 CFR; and Arts. 7 and 10 UDHR.

concerning serious crimes.[17] Even though the circumstances are somewhat different in civil law cases, the same concern is relevant.

Compared to enforcement by civil procedure, it is obvious that one of the critical issues in relation to content moderation by intermediaries is to balance the procedural safeguards to ensure a fair trial with the required resources.[18] The same balance is relevant in civil procedures, however, different weights turn the scale. Further, a certain amount of roughness characterizes formal civil justice as well.[19]

3. The human rights approach to justice

The human rights concept of fair trial could be a reference point for the development of a model of rough justice. In the following sections, the essential elements of the human rights concept of fair trial will be examined.

It follows from Art. 6(2) of The Treaty on European Union (TEU) that the Union shall accede to the European Convention for the Protection of Human Rights (ECHR) and Fundamental Freedoms. In addition, Art. 6(3) TEU provides that fundamental rights, as guaranteed by the ECHR and as they result from the constitutional traditions common to the Member States, shall constitute general principles of the Union's law. The significance of the ECHR on the protection of human rights is reiterated in Art. 52(3) CFR that stipulates that the EU Charter[20] in so far as it contains rights which correspond to rights guaranteed by the (ECHR), the meaning and scope of those rights shall be the same as those laid down by the ECHR. However, Art. 52(3) CFR does not prevent Union law providing more extensive protection. Accordingly, a very close relationship is established between the EU Charter and the ECHR.

The reference to the ECHR also covers the interpretation of the provisions of the ECHR as made by the European Court of Human Rights (ECtHR) which means that the Court of Justice of the EU (ECJ) shall interpret the provisions of the EU Charter in accordance with ECtHR case law on the corresponding provisions of the ECHR.[21] Contrary to Art. 6(1) ECHR, Art. 47 CFR is not limited to the determination of civil rights or obligations or of a criminal charge which means that administrative disputes are also included in its scope.[22]

The ECJ has specified that the right to a fair trial derives *inter alia* from Art. 6 ECHR and constitutes a fundamental right which the EU respects as a general principle under Art. 6(2) TEU.[23] Title VI of the EU

[17] E.g. Pyle, *The Economics of Crime and Law Enforcement*, (Macmillan Press, 1983), chapters 6-9.
[18] In that direction, Bunting, "From editorial obligation to procedural accountability: policy approaches to online content in the era of information intermediaries", (2018) *Journal of Cyber Policy*, 175.
[19] Goodin, "Rough Justice", (2019) *Jus Cogens*, p. 83 et seq.
[20] The EU Charter is written into art. 6(1) of the TEU.
[21] Cf. "Explanations relating to the EU Charter of Fundamental Rights", O.J. 2007, C 303/17. See also Hazelhorst, "The Right to a Fair Trial in Civil Cases", (Springer – Asser Press, 2017), pp. 126 and 130.
[22] Ibid., p. 130; and Ward, "Article 47 – Right to and Effective Remedy and to a Fair Trial", in Peers, Hervey, Kenner and Ward (Eds.), *The EU Charter of Fundamental Rights: A Commentary* 2nd ed. (Bloomsbury Publishing Plc, 2022), at 47.19.

[23] Case C-305/05, *Ordre des barreaux francophones et germanophone, et al. v. Conseil des ministres,* EU:C:2007:383, para. 29. See also case T-351/03, *Schneider Electric SA v. Commission*, EU:T:2007:212, paras. 181-182. See also Art. 2 TEU.

Charter has the headline 'Justice' and the primary provision in this title is Art. 47 on right to an effective remedy and to a fair trial. Referring to the case-law of the ECtHR,[24] in Case C-305/05, *Ordre des barreaux francophone,* the ECJ explains that the concept of 'a fair trial' referred to in Art. 6 ECHR consists of various elements, which include, inter alia, the rights of the defence, the principle of equality of arms, the right of access to the courts, and the right of access to a lawyer both in civil and criminal proceedings.[25]

Art. 47 CFR consists of three parts. The first part establishes a right to an effective remedy.[26] The third part ensures legal aid to those who lack sufficient resources. The second part, that is the most pertinent for the purpose of this article and for the concept of justice, reads:

> *'Everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal previously established by law. Everyone shall have the possibility of being advised, defended and represented.'*[27]

3.1. Fair hearing and information

The ECJ has held that the right to a fair hearing is, in all proceedings initiated against a person which are liable to culminate in a measure adversely affecting that person, a fundamental principle of Community law and must be guaranteed even in the absence of any rules.[28] The right to a hearing is an essential element in the rights of the defence which according to the ECJ is a fundamental principle of Community law. That principle is infringed where a judicial decision is based on facts and documents which the parties themselves, or one of them, have not had an opportunity to examine and on which they have therefore been unable to comment.[29]

According to the practice of ECtHR the requirements of a fair hearing are stricter in the sphere of criminal law than under the civil and. In general, the states have greater latitude in civil law matters.[30] In addition, in the case law of ECtHR, it has been emphasized that the assessment of the proceedings shall be made of the proceedings as a whole in order to determine whether the applicant was guaranteed a fair hearing.[31] The implication of the requirement of assessing the proceeding in its entirety, is that there is not one correct way of conducting a fair civil trial.[32]

---

[24] ECtHR, *Golder v United Kingdom*, Appl. No. 4451/70, judgment of 21 February 1975, paras. 26 to 40; ECtHR, *Campbell and Fell v United Kingdom*, Appl. No. 7819/77; 7878/77, judgment of 28 June 1984, paras. 97 to 99, 105 to 107 and 111 to 113; and ECtHR *Borgers v Belgium*, App. No. 12005/86, judgment of 30 October 1991, para. 24.
[25] Para. 31. Cf. Borraccetti, "Fair Trial, Due Process and Rights of Defence in the EU Legal Order" in Di Federico (Ed.), *The EU Charter of Fundamental Rights from declaration to binding instrument*, (Springer, 2011), pp. 95-107, 100.
[26] Cf. Art. 13 ECHR.
[27] The provision corresponds to Art. 6(1) ECHR that i.a. entitles everyone 'to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law'.
[28] Joined cases C-439/05 P and C-454/05 P, *Land Oberösterreich and Austria v. Commission,* EU:C:2007:510, para. 36.
[29] Case C-199/99, *P Corus UK Ltd,* paras. 19 and 41. See also C-348/16, *Sacko*, EU:C:2017:591, para. 37.
[30] ECtHR, *Peleki v. Greece,* Appl. No. 69291/12, Judgment of 7 September 2020, para. 70.
[31] ECtHR, *Centro Europa 7 S.r.l. and di stefano v. Italy*, Appl. No. 38433/09, judgment of 7 June 2012, para. 197.
[32] Hazelhorst, *The Right to a Fair Trial in Civil Cases,* (Springer – Asser Press, 2017), p. 132.

3.2. Equality of arms

The principle of equality of arms basically means that there should be ensured a 'fair balance' between the parties in a dispute resolution process and that both parties are treated equally.[33] That implies a fair balance in respect of opportunity to present each party's case, access to relevant information, available procedural means and available expertise and resources.[34] In Case C-682/15, *Berlioz Investment Fund SA v. Directeur de l'administration des contributions directes*,[35] the ECJ states that the principle of equality of arms, which is a corollary of the very concept of a fair hearing, implies that each party must be afforded a reasonable opportunity to present his case, including his evidence, under conditions that do not place him at a substantial disadvantage vis-à-vis his opponent.[36] In the application of the principle, the ECJ provides a certain amount of flexibility, which can prove particular relevant in respect of content moderation on online platforms because the practical circumstances and considerations are very different from other dispute resolution systems. The Court has thus held that the question whether there is an infringement of the rights of the defence and of the right to effective judicial protection, must be examined in relation to the specific circumstances of each case, including the nature of the act at issue, the context of its adoption and the legal rules governing the matter in question.[37]

3.3. The right to a reasoned judgment

The practical circumstances surrounding decisions on moderating content on online platforms may create a risk that decisions are not sufficiently reasoned.

The right to a reasoned judgment is established by the case law of the ECJ and the ECtHR. The former has stated that the observance of the right to a fair trial requires that all judgments be reasoned to enable the defendant to see why judgment has been pronounced against him and to bring an appropriate and effective appeal against it.[38] However, according to the ECJ fundamental rights do not constitute unfettered prerogatives and may be subject to restrictions, provided that the restrictions in fact correspond to objectives of general interest pursued by the measure in question and that they do not constitute, with regard to the objectives pursued, a manifest and disproportionate breach of the rights thus guaranteed.[39] Thus also in this respect the ECJ provides for a certain degree of flexibility.

---

[33] ECtHR, *Feldbrugge v. The Netherlands*, Appl. No. 8562/79, judgment of 29 May 1986, para. 44.
[34] ECtHR, *Steel and Morris v. The United Kingdom*, Appl No. 68416/01, judgment of 15 May 2005, para. 72. See Bar-Ziv & Elkin-Koren, "Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown", (2018) *Connecticut Law Review*, 380-382: 'In other words, the online adjudication of copyright disputes is still dominated by repeat players that acquired the expertise in managing online disputes, while end users may still suffer from the power/knowledge gap.'
[35] EU:C:2017:373.
[36] Para. 96. Similar in ECtHR, *Kress v. France*, Appl. No. 39594/98, judgment of 7 June 2001, para. 71.
[37] Ibid., para. 97, and Joined Cases C-514/07 P, C-528/07 P and C-532/07 P, *Sweden and Others v. Association de la presse internationale ASBL*, EU:C:2010:541, para. 102.
[38] Case C-619/10, *Trade Agency Ltd v. Seramico Investments Ltd,* EU:C:2012:531, para. 53.
[39] Ibid., para. 55.

On the same lines the ECtHR has stipulated that national courts must indicate with sufficient clarity the grounds on which they based their decision. The reasoned judgement, inter alia, makes it possible for the accused to exercise usefully the rights of appeal available to him.[40]

Another rationale behind the right to a reasoned judgment is to ensure publicity of the legal reasoning that enables the public to predict valid law.[41]

4. Codes and rules on internet content moderation

A number of attempts to establish codes for fair trial etc. on the internet have been presented. In the following, I will examined three such non-binding codes in addition to the provisions on content moderation in the recently adopted Digital Services Act. There are quite diverse in their scope and substance.

'The Santa Clara Principles On Transparency and Accountability in Content Moderation' are issued by a coalition of organizations, advocates, and academics. Originally, the Santa Clara Principles 1.0 from 2018 were purely procedural principles that did not address substantial principles and norms. They were not very detailed and had a limited scope. That changed somewhat when the elaborated Santa Clara Principles 2.0 were published in 2021. The Principles outline minimum standards that tech platforms must meet in order to provide adequate transparency and accountability around their efforts to take down user-generated content or suspend accounts that violate their rules.[42] Despite a number of the very large global tech companies endorse the principles, it is mentioned on the website of the Santa Clara Principles that very few companies have fully met the demands of the principles.

The Aequitas Principles on Online Due Process are issued by 'DigiScholar - The Digital Scholarship Institute' which mostly consists of academics. The purpose of the Principles is to offer guidance on due process principles, including procedural fairness and natural justice, and guide the design and implementation of platform processes so that human rights, fundamental freedoms, and the rule of law are restored and maintained in the online environment.[43]

The establishment of a proper balance among different fundamental rights, including but not limited to the freedom of expression, protection of personal data, and protection of intellectual property rights – as well as upholding the rule of law in preventing crime online – all depend on the proper application of due process in the online environment.

Finally, the Council of Europe adopted on 7 March 2018 a recommendation on the roles and responsibilities of internet intermediaries.[44] The text of the full Recommendation is relative

[40] ECtHR, *Hadjianastassiou v. Greece*, Appl. No. 12945/87, judgement of 16 December 1992, para. 33.
[41] Hazelhorst, *The Right to a Fair Trial in Civil Cases*, (Springer – Asser Press, 2017), p. 150 f.
[42] <https://santaclaraprinciples.org/history/> (last visited 27 January 2023).
[43] <https://aequitas.online/> (last visited 27 January 2023).
[44] Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.

comprehensive and covers a broad range of different intermediaries.[45] Paragraph 2.3. of the Recommendation is dedicated to Content moderation.

4.1. A substantial norm to prevent over-enforcement

A substantial norm[46] for content moderation is needed in order to prevent platforms from moderating legal and compliant content. In the absence of such a norm, the platforms will have incentives to rather moderate too much than too little, to be on the safe side. A substantial norm thus provides a countervailing obligation for platforms not to over-enforce. In the various codes this substantial norm is based on human rights.

It is stated in The Santa Clara Principles that companies shall ensure that human rights are integrated at all stages of the content moderation. This implies, that users of a platform service shall be informed of inter *alia* how the company that supply the service has considered human rights—particularly the rights to freedom of expression and non-discrimination—in the development of its rules and policies. The company shall also inform users on how the company has considered human rights in automated processes in content moderation in situations where a company uses such techniques.

The Aequitas Principles express a bit more firm obligation to respects human right. The principles thus stipulate that human rights and fundamental freedoms must be upheld everywhere and should permeate all activities of platforms. In addition, platforms should take an active role in preserving human rights and fundamental freedoms in the online world. However, the obligation to respect human right is included in the first part of the Principles on 'Scope of principles' and the obligation is not made operational and serves merely as a non-binding guideline for supporting human rights values.

The only mentioning of human rights in paragraph 2.3. of the Council of Europe Recommendation on content moderation is found in paragraph 2.3.4. and 2.3.5. Paragraph 2.3.4. stipulates that all members of staff of intermediaries who are engaged in content moderation shall be given adequate initial and ongoing training on the applicable laws and international human rights standards. In addition paragraph 2.3.5. stipulates that intermediaries should carefully assess the human rights impact of automated content management. In other parts of the Recommendation it is stated that member states shall take all necessary measures to ensure that internet intermediaries fulfil their responsibilities to respect human rights.[47]

Neither of the codes specify in which way human rights shall be respected and it seems that the codes merely suggest that the internet platforms take human rights into consideration. The codes are non-binding and international human rights law is binding on states only, not on individuals or companies. Accordingly, states are the primary duty bearers. However, when a state fulfil its obligations under international human rights law, international human rights standards have legal impact on companies.[48]

---

[45] The Preambel of the Recommendation, para. 4.
[46] As opposed to a procedural rule.
[47] The Preambel of the Recommendation, para. 12.
[48] Cf. Laidlaw, *Regulating Speech in Cyberspace Gatekeepers, Human Rights and Corporate Responsibility*, (Cambridge University Press, 2015), p. 48, who suggest that an internet information gatekeeper's '*human rights responsibilities should increase or decrease based on the extent that its activities facilitate or hinder democratic*

A binding rule that resembles a substantial norm to prevent over-enforcement is now found in the Digital Services Act.[49] Art. 14(4) stipulates that providers of intermediary services shall act in a diligent, objective and proportionate manner in applying and enforcing the restrictions that they impose in relation to the use of their service in respect of information provided by the recipients of the service, with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the EU Charter.[50]

The wording of the provision resembles the corresponding provisions in the codes and it is doubtful whether the provision will have any significant effect on over-enforcement. Due regard to fundamental rights, namely freedom of expression, does not restrict the platforms possibilities of deciding for themselves what type of otherwise legal content they deem as incompatible content in their terms and conditions, jf. Art. 14(1) Digital Services Act.

4.2. Transparency

Transparency into the processes of content moderation plays an important role in various contexts. Firstly, it enables a person whose content has been moderated to obtain an explanation for the reasons behind the moderation.[51] Secondly, it provides an opportunity to detect errors and enhance accuracy. Thirdly, in cases of automated and semi-automated content moderation it provides an opportunity to detect and address biases. Fourthly, it enables persons to assess the quality of the moderation processes and thus the legitimacy of the result of the moderation processes. In addition to accuracy and biases, quality also relies on whether a human review has been involved, and if so, the time the review has allocated to the review and the qualifications of the reviewer.

The platforms being commercial enterprises are reluctant to provide a high degree of transparency[52] and they often claim in respect of automated and semi-automated content moderation that the pertinent algorithms and data are protected as trade secrets and shall be inaccessible to the general public.[53]

---

culture. This scale of responsibility is reflected not only in the reach of the gatekeeper, but also in the infiltration of that information, process, site or tool in democratic culture'. In general see, Jørgensen, "Human Rights and Private Actors in the Online Domain", in Land and Aronson (Eds), *New Technologies for Human Rights Law and Practice*, (Cambridge University Press, 2018), pp. 243-269, 253 and 260-261; and Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, 15 (Apr. 6, 2018).

[49] The aim of the Digital Services Act is to contribute to the proper functioning of the internal market for intermediary services, cf. art. 1(1).

[50] Cf. para. 52 of the preamble, that specifies that these safeguards shall be "robust".

[51] Grimmelmann, "Regulation by Software," (2005) *Yale Law Journal* , 1737. See e.g. Suzor, Myers West, Quodling and York, "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation", (2019) *International Journal of Communication*, 1526–1543.

[52] Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, 15 (Apr. 6, 2018), para. 40.

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

The Santa Clara Principles endorse transparency in its 'Foundational Principles' but under its 'Operational Principles' only provides a very limited degree of transparency in the moderation practices and the appeal procedures. According to Operational Principles 1, companies should publish information on total number of pieces of content actioned and accounts suspended, number of appeals, share of successful appeals etc. Companies should also publish information on when and how automated processes are used, the key criteria used by automated processes for making decisions and the accuracy of the processes. Basically, the transparency obligation primarily apply to the results of the platforms content moderation practices and not to how the practices function and the rationales underlying them.

The Aequitas Principles contain much more elaborated provisions on transparency compared to the Santa Clara Principles and part. 5 of the Principles is dedicated to 'Algorithmic Transparency'. Accordingly, at a minimum, a platform shall provide:

- descriptions of the algorithms that the platform utilizes,
- the logic underpinning algorithms,
- the method in which the algorithms process or use personal data,
- the characteristics of individual persons (gender, race, religion, age, etc.) that the algorithms might utilize,
- detailed logs and auditable data concerning the ordinary operation, output etc.
- details of any reported or identified algorithmic bias, the actual or potential causes of the bias, steps taken to mitigate any harmful bias and a recognition of any unmanageable bias effects,
- engagement rates for all instances where the algorithms have amplified misinformation,
- open source and/or non-confidential elements of the algorithms that may assist in the understanding of platform decision-making,
- any gaps in transparency concerning the functioning of the algorithms that might be caused by technical limitations.

According to the Council of Europe Recommendation intermediaries' transparency obligations, firstly, apply to service agreements and policies, and the process of drafting and applying terms of service agreements, community standards and content-restriction policies should be transparent, accountable and inclusive. Secondly, internet intermediaries should clearly and transparently provide meaningful public information about the operation of automated data processing techniques in the course of their activities. Thirdly, intermediaries should regularly publish transparency reports that provide clear (simple and machine-readable), easily accessible and meaningful information on all restrictions to the free and open flow of information and ideas and all requests for such restrictions.[54]

All three codes acknowledge the importance of transparency obligations, however, it is only the Aequitas Principles and the Council of Europe Recommendation that actually stipulate firm and detailed transparency obligation relation to platforms.

In the Digital Services Act the transparency obligations in Arts. 14 and 15 seem to be less strict than at least the Aequitas Principles. This concerns especially the transparency reporting obligation related to automated content moderation in Art. 15(1)(e) that requires information on 'any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the

[54] Para. 2.2.

automated means used in fulfilling those purposes, and any safeguards applied'. The provision does not seem to require transparency into how the algorithms are working and the logic underlying them which is essential for namely the legitimacy of automated content moderation.

4.3. Fair trial/due process

The Santa Clara Principles state that companies should ensure that due process considerations are integrated at all stages of the content moderation process. This foundational principle is implemented by informing the users on how the company has considered the importance of due process when enforcing its rules and policies, and in particular how the process has integrity and is administered fairly.[55]

Under the Santa Clara Principles companies must provide notice to each user whose content is moderated, about the reason for the moderation. The notice shall i.a. express the specific clause of the guidelines that the content was found to violate and include an explanation of the process through which the user can appeal the decision, including any time limits or relevant procedural requirements. Companies should ensure that appeal includes human review by a person or panel of persons who were not involved in the initial decision. Furthermore, users should be offered an opportunity to present additional information in support of their appeal that will be considered in the review. Finally, the companies should ensure that the user receives notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision.

Part 2 of the Aequitas Principles requires that due process should be reflected in platforms' decision-making concerning user-generated content. 'Due process' include the following principles:

- A fair and transparent review within a reasonable time and by an independent and impartial decision-maker
- Proper prior notification to the parties
- An opportunity for the parties to respond and present evidence
- The right to legal representation
- The right to appeal to an internal appeals panel, alternative dispute resolution panel or competent court, and
- The right to receive a decision in writing which clearly articulates the reason for the decision.

The Council of Europe Recommendation is not very specific on how to ensure a fair trial but basically states that intermediaries should make available effective remedies and dispute resolution systems that provide prompt and direct redress in cases of user, content provider and affected party grievances. While the complaint mechanisms and their procedural implementation may vary with the size, impact and role of the internet intermediary, all remedies should allow for an impartial and independent review of the alleged violation. These should – depending on the violation in question – result in inquiry, explanation, reply, correction, apology, deletion, reconnection or compensation.[56]

[55] Para 1.
[56] Para. 2.5.1, cf. para. 1.5.2.

All three codes require a certain degree of human involvement in the decision-making process associated with content moderation, particularly at the appeals stage. The Aequitas Principles and the Council of Europe Recommendation also require that platforms should provide for proper, adequate and professional qualifications and ongoing training for reviewers and other decision-makers.[57] As the only one of the three codes, the Council of Europe Recommendation adds that staff should also be provided with appropriate working conditions which includes the allocation of sufficient time for assessing content and opportunities to seek professional support and qualified legal advice where necessary.[58] These are all very sound principles, but the wording is imprecise and not very operational.

In the Digital Services Act, the fair trial dimension in content moderation is primarily addressed in Arts. 16, 17 and 20. Art. 16(2) specifies the required content of a notice of which the most important issue is a sufficiently substantiated explanation of the reasons why the individual or entity alleges the information in question to be illegal content.[59] The provider shall also, without undue delay, notify that individual or entity of its decision in respect of the information to which the notice relates, providing information on the possibilities for redress in respect of that decision.[60] The platform shall provide a clear and specific statement of reasons to any affected recipients of the service for the restrictions imposed.[61] Art. 17(2) specifies in some details the minimum information contained in the statement of reasons which i.a. includes information on the redress possibilities available to the party affected in respect of the decision, in particular, where applicable through internal complaint-handling mechanisms, out-of-court dispute settlement and judicial redress.

Art. 20 in the Digital Services Act stipulates requirements to platforms' and other service providers' appeal process which in Art. 20 is called 'Internal complaint-handling system'. After the platform has taken a decision on moderating or not moderating content, all affected parties can free of charge bring a complaint. The provision does not specify how much evidence is expected or required but simply states that the complaint shall be sufficiently precise and adequately substantiated.[62] Online platforms shall ensure that the decisions on the submitted complaints are taken under the supervision of appropriately qualified staff, and not solely on the basis of automated means.[63] The online platforms are only required to inform complainants of their reasoned decision on the complaints.[64]

Art. 17 ensures that a reasonably detailed explanation of the decision taken is sent to the party affected and thus complies with the fundamental right to a reasoned judgment. The Digital Services Act contains no provisions on presentation and amount of evidence which indicates that the platform in that respect can make its own guidelines.

The process of content moderation as envisaged in the Digital Services Act is problematic in view of the adversarial principle and the equality of arms principle. The result of the notice and action mechanism in Art. 16 functions as a decision of first instance and that decision is exclusively made on the basis of the

[57] The Santa Clara Principles differentiates from the two other codes by only requiring the humans involved in content moderation possess cultural competence, cf. para. 3.
[58] Para. 2.3.4.
[59] Art. 16(2)(a).
[60] Art. 16(5).
[61] Art. 17(1).
[62] Art. 20(2).
[63] Art. 20(6)
[64] Art 20(5).

information submitted in the notice. The party making the content available is not heard at this stage. It can be expected that in most cases, a notice on illegal content is assumed to be correct because the underlying information is not contested[65] and, for the same reason, that more errors occur when platforms decide to moderate content compared to the situation when they decide not to moderate content. This imbalance leads to over-enforcement and is further exacerbated by studies that show that where a right to appeal is available, in practice it is rarely exploited, and overall has failed to provide effective remedies.[66] These circumstances, create substantial inequality between the parties for the benefit of the parties submitting notices.[67]

No less troubling is the overall structure of the notice and action model in Arts. 16 and 17. A notice is filed on alleged illegal content. Either the platform decides to moderate or not to moderate. In the former situation, the party whose content has been moderated is given one chance, by means of a complaint, to have the content re-established into its original form. In the latter situation, the party submitting the notice is given two chances to have the content moderated. First chance is the notice and the second chance is the complaint if the platform decides not to moderate in the first instance. This additional imbalance between the parties further challenges the equality of arms principle and leads to over-enforcement.

5. Liability of sharing platforms

Cases where it is obvious that online content is illegal or incompatible are, as a point of departure, unproblematic and should as a matter of course be moderated.[68] However, borderline cases exist and errors occurs no matter whether the decisions are automated or made under human review.[69] Borderline cases increases the risk of errors in content moderation. In relation to the assessment of illegality of content the final say on the matter lies at the ordinary courts. Incompatible content is somewhat different because the platforms design their own policies on incompatible content and the policies are practiced by each individual platform, and it is likely that there are no comparable practice outside each platform that can guide the assessment of compatibility. For that reason, it must be assumed that more errors occur in decisions concerning legality than concerning compatibility.

The provision on content moderation in Art. 17 of the Directive on copyright in the digital single market[70] (DSM Directive) on one hand requires that sharing platforms that have not entered into a

[65] Van Loo, "Federal Rules of Platform Procedure", (2021) *The University of Chicago Law Review*, 850.
[66] Elkin-Koren, "Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence", (2020) Big Data & Society, 7(2), available at <https://doi.org/10.1177/2053951720932296> (last visited 27 January 2023).
[67] See also Bar-Ziv & Elkin-Koren, "Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown, (2018) Connecticut Law Review, 381-382: '*… a surprising finding emerged from this study: N&TD serves mainly large players; more specifically, multinational companies*'.
[68] This is only the point of departure. Below, I present an argument of less moderation of incompatible content.
[69] Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation", (2020) *Fordham International Law Journal*, 968 and 1005; and Douek, "Facebook's oversight board: move fast with stable infrastructure and humility", (2019) *North Carolina Journal of Law & Technology*, 6.
[70] 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.

licensing agreement with the right holders shall be liable for copyright infringement, unless the service providers demonstrate that they have made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of infringing specific works which the right holders have provided the service providers with the necessary information, and have acted expeditiously with respect to taking down the infringing works.[71] On the other hand, the cooperation between sharing platforms and right holders shall not result in the prevention of the availability of works uploaded by users, which do not infringe copyright, including where such works are covered by an exception or limitation. Furthermore, member states shall ensure that users are able to rely on existing exceptions or limitations on quotation, criticism, review, caricature, parody and pastiche, when uploading and making available content generated by users on sharing platforms.[72]

The DSM-Directive's Art. 17(4) is an exemption from copyright liability established by Art. 17(1) and platforms are liable for copyright infringement if the conditions of Art. 17(4) are not satisfied. The open question is whether platforms could be liable for taking down content that does not infringe copyrights. The immediate answer is no. There is no in-built liability standard in Art. 17(7) and, in principle, international human rights law is binding on states only.[73] States then have to adopt a national provision on liability to cover that situation. Finally, platforms are free to make their own policies also on content moderation[74] and therefore they can word their terms of service in a way that allows them to make errors and over-enforce in borderline cases.[75]

That being said, the CJEU has made a very strict interpretation of Art. 17(7) in Case C-401/19, *Republic of Poland v. European Parliament and Council of the European Union*.[76] The Court states that contrary to the requirement of making their 'best effort' to take down copyright infringing content in Art. 17(4), Art. 17(7) prescribes a specific result to be achieved, not just a best effort.[77] Furthermore, the Court reiterates that although the protection of intellectual property rights is enshrined in Art. 17(2) CFR, there is nothing whatsoever in the wording of that provision or in the Court's case-law to suggest that that right is inviolable and must for that reason be protected as an absolute right.[78] According to the Advocate General, the EU legislature by means of Art. 17(7) has expressly recognised that users of sharing services have subjective rights under copyright law. Those users now have the right, which is enforceable against the providers of those services and rightholders, to make legitimate use, on those services, of protected subject matter, including the right to rely on exceptions and limitations to copyright and related rights.[79] In his opinion, the Advocate General also finds that platforms may no longer exclude the application of exceptions and limitations in their terms of service or in contractual agreements with right holders by providing, for example, that a mere allegation by right holders of infringement of copyright will be sufficient to justify such blocking or removal. Following this argument,

[71] Art. 17(4).
[72] Art. 17(7), cf. the preamble recital 70.
[73] E.g. Land, "Against Privatized Censorship: Proposals for Responsible Delegation", (2020) *Virginia Journal of International Law*, 389-393.
[74] Cf. Art. 14(1) of the Digital Services Act.
[75] The Advocate General in Case C-401/19, *Republic of Poland v. European Parliament and Council of the European Union*, EU:C:2021:613, disagrees on the last issue.
[76] EU:C:2022:297.
[77] Para. 78, cf. the Opinion of the Advocate General, para. 165.
[78] Para. 92.
[79] Para. 161.

the platforms should only remain free to remove content which falls within the scope of exceptions or limitations on grounds other than copyright issues, for example if it is insulting or contravenes their nudity policy.[80] However, the CJEU does not refer to this line of argument. Hence, it is not at all clear whether the Court shares the opinion of the Advocate General or whether the platforms are free to include in their terms of service provisions according to which a mere allegation by right holders of infringement of copyright will be sufficient to justify such blocking or removal or provisions that mandate the platforms to block or remove content that give rise to borderline copyright infringement cases without further examination.

Within the scope of the DSM Directive there is now in Art. 17(7) a harmonized rule stating that sharing platforms shall not take down copyright protected content that does not infringe copyright. The provision does not expressly stipulate a liability norm which means that if a platform should be liable for taking down content that does not infringe copyright, possibly it requires a liability norm in national law. International human rights law is inapplicable in relation to private parties unless national law provides for this. In a few special instances, ordinary principles of tort law may perhaps provide a liability norm that does not infringe copyright. However, a platform is not liable if it takes reasonable care not to over-enforce in content moderation and even if it does not take reasonable care it will in most cases be very difficult for the users to demonstrate injury or harm to their property. This creates an argument for adopting a national rule on liability for internet platforms over-enforcement.

6. Privatization of justice

Privatization of justice occurs when rights-enforcement is left for private parties and when private parties substitute public rules with private rules. The reason why this is problematic is that public rules pursue societal objectives and values whereas private rules must be assumed to pursue private objectives and values. Accordingly, privatization of justice and the associated content moderation must be assumed primarily to be governed by corporate philosophy and the maximization of the wealth of the entity that privatizes the justice, in the present case internet platforms.[81] It is claimed that in automated content moderation, the algorithms are designed to maximize profits for their facilitators and are therefore likely to be commercially biased in non-transparent ways.[82]

The problem of privatization of justice creates an argument for adopting rules for the purpose of aligning the internet platforms terms of service on incompatible content with public rules.[83] According

[80] Para. 163.
[81] Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation", (2020) *Fordham International Law Journal*, 948 ff; Land, "Against Privatized Censorship: Proposals for Responsible Delegation", (2020) *Virginia Journal of International Law*, 410-411; Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, 15 (Apr. 6, 2018), para. 17.
[82] Elkin-Koren, "Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence", (2020) *Big Data & Society*, 6.
[83] Bunting, "From Editorial Obligation to Procedural Accountability: Policy Approaches to Online Content in the Era of Information Intermediaries", (2018), *Journal of Cyber Policy*, 165, 174; Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation", (2020) *Fordham International Law Journal*, 970.

to Evelyn Douek, the governance of private dispute resolution requires weighing economic, social, and moral factors.[84]

7. Conclusion: Rough justice on online platforms

For practical reasons it is obvious that the full range of fair trial principles and values cannot be upheld in the content moderation activities of internet platforms because for the parties involved it is too burdenful and costly.[85] It must be fair trial in a modified and reduced form. In another context than internet platforms' content moderation, Peter Linzer states the obvious general principle that 'rough justice is infinitely better than no justice'.[86] When Linzer refers to 'rough justice' he means justice that is driven more by general standards of fairness than by structured (or formal) systems of rules and neat categories, justice that is often untidy, that may be second-best where the best is unachievable.[87]

The concept of 'rough justice' has also been used in relation to the ICANN's Uniform Domain-Name Dispute-Resolution Policy, which is a policy on entitlement to domain names, primarily, enforced by approved dispute-resolution service providers.[88] However even that procedure is too comprehensive and resource-demanding to translate into content removal and moderation by internet platforms.

By taking bits and pieces from the different codes and rules on content moderation and adding a few more a model for rough justice on internet platforms emerges. The following model of rough justice is divided into 3 different parts: 1) Procedural rules, 2) substantive rules, and 3) competences.

7.1. Procedural rules

In respect of procedural rules, there is a need for much more transparency in the functioning of algorithms and working conditions and qualifications of humans where human review is involved. At the moment nothing suggests that platforms voluntarily are willing to ensure transparency in respect of their algorithms. Typically, internet platforms assert that their algorithms are legally protected as trade secrets. The Digital Services Act contains provisions that aim at ensuring protection of very large online platforms' and search engines' trade secrets.[89] Similarly, the Aequitas Principles exclude confidential elements of the algorithms from the transparency obligations.[90]

If the transparency obligations of internet platforms shall be sufficiently effective, the platforms ought not to be able to prevent transparency with reference to trade secrets protection. At first glance, it may seem as a radical proposition to eliminate trade secrets protection in content moderation algorithms all

[84] Douek, "Facebook's oversight board: move fast with stable infrastructure and humility", (2019) *North Carolina Journal of Law & Technology*, 16-17.
[85] E.g. Van Loo, "Federal Rules of Platform Procedure", (2021) *The University of Chicago Law Review*, 864 and 889.
[86] Linzer, "Rough Justice: A Theory of Restitution and Reliance, Contracts and Torts", (2001) *Wisconsin Law Review*, 766.
[87] Ibid., 695 fn.
[88] Mueller, "Rough Justice: A Statistical Assessment of ICANN's Uniform Dispute Resolution Policy", (2001) *The Information Society*, 151–163.
[89] E.g. Art. 40.
[90] Part 5.

together. However, it is necessary because the conditions for protection as a trade secret under the EU Trade Secrets Directive[91] can easily be satisfied by all parts of the algorithms. It simply requires that the content is secret, has commercial value because it is secret and it has been subject to reasonable steps to keep it secret.[92]

Transparency could be taken one step further. Danielle Keats Citron suggests that automated systems must be designed with transparency and accountability as their primary objectives and vendors should release systems' source codes to the public because opening up the source code would reveal how a system works.[93]

As to transparency into working conditions where moderation is under human review, David Kaye reports on automated content moderation supplemented by human review and how the biggest social media companies developing large teams of content moderators to review flagged content and they will typically be authorized to make a decision often within minutes about the appropriateness of the content and to remove or permit it.[94] Transparency into the working conditions and qualifications of humans involved in content moderation is the least controversial proposal because it does not incur costs on the internet platforms.

The Aequitas principles is the code with the most detailed principles on due process and refer to due process in all parts of the content moderation process, but the principles are not very specific on exactly what 'due process' in content moderation implies. The principles seem to assume that the normal understanding of 'due process' applies which obviously cannot be the case

There is a need for establishing a minimum level of measures that can satisfy the concept of 'due process in content moderation'. However, a model of rough justice must accept substantial limitations i.a. in regards to types of evidence, extent of evidence and number of pleadings compared to an ordinary court case.

It does not follow explicitly from the various code etc. how many stages in the moderation and appeal process are needed in order to comply with the codes etc. A process that involves two stages a complaint/automatic removal and a subsequent counter notice which works as an appeal may seem like a reasonable and manageable solution. The most problematic part of such a process is as mentioned above in relation to the Digital Services Act that in the first stage the user's content will be moderated before the user is heard which fits uneasily with the equality of arms principle. The hearing of users before content is moderated requires sharing platforms to invest more resources in content moderation and, hence, it is unlikely that the platforms will do this voluntarily which suggests legislative intervention.[95]

[91] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
[92] Art. 2(1).
[93] Keats Citron, "Technological Due Process", (2008) *Washington University Law Review*, 1308 f.
[94] Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, 15 (Apr. 6, 2018), para. 35.
[95] Cf. Van Loo, "Federal Rules of Platform Procedure", (2021) *The University of Chicago Law Review*, p. 849.

The human rights principle of a reasoned judgment seems to be adequately addressed by the provisions of the Digital Services Act and in the codes.

7.2. Substantive rules

The purpose of substantive rules on what to moderate is twofold. Firstly, it shall create a counter-weight to internet platforms' tendency to over-enforce. Secondly, it shall reduce moderation of incompatible but legal content. The latter purpose is to a certain extent controversial because platforms are private enterprises and as the point of departure, they have the freedom to enter into whatever agreements they like with their users. However, as suggested by Emily B. Laidlaw, an internet information gatekeeper's '*human rights responsibilities should increase or decrease based on the extent that its activities facilitate or hinder democratic culture*'.[96] And at least the very large internet platforms play a substantial role in facilitating or hindering democratic culture. Therefore substantial rules based on human rights would be an important means to align the platforms' terms of service to societal objectives and values and thus counteract the adverse effect of privatization of justice.

The fluffy references to human rights in the various codes and the Digital Services Act are not likely to have a substantial impact on facilitating democratic culture. Therefore, there is a need for legislation for the purpose of making the EU Charter and other parts of human rights direct applicable to large sharing platforms. There is a delicate balance between making human rights directly applicable to internet platforms and the platforms' freedom to issue the terms of service they seem appropriate. The platforms' freedom to decide their terms of service is acknowledge in Art. 14(1) of the Digital Services Act. Unless that provision is repealed, the alignment of the platforms' terms of service with societal objectives and values will, at best, be moderate. A compromise solution that does not require any statutory amendments is to adopt the advocate General's interpretation in his opinion in Case C-401/19 *Republic of Poland v. European Parliament and Council of the European Union*[97] according to which the platforms should remain free to remove content which falls within the scope of copyright's exceptions or limitations on grounds other than copyright issues.

The application of human rights often involves a balancing of the interests underlying the different fundamental rights. Thus, in relation to Art. 17 of the DSM Directive, the ECJ has stated that in the application of Art. 17, a fair balance has to be struck between the various fundamental rights protected by the EU Charter.[98]

In the context of copyright law, the balancing of interest is typically illustrated by the protection of intellectual property on one hand and access to culture and freedom of expression and access to culture on the other hand. In UNDHR article 17(2) the protection of 'intellectual property' is worded as 'the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.' Protecting the 'material interests' is not only about the interest of the author in the effective enforcement of right and prevention of infringements. The material interests may also

[96] Laidlaw, *Regulating Speech in Cyberspace Gatekeepers, Human Rights and Corporate Responsibility,* (Cambridge University Press, 2015), p. 48.
[97] EU:C:2022:297.
[98] Case C-401/19, *Republic of Poland v. European Parliament and Council of the European Union*, EU:C:2022:297, para. 99.

be the interest in the results of any scientific, literary or artistic production being disseminated. Artists, and particularly, upcoming artists, sometimes choose a business model that aims at disseminate the works on the internet as wide as possible without claiming remuneration in order to become renown and use that status to generate revenue in other ways. If such an artist's works are taken down from the internet platform, the material interests of the author is harmed. The acknowledgement that the dissemination of copyright protected works as wide as possible may be the material interests of the authors creates another factor in the balancing of human rights that may turn the scale towards under-enforcement rather than over-enforcement.

An alternative to making international human rights directly applicable to sharing platforms is to adopt a clear an adequate liability norm that covers sharing platforms taking down legal content. Such a norm is lacking at the moment[99] and is needed in order to equalize the platforms' incentives to under-enforce with the incentives to over-enforce.[100]

7.3. Competences

The last part of the model on rough justice concerns the competences of humans who are involved in content moderation. Most of the codes and the Digital Services Act contain provisions to that effect. According, to e.g. Art. 20(4) of the Digital Services Act platforms shall ensure that the decisions on the submitted complaints are taken under the supervision of appropriately qualified staff. Similar wordings are found in the codes. The European Council Recommendation suggest that staff should also be provided with appropriate working conditions.

One may critizise the various instruments for not being precise in the provisions on the professional qualifications and working conditions for humans involved in content moderation activities. On the other hand, it would clearly be very difficult to establish appropriate and precise criteria for professional qualifications and working conditions. Obviously, a trained judge or lawyer is not required, much less would suffice. More important than setting up precise standards for qualifications and working conditions, is to impose an obligation on the platforms to inform on the internal criteria for appropriate qualifications and working conditions (transparency), so the users of the platform themselves are able to assess the legitimacy of the content moderation process.

Algorithmic content moderation involves a risk of biases both original biases and developed biases. To reduce the biases and ensure accuracy, there must be a certain level of human involvement. The existing instruments suggest human involvement in the appeal stage which seems to be a sound principle. In addition to this, there is also a need of a certain degree of human involvement in the first stage of automated content moderation to rectify biases. It is claimed that the design of the algorithms always is the result of subjective choice and judgment on the part of programmers and, thus, that all

[99] Cf. above in para. 5.

[100] Cf. Sander, "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation", (2020) Fordham International Law Journal, 968.

algorithms are originally biased.[101] In addition, it is well-known from a number of studies that algorithms develop biases.[102]

A means to rectify biases is to impose a duty on platforms to do random tests of accuracy in the first stage of automated content moderation. The quality can be ensured if the platforms each establishes an oversight board comprised of person with special qualifications that substantially exceed the qualifications of the ordinary staff that moderate content.[103]

[101] Cf. *Artificial Intelligence A National Strategic Initiative*, Tencent Research Institute, CAICT Tencent AI Lab, Tencent open platform (Palgrave Macmillan, 2021), 199: '*The design of the algorithms is always the result of subjective choice and judgment on the part of programmers. Whether they can impartially write existing legal or moral rules into their programs is questionable. Algorithmic bias has become a problem that needs to be tackled. The translation of rules into code brings issues of opacity, inaccuracy, unfairness, and difficulty of investigation that require serious consideration and research*', p. 199.

[102] See in general Keats Citron, "Technological Due Process", (2008) *Washington University Law Review*, 1249-1313.

[103] Cf. Van Loo, "Federal Rules of Platform Procedure", (2021) *The University of Chicago Law Review* 870: *'Congress should thus consider mandating that each large platform fund an external oversight board comprised of salaried judges.*'

DRAFT 30 November 2022

Errors and the Quality of Automated Content Moderation: Regulatory Routes for Mitigating Errors

Sebastian Felix Schwemer[1][2]

Content

1. Introduction

In a(n overly) simplistic worldview, in any content moderation scenario there should be a "right" and a "wrong" outcome. In a *copyright* content moderation context, this is because such decision basically answers the question "is this illegal" and there should be only one answer.[3]

[1] Associate Professor, Centre for Information and Innovation Law (CIIR), University of Copenhagen, and Adjunct Associate Professor, Norwegian Research Center for Computers and Law (NRCCL), University of Oslo. E-mail: sebastian.felix.schwemer@jur.ku.dk.

[2] Acknowledgements: This research is conducted under the reCreating Europe project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870626. I thank student assistant Philip and Heidi for their help and Thomas Riis for valuable comments.

[3] This is disregarding the fact that, empirically, the perception by creative content producers may not correspond to the legal reality, see Dergacheva, D. Katzenbach, C., and Otero, Paloma (2022). "Losing authenticity: social media creators' perspective on copyright restrictions in the EU" *submitted to ICA 2023.* Furthermore, content

Sometimes such decision might be straightforward, for example, in instances where content – in the words of the European Commission– is "manifestly illegal".[4] Sometimes such decision might require a detailed assessment by domain experts or even call for the involvement of the courts. In any case, however, we should be able to assess the "quality" of such content moderation decision.

The moderation of content has been increasingly automated and the reliance on algorithms is not only here to stay but likely to continue to increase. Already in 2018, for example, the European Commission "encouraged" online platforms to take proactive measures to detect illegal content, including by automated means.[5] The Digital Services Act, even though setting outer boundaries via due diligence obligations, continues this tendency, for example, with the inclusion of a good Samaritan clause in Article 7 DSA.[6] Algorithmic copyright content moderation might by volume very well be one of today's biggest use cases of automated (micro) legal decision making (or, if you prefer a fancier term, the use of "AI"). In a manner of speaking, content moderation is legal decisions on steroids.

The Digital Services Act is the first European framework to provide a legal definition of "content moderation" (Article 3 lit. t DSA): it refers to the "the activities, automated or not, undertaken by providers of intermediary services aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or the recipients' ability to provide

---

moderation in practice is not only concerned with *illegal* information but also information that might be considered "lawful but awful" where no clear benchmark is available.

[4] See, e.g. in the context of See European Commission, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, Brussels, 4.6.2021 COM(2021) 288 final, pp. 20-23. See also AG Øe, who refers to "information the unlawfulness of which *is obvious from the outset*, that is to say, it is *manifest*, without, inter alia, the need for contextualization", Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 198 and also para. 201.

[5] Point 18 in Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, C/2018/1177. This specific recommendation, however, is only directed to instances where such measures are "appropriate, proportionate and specific" and furthermore, in the context, of algorithmic content moderation, subject to certain safeguards. See also, T Riis and SF Schwemer, "Leaving the European Safe Harbor, Sailing Towards Algorithmic Content Regulation", *Journal of Internet Law* (2019), 1–21, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3300159>.

[6] See, e.g. A Kuczerawy, "The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act" (*Verfassungsblog*, 12.1.2021) < https://verfassungsblog.de/good-samaritan-dsa/> (last accessed 10 November 2022); SF Schwemer, "Digital Services Act: A Reform of the e-Commerce Directive and Much More" *prepared for* A Savin, *Research Handbook on EU Internet Law (Edward Elgar 2022)*, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213014>.

that information, such as the termination or suspension of a recipient's account."[7] In essence, content moderation can contain a large variety of activities that address content that is illegal or deemed incompatible with the private regulatory framework by that respective platform in form of terms and conditions.[8] It is also distinct from content *recommendation*, which is not addressed in this article.

Much has been discussed on the multiple issues related to the large-scale automation both within copyright and beyond. In this context, "false positives" and "false negatives" are regularly used as examples in legal and policy discourses. This article aims at conceptualizing these notions and explore the implications. I argue that we need to increasingly focus on decision quality in (copyright) content moderation. First, I propose the benchmark for decision quality. Then, I explore the various mechanisms proposed in the context of Article 17 of the CDSM Directive and the Digital Services Act with a view to introduce where exactly in the decision-making process mitigation mechanisms come into play. Finally, I structure to what extent these safeguards or mitigation mechanisms influence decision quality.

2. What's the benchmark for decision quality in copyright content moderation?

In order to assess the quality of a decision, we need to rely on a benchmark. So, what is the "right" decision? In the context of content moderated by online platforms, I see two main starting points:[9] firstly, the existing substantive legal rules which are "operationalized" by (automated or manual) content moderation. Secondly, the existing private rules *inter partes* based on contract, be it in the form of terms and conditions or community guidelines. Related to this is compliance with those rules (i.e., their enforcement by the platform) including any secondary legislation regarding this. A final route, namely users' normative perception of either legal rules or policies by online platforms is outside the scope of the following analysis.

[7] On further definitions and conceptualisations of content moderation, see, João Pedro Quintais, Péter Mezei, István Harkai, João Carlos Magalhães, Christian Katzenbach, Sebastian Felix Schwemer, and Thomas Riis, "Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis", reCreating Europe Report (August 2022), p. 33ff.

[8] Note also the broad definition of terms and conditions in the Art. 2 lit. u DSA, which covers "all clauses, irrespective of their name or form, which govern the contractual relationship between the provider of intermediary services and the recipients of the service".

[9] A third perspective is omitted in this article: namely, instead of taking existing rules as starting point, requiring content moderation to consider how rules ought to be in a *de lege ferenda* or political perspective.

"Regulation in the real world is far from optimal, and it is perhaps unrealistic to believe that it ever will be"[10] notes Train (1991). Nonetheless, for this analysis, let me set out the following assumption for our basic analytical model: Let us assume that the European copyright framework regarding substantive rights represent the starting point of "optimal" regulation.[11] The intermediary liability (exemption) framework then is a second layer that can be adjusted for changing intermediaries'/platforms' behaviour, either by adjusting conditions for liability exemption themselves or by adjusting the procedural rules around it.

Transferred to the context of copyright content moderation[12] by online platforms and the impact on access to culture this means the following. The "quality" of copyright content moderation is correlated to access to culture, because access to culture is considered embedded in the existing copyright framework.[13] Since the existing framework is assumed to strike the appropriate balance between exclusivity in copyright protection and access to culture, any variation in that balance – beyond the margin of interpretation allowed by law – will impact on access to culture. Consequently, both excessive and insufficient content moderation will have a negative impact on access to culture. Simply put, excessive content moderation by platforms restricts access to culture.[14] Conversely, insufficient content moderation increases access to culture in the short run, but in a harmful way because it encroaches on the legitimate interest of copyright holders and thus distorts the optimal balance and concomitantly harms access to culture in the long run. In other words: the smaller the difference between actual content moderation performed by intermediaries and the correct application of the legal framework, the smaller the negative impact on access to culture.

The case becomes trickier when private regulation is at work. Per today, online platforms enjoy wide contractual freedom[15] and may, for example, voluntarily go "beyond" what is required by law. A popular non-copyright example of this is Instagram's policy on nudity and the

[10] Kenneth Train, *Optimal Regulation: The Economic Theory of Natural Monopoly* (MIT Press 1991) 297.

[11] See also Quintais et al. (2022), p. 55. By optimal regulation we mean that the legislative framework strikes the appropriate balance between conflicting interests and fundamental rights, namely by recognizing time-restricted exclusive rights and corresponding exceptions and limitations with a utilitarian view to incentivize creators. By "appropriate" we mean the balance that was struck as a result of the normal operation of a democratic legislative process. In other words, we do not mean to pass a value judgment on the desirability of such balance from the perspective of any normative theory or viewpoint about copyright law.

[12] The use of copyright moderation for non-copyright purposes and the use of non-copyright moderation (e.g. privacy) for copyright purposes is not addressed in the following.

[13] See Quintais et al. (2022), p. 53 f.

[14] Note, however, that AG Øe for example in light of the CJEU's case law argues that "the effectiveness of the protection of the rights of rightholders may justify certain cases of 'over-blocking'", see Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 183.

[15] But see, e.g., Article 14 DSA on Terms and Conditions.

differential treatment of male and female nipples.[16] Terms and conditions (i.e., the "contract" between user and platform) may also in certain instances where permitted deviate from the fallback substantive rules in the EU acquis. Since thus private regulation is permitted under the democratically legitimized legal framework, however, for the sake of argument, let us here also assume that this private regulation itself is an eligible benchmark for content moderation decision quality.[17]

3. Quality and Errors

In any case then, the "quality" of content moderation can in simple terms be described in terms of correct and false results. For simplicity, let us in the following differentiate between illegal content (i.e., infringement of copyright) and legal content (i.e., no infringement of copyright).[18]

The following attempt to describe outcomes is borrowed from statistics and popularly referred to in the context of content moderation (e.g., Sartor 2020[19]). There are four theoretical outcomes that need to be distinguished, as displayed in Figure 1 below. For simplicity, the following focusses on making illegal content unavailable without going into technical details or considering other measures such as, e.g., demotion, demonitisation or measures related to the user's account.[20]

Figure 1. Error Types in Copyright Content Moderation

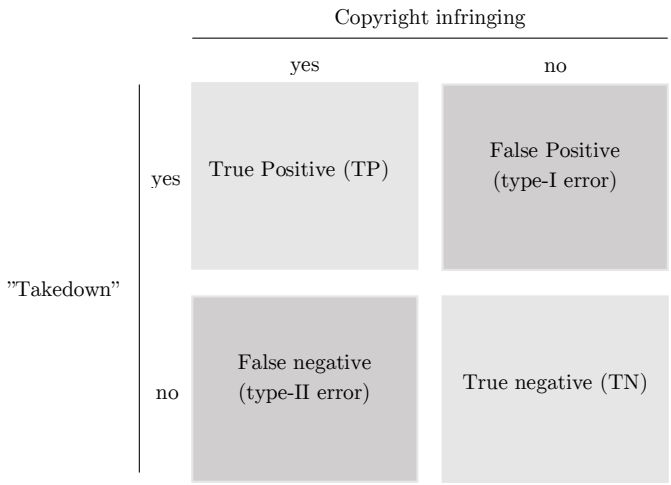
[16] https://www.nytimes.com/2019/11/22/arts/design/instagram-free-the-nipple.html

[17] That is despite that private regulation to the extent permissible by law, too, can restrict or distort "access to culture".

[18] Outside of copyright, we would also have to consider harmful content which is not necessarily illegal (i.e. "lawful but awful").

[19] Sartor, Giovanni and Loreggia, Andrea (2020). *The impact of algorithms for online content filtering or moderation.* European Parliament, Study requested by the JURI committee, PE 657.101.

[20] Cf. content moderation definition in Art. 3(t) DSA.



No error

The first set of outcomes relates to correct result of content moderation (i.e., the absence of error): if *illegal* content is taken down, there is no error in the moderation (true positive). The following example illustrates this scenario: Imagine that a copyright-infringing (i.e., not covered by a limitation or exception) musical work uploaded in its entirety by a user to the platform that is identified and removed by said platform's content moderation tools and practices. The platform's content moderation comes with the correct result.

Similarly, if *legal* content is *not* taken down, there is no error present in the moderation (true negative). The following example illustrates this scenario: Imagine the upload of a copyright-protected work that is covered by a limitation or exception and thus is not copyright-infringing and consequently not identified and reacted upon by said platform's content moderation tools and practices. Also in this scenario, the platform's content moderation comes with the correct result.

This set of outcomes (true positive and true negative) represents the optimal state of content moderation *based on benchmark for decision quality*; naturally, since this benchmark, as per above, is building upon the legal framework *de lege lata* (e.g., what is permitted under copyright law or not) or the existing private regulatory framework, it does not necessarily represent the optimal state of regulation (i.e., how the rules and private policies should be).

Error

The second set of outcomes relates to false results of content moderation, i.e., the presence of error.[21] Error is present, firstly, in instances where legal (i.e., non-infringing) content is taken down. This is also referred to as false positive (or type-I error). The following examples illustrates this false positive copyright content moderation scenario: Imagine a (copyrightable) work which is in the public domain but (falsely) identified as copyright-protected work and taken down. A second, far trickier copyright-specific example relates to a copyright-protected work, where its use is permitted because it falls under a limitation or exemption.[22] In the context of online content sharing service providers, Article 17 (7) CDSM Directive introduces a specific mandatory regime for selected limitations and exceptions, namely: (i) quotation, criticism, review; (ii) use for the purpose of caricature, parody or pastiche.[23] These limitations and exception –to varying degree– require a contextual analysis[24], profound knowledge of national and EU copyright law and sometimes the involvement of the CJEU, as, e.g., in the case of parody.[25]

Secondly, error is present in instances where illegal content is not taken down. This is also referred to as false negative (or type-II error). An example of this would be the unlicensed use of a copyright-protected work where the platform's content moderation fails to detect the copyright-infringing material.

The "upstream" question of the "right" balance

Based on the above assumption, this implies firstly that (1) any moderation by intermediaries that comes with type-I errors (false positives) or (2) type-II errors (false negatives) is assumed

[21] These errors can be introduced at various stages by the platform, for example, when its policies and practices deviate from substantive copyright rules or when the concrete decision fails to correctly assess limitations and exceptions.

[22] Article 17(7) first sub-paragraph CDSM Directive includes a general clause on exceptions and limitations, according to which the preventive obligations in 4(b) and (c) should not prevent that content uploaded by users is available on OCSSPs if such an upload does not infringe copyright, including if it is covered by an exception and limitation.

[23] These were optional in Articles 5(3)(d) and (k) InfoSoc Directive. See also MAPPING REPORT p. 99 ff. See on the special role of these selected limitations and exemptions also recital 70.

[24] See on this point also the European Commission's Guidance on Article 17 which notes that the scope and meaning of the notions of pastiche, criticism and review in Article 17(7) CDSM Directive "must be determined by considering their usual meaning in everyday language, while also taking into account the context in which they occur and the purposes of the rules of which they are part" with reference to the CJEUs case law C-476/17 Pelham, para 70, and C-201/13 Deckmyn, para 19. See European Commission, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, Brussels, 4.6.2021 COM(2021) 288 final, p. 19.

[25] C-201/13 – Deckmyn, Judgment of the Court (Grand Chamber), 3 September 2014, ECLI:EU:C:2014:2132.

to have a negative impact on access to culture.[26] This suggests, in other words, that the pure availability of content is not always optimal for access to culture at least in a broader copyright context.[27] Secondly, if the basic hypothesis and framework are accepted, then true positives and true negatives are not detrimental to access to culture; their impact on access to culture is neutral.[28] The simplified nature of this descriptive model, however, comes with several drawbacks. Chief among them the fact that it does not consider the normative aspects of copyright law. It is, for example, unclear whether true positive outcomes (i.e., takedown of illegal content) or under-enforcement (i.e., false-negative) will always have a negative effect on access to culture. Furthermore, this model does not account for the uncertainty associated with the margin of discretion that platforms may have when designing their content moderation terms and conditions or other means such as licensing. This normative "upstream" dimension of access to culture challenges the assumption on which this present model is on based and regards the copyright balance struck in the existing legal framework as such. It is, however, outside the scope of this article.

4. How much error is acceptable de lege lata?

Now, based on the above, the principal question should be: what error rate is acceptable under the legislative framework (and thus, society)? This question of the quality of content moderation –or the question of error– is (implicitly) addressed in several dimensions in the legislative framework.

The Digital Services Act[29], a horizontal (i.e., not-copyright-specific) framework that will apply to all intermediary service providers, for example, addresses content moderation error rates in the context of several provisions. In relation to *voluntary* measures by online platforms[30] to ensure the unavailability of illegal (in our context copyright-infringing[31]) content, recital 26

[26] Seen in isolation and outside the broader copyright context, a false negative could be considered to have a positive impact on access to culture if focusing exclusively on the accessibility dimension, even if it is harmful in other ways.

[27] If one accepts the premise of the incentive doctrine of copyright, one could for example ask whether the wide availability of copyright-infringing content, i.e. under-enforcement, would lead to a reduction in the production of copyright-protected works.

[28] That is because technically it enforces the –presumed– optimal regulation.

[29] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277, 27.10.2022, p. 1–102.

[30] Also referred to as good Samaritan actions, see Article 7 DSA.

[31] On the applicability of the horizontal DSA vis-à-vis the sector-specific regulation of online content sharing service providers in the CDSM Directive, see, JP Quintais and SF Schwemer, The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright? *European Journal of Risk Regulation, 13*(2), 191-217.

DSA states that automation technology must be "sufficiently reliable to limit to the maximum extent possible the rate of errors". The recital refrains from further specifying what would be considered "sufficiently reliable" but note the superlative in relation to the limitation of error. At first glance, both parameters (i.e., "sufficiently" and "maximum") seem somewhat self-contradictory. In another content moderation context, Article 17 of the CDSM Directive, further explored below, for example, requires certain measures to be "in accordance with high industry standards of professional diligence". Recital 26 DSA does not put forward a requirement of a high industry standard but in lieu of other benchmarks, a high industry standard could be the relevant concept for understanding the "maximum extent possible".

In yet another context in relation to transparency reporting, Article 15(1)(e) DSA obliges intermediary service providers[32] to include in their transparency reporting information on "any use made of automated means for the purpose of content moderation, a qualitative description, a specification of the precise purposes, indicators of the *accuracy and the possible rate of error* of the automated means used in fulfilling those purposes, and any safeguards applied".[33] It is not further specified how this *possible* rate of error should be calculated.[34] Furthermore, it can be assumed that once a platforms discovers erroneous decisions in the preparation of reporting, it is likely that such error would be corrected. The number of complaints received through an internal complaint-handling (cf. Article 15(1) lit. d DSA) can provide only one starting point, since it is unlikely that all wrong content moderation decisions necessarily lead to such a complaint.[35] In any case, both examples underline the crucial role of error in content moderation. They, however, also imply, that error rates are not (and need not) equal zero. Importantly, the DSA does not differentiate between type-I (false-positive) or type-II errors (false-negative).

---

doi:10.1017/err.2022.1; A Peukert et al, "European Copyright Society – Comment on Copyright and the Digital Services Act Proposal" (European Copyright Society 2022) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4016208> (last accessed 8 February 2022); E Rosati, "The Digital Services Act and Copyright Enforcement: The Case of Article 17 of the DSM Directive" in *Unravelling the Digital Services Act Package* (Strasbourg, European Audiovisual Observatory 2021).

[32] Which includes not only (very large) online platforms but notably also "regularly" hosting service providers and even mere conduit or caching providers; on content moderation outside the platform see, e.g., SF Schwemer, "Location, location, location! Copyright content moderation at non-content layers" in E Rosati, *The Routledge Handbook of EU Copyright Law* (Routledge, 2021), 378–395.

[33] Emphasis added.

[34] In the context of very large online platforms Article 42 (2) lit. c DSA adds that these indicators need to be broken down by each official EU language. This is of interest, for example, with regards to potential biases or differences in accuracy between certain language markets.

[35] Furthermore, an internal complaint mechanism is according to Article 20 DSA only required by online platforms, however, not by those intermediary service providers that only provide hosting, caching or mere conduit.

Another starting point in the DSA relates to content moderation activities in light of fundamental rights.[36] All intermediary service providers, including online platforms are according to Article 14(1) DSA required to inform users in their terms and conditions of inter alia of "measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review". In the context of error, Article 14(4) DSA is of special interest. It obliges intermediary service providers to "act in a diligent, objective and proportionate manner" and "with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter" when *applying and enforcing* their terms and conditions. Its exact extent remains vague but this implies a connection between error rates and fundamental rights. Even more relevant, however, only in the context of very-large online platforms (VLOPs) such as Youtube or Instagram, i.e. those that have more than 45 million active users.[37] Under certain circumstances, those platforms will be required to perform a risk assessment and risk mitigation.[38] Under Article 35(1) DSA, VLOPs would be required to mitigate risks, e.g., by "adapting content moderation processes, including the speed and *quality of processing notices* related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision-making processes and dedicated resources for content moderation"[39] (lit. c). As can be seen, copyright is not mentioned as a specific area of attention. This, however, does not mean that such adaption of content moderation would not be relevant in a copyright context.[40]

The copyright-specific framework for online content sharing service providers (OCSSP), too, can be understood as addressing decision quality. By comparison, however, it seems to have a somewhat more idealistic starting point: With regards to content moderation by OCSSPs like YouTube, Twitter, Pornhub or similar, Article 17 CDSM Directive appears to have a stunningly clear and equally surprising answer by the legislator: taking the text of the Article at face value, it seems that the acceptable rate of error is very close to "zero".

Firstly, Article 17(4) lit. b CDSM Directive requires OCSSPs' best efforts to ensure the unavailability of specific works in accordance with high industry standards of professional

[36] See SF Schwemer, "Digital Services Act: A reform of the e-Commerce Directive and much more" (2022).

[37] cf. Article 33(1) DSA.

[38] Cf. Articles 34 and 35 DSA.

[39] Emphasis added.

[40] On the interplay between the DSA and CDSM Directive, see Quintais, J., & Schwemer, S. (2022). The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright? European Journal of Risk Regulation, 13(2), 191-217. doi:10.1017/err.2022.1.

diligence, as noted above.[41] Article 17(7) CDSM Directive states that the cooperation between OCSSPs 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." Since this cooperation has direct influence on a platforms' content moderation practices, the question is what standard exactly is set by the requirement that lawful uses of copyright-protected works may not be prevented. Read alone, Article 17(7) CDSM Directive could be interpreted in a way that merely prohibits systematic over-enforcement. Read in conjunction with Article 17(9) para. 3 CDSM Directive, however, it seems that the standard might be stricter than that: this provision notes in a more straight-forward fashion that the Directive "shall in no way affect legitimate uses, such as uses under exceptions or limitations provided for in Union law (...)". On the one hand, it could be argued that Article 17(9) paras. 1 and 2 CDSM Directive merely deals with complaint and redress mechanisms and that therefore also para. 3 only relates to the ex post mitigation of errors. On the other hand, however, the very wording of Article 17(9) para. 3 CDSM Directive points to a more holistic standard. This reading is also supported by the requirement that it shall "shall not lead to any identification of individual users nor to the processing of personal data" unless in compliance with the GDPR. It is unconvincing to argue that such data protection consideration should only apply to redress mechanisms but not the content moderation decision in the first place.

Furthermore, as mentioned above, the provision in Article 17(7) CDSM Directive also harmonises the mandatory limitations and exceptions for quotation, criticism, review and the use for the purpose of caricature, parody or pastiche. Since this in practice is unfeasible – whether involving automation or not–, redress mechanisms are put in place to mitigate errors. In this context, the European Commission's Guidance on Article 17, however notes that "to restore legitimate content ex post (...) once it has been removed or disabled" would "not be enough for the transposition and application of Article 17(7)".[42] Therefore, "automated blocking, i.e. preventing the upload by the use of technology, should in principle be limited to manifestly infringing uploads".[43]

The question of copyright content moderation quality and error in the context of this provision was also touched upon by Advocate-General Øe in his opinion in Case C-401/19, *Poland v Parliament and Council*. Øe points out that Article 17(7) CDSM Directive "does not mean that the mechanisms which lead to a negligible number of cases of 'false positives' are automatically

[41] In other words, it sets the standard for false-negatives.

[42] See European Commission, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, Brussels, 4.6.2021 COM(2021) 288 final, p. 20.

[43] See European Commission, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, Brussels, 4.6.2021 COM(2021) 288 final, p. 20.

contrary to that provision".[44] Yet, the AG notes that error rates "should be as low as possible".[45] Therefore, AG Øe argues that in situations where the current technological state of the art for automatic filtering tools is not sufficiently advanced to prevent a significant false-positive rate, the use of such tool should be precluded.[46] In conclusion, Article 17 CDSM Directive contains both indicators as to the acceptable error rate for false-negatives and false-positives. It is, however, noteworthy that over-blocking –i.e. a higher false-positive rate– according to AG Øe may be justified in certain cases in in light of "effectiveness of the protection of the rights of rightholders" in light of the CJEU's case law.[47] Thus, the acceptable error rate for false-positives does not necessary correspond to that of false-negatives in copyright content moderation by OCSSPs.

5. But when does error rate say something about decision quality?

In any case, however, the issue of error rates in all above examples can only consist of a concrete contextual analysis in order to provide meaningful information on decision quality.

First, it is necessary to have information available on how error rates are calculated, i.e. how much error is present. As noted above, the number of complaints received through an internal complaint-handling (cf. Article 15(1) lit. d DSA) can provide only one starting point, since it is unlikely that all wrong content moderation decisions necessarily lead to a complaint. There may be instances where a user simply decides not to appeal the decision. The question is then what other aspects could be taken into consideration for identifying error. One way could be to work with an estimate of what percentage of (wrong) decisions will not be overturned, to perform random (statistically significant) samples where the legality of content is assessed, i.e. the course of pre-flagging.

Second, whereas the mere percentage of error in content moderation might provide a first insight into whether the moderation activities are somewhat precise, it is per default only a superficial metric. In the example of large-scale content moderation, a low percentage of error (error rate) would still constitute a high number of actual "wrong" content moderation decisions. Consider, the following: if bots would post millions of evidently infringing copyright-protected works on a platform, which are automatically removed, the error rate for uploads by

[44] Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 214.

[45] Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 214.

[46] Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 214.

[47] See Case C-401/19, Opinion of Advocate General Saugmandsgaard Øe delivered on 15 July 2021, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para. 183.

users which constitute transformative uses, would be lower than for a platform where less evidently infringing copyright-protected works are posted. A second factor should relate to the actual volume of content moderation decisions taken.

It is also questionable whether error rate is a one-size-fits all or whether there are varying acceptable error rates. A third factor should therefore relate to the "harm" caused by the error, i.e. the wrong content moderation decision, and whether and to what extent such harm can be mitigated ex post. In the case of hate speech or child sexual abuse material, for example, it might be societally more acceptable to have a higher false-positive rate (over-removal), simply because of the seriousness of the potential infringement and the difficulty to mitigate the harm for the object of protection. In any case, the negative effect of a delayed posting due to pre-flagging might be relatively small. In as far economic rights are concerned (including copyright-protected content), on the other hand, consideration should be given to the fact that economic damage can be remedied.[48] Thus, it could be assumed that the acceptable rate of error in light of fundamental rights is lower.

5. Conclusions

This present exercise of oversimplification underlines the importance of analyzing and differentiating the different strategies aiming at mitigating issues of automated (copyright) content moderation. The simplified model introduced above allows us to compartmentalize the specific issues of copyright content moderation by online platforms: in this model, the focus is consequently on the "downstream" issue of mitigating of (type-I and type-II) errors in content moderation. In this downstream mitigation, both ex ante obligations as well as ex post procedural redress mechanisms become relevant. As mentioned above, this paper does not address the "upstream" question of the "right" balance[49], e.g., by introducing or interpreting a broad liability exemption, thereby minimizing platforms' incentives to over-enforce.[50] In other words, this perspective just took the "existing" framework as starting point; the more perfect automatic enforcement, the better. But it is (and will remain) unrealistic to achieve "0" error. What becomes clear though is that ex post mechanisms such redress mechanisms as well as transparency mechanisms do not have a direct effect on error on the initial content moderation decision. Concomitantly, not all are equally fit for minimizing error (or, increase quality of

[48] Note in this context also the relevance of Article 17(1) CDSM Directive.

[49] see above point 2 on benchmarks

[50] See also Kuczerawy, Aleksandra: *Remedying Overremoval: The Three-Tiered Approach of the DSA, VerfBlog,* 2022/11/03, https://verfassungsblog.de/remedying-overremoval/

decision-making). The following table lists measures and their expected effect on error in the (automated) content moderation.

Measure	**Effect on error in decision making**
Changing substantive copyright rules [not addressed in this paper]	
Introduce broad liability exemptions	Minimising error because of decreased complexity and concomitantly fewer grey zones
Increase / reduce copyright protection	Minimising error because less need for content moderation in the first place
Other measures	
Introduce presumption rules	Ex ante: Minimising error because of legal fiction
Better technology	Ex ante: Minimising errors; also: minimizing need for copyright experts
Redress mechanism	Ex post: No effect on error in CoMo
Transparency mechanism	Ex post: No effect on error in CoMo

Table 1: *Measures and anticipated effect on error in content moderation*

A further simplification of this perspective lays in the fact that content moderation consists of activities that go beyond takedown, namely measures that "affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or the recipients' ability to provide that information, such as the termination or suspension of a recipient's account".[51] In the context of copyright *infringement*, however, takedown may be the only appropriate response to the exclusive right of right holders. In order to minimise error, however, a mix of moderation techniques might be able to strike a more appropriate balance.[52] If the legal status of an uploaded work is uncertain, its visibility could for example be lowered until a final decision is taken. Furthermore, it is not necessarily a case of either full automation or manual moderation. The importance of human involvement (in form of review) in the automated content moderation process (and not merely ex post) is exemplified by the case YouTube, which reduced its workforce in response to COVID-19. This reduction in human reviewers, according to YouTube means it removes "more content that may not be violative of our policies".[53]

Arguably, the quality of decision making should be a central perspective when regulating how copyright-protected material is enforced. With regards to copyright content, it has been argued that Article 17 CDSM Directive has tipped the balance in favour of rightsholders and OCSSPs

[51] Article 3 DSA.

[52] However, difficult to reconcile with Article 17(1) DSM Directive; see also AG Øe and CJEU in its caselaw.

[53] See Google, "YouTube Community Guidelines enforcement", https://transparencyreport.google.com/youtube-policy/removals?hl=en (last accessed 15 November 2022)

may be incentivised to over-enforce.[54] Also outside the copyright-specific special regime of Article 17 CDSM Directive, the Digital Services Act's rules, applying from 17 February 2024[55], further incentivises online platforms (and in fact all intermediary service providers[56]) to conduct voluntary own-initiative investigations including to "take other measures aimed at detecting, identifying and removing, or disabling access to, illegal content" in Article 7 DSA.[57]

Ex post mitigations (redress mechanisms) or transparency do not reduce error but merely mitigate effect of error, which may vary from instance to instance. Thus, it seems that "users" bear the larger risk of error/low decision quality. In conclusion therefore it is necessary to focus on *decision quality* as distinct aspect in addition to ex post mitigation mechanism.

[54] Schwemer, S. F., & Schovsbo, J. (2019). What is Left of User Rights?–Algorithmic Copyright Enforcement and Free Speech in the Light of the Article 17 Regime. Intellectual Property Law and Human Rights, 4th edition (Wolters Kluwer, 2020), 569-589. Just as Article 17(1) CDSM Directive establishes direct liability for online platforms vis-à-vis rights holders, it could for example be argued that there needs to be a liability basis for overenforcement. In this context, Article 54 DSA might provide an interesting train of thought, which establishes that users "shall have right to seek, in accordance with Union and national law, compensation from providers of intermediary services, in respect of any damage or loss suffered due to an infringement by those providers of their obligations under this Regulation.

[55] See Article 93(2) DSA.

[56] On the issue of copyright content moderation on the DNS layer, see, e.g., Schwemer, S. F. (2021). Location, location, location! Copyright content moderation at non-content layers. In *The Routledge Handbook of EU Copyright Law* (pp. 378-395). Routledge. See also the recent study prepared for WIPO by Dean Marks and Jan Bernd Nordemann, *The Role of the Domain Name System and Its Operators in Online Copyright Enforcement*, 2022, https://www.wipo.int/edocs/mdocs/enforcement/en/wipo_ace_15/wipo_ace_15_7-executive_summary1.pdf

[57] See, however, the arguments above with regards to terms and conditions of intermediary service providers.

Kluwer Copyright Blog (http://copyrightblog.kluweriplaw.com/)

COPYRIGHT (HTTP://COPYRIGHTBLOG.KLUWERIPLAW.COM/CATEGORY/COPYRIGHT/), EUROPEAN UNION (HTTP://COPYRIGHTBLOG.KLUWERIPLAW.COM/CATEGORY/JURISDICTION-2/EUROPEAN-UNION/)

Algorithmic propagation: do property rights in data increase bias in content moderation? Part I (http://copyrightblog.kluweriplaw.com/2022/06/08/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/)

Thomas Margoni (http://copyrightblog.kluweriplaw.com/author/thomas-margoni/) (Centre for IT and IP Law (CiTiP), Faculty of Law, KU Leuven (https://www.kuleuven.be/english/kuleuven/index.html)), João Pedro Quintais (http://copyrightblog.kluweriplaw.com/author/joao-pedro-quintais/) (Institute for Information Law (IViR) (https://www.ivir.nl/)), and Sebastian Felix Schwemer (http://copyrightblog.kluweriplaw.com/author/sschwemer1/) (Københavns Universitet Centre for Information and Innovation Law (CIIR), University of Copenhagen (https://jura.ku.dk/ciir/english/staff/?pure=en/persons/389492)) / June 8, 2022 (http://copyrightblog.kluweriplaw.com/2022/06/08/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/) / Leave a comment (http://copyrightblog.kluweriplaw.com/2022/06/08/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/#respond)

Introduction

This two-part blog post offers a reflection on the topic of content moderation and bias mitigation measures in copyright law. It explores the possible links between conditional data access regimes and content moderation performed through data-intensive technologies such as fingerprinting and, within the realm of artificial intelligence (**AI**), machine learning (**ML**) algorithms. More specifically, this post explores whether current EU copyright rules may have the effect of favoring the propagation of bias present in input data to the algorithmic tools employed for content moderation and what kind of measures could be adopted to mitigate this effect.

Our analysis explores the dynamic of "bias propagation" in relation to the obligations stemming from Article 17 **CDSM Directive**

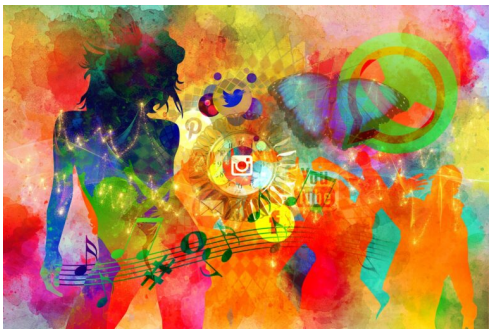


Image by **Gerd Altmann** (https://pixabay.com/da/users/geralt-9301/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=3998128) from (https://pixabay.com/da/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=3998128)

(https://eur-lex.europa.eu/eli/dir/2019/790/oj). In simple terms, Article 17 incentivizes certain platforms to filter content uploaded by users to comply with their "best efforts" obligations to deploy preventive measures against infringing content. Prior to the introduction of this legal regime, however, some platforms already "voluntarily" relied on similar automated content moderation (e.g., YouTube's' ContentID). At the current state of technology, filtering appears to be done mainly through matching and fingerprinting. However, these tools are incapable of assessing contextual uses (see, e.g. **here (https://www.communia-association.org/2019/12/03/article-17-stakeholder-dialogue-day-3-filters-not-meet-requirements-directive/)**). Therefore, they are not suitable to ensure the required protection of freedom of expression-based exceptions like parody, criticism and review, as required by Article 17(7). Accordingly, more sophisticated tools seem necessary to enable preventive measures while respecting users' rights and freedoms, as recently confirmed by the CJEU in case **C-401/19 (https://curia.europa.eu/juris/liste.jsf?language=en&num=C-401/19)** (see **here (https://verfassungsblog.de/filters-poland/)**). This suggests that ML algorithms may increasingly be employed for copyright content moderation given their alleged superiority in identifying (understanding?) contextual uses.

Against this background, a crucial question emerges for the future of online content moderation and fundamental rights in the EU: what happens when these tools are based on "biased" datasets? More specifically, if it is plausible that any bias, errors or inaccuracies present in the original datasets be carried over in some form onto the filtering tools developed on those data: (1) How do property rights in data influence this "bias carry-over effect"? and (2) what measure (transparency, verifiability, replicability, etc.) can and should be adopted to mitigate this undesirable effect in copyright content moderation in order to ensure an effective protection of fundamental rights?

Part I of this post briefly discusses the concept of bias and examines the role of property rights in data and factual information, with a focus on copyright. Part II explores the potential of property rights to increase bias in content moderation by looking at the topic from the perspective of Article 17 CDSM Directive.

Bias, data, and property rights

Looking at the common meaning of the word, *bias* may be defined as the tendency to favor or dislike a person or thing, especially as a result of a preconceived and often unfairly formed opinion that affects behavior. In scientific literature, bias is usually defined in relation to a specific field, area or group (e.g., cognitive bias, algorithmic bias, confirmation bias, gender bias, etc.). **Common denominators (https://onlinelibrary.wiley.com/doi/full/10.1002/9781119125563.evpsych241)** are the presence of an error that is due to systematic imprecisions, or to deviations from standards in judgement which lead to unfair, inaccurate, illogical or discriminatory conclusions. Wikipedia **lists more than 200 types of bias (https://en.wikipedia.org/wiki/List_of_cognitive_biases)**, just within the field of cognitive sciences.

In the field of AI, bias may be present at various stages of the development of an application including the *design* of the algorithms, the *designers* of the algorithm, the identification and sampling of the learning information and the curation, **annotation (https://arxiv.org/abs/2007.14886)**, and verification of the input data (see, e.g., **here (https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai)** and **here (https://www.turing.ac.uk/research/interest-groups/fairness-transparency-privacy)**). Bias in AI may be particularly treacherous due to the specific technical and societal characteristics that this technology has acquired. On the one hand, it is often said that **AI operates as a black box (https://cyber.harvard.edu/story/2019-07/intellectual-debt-great-power-comes-great-ignorance)** in the sense that it is not possible for humans to really *understand* the learning process that happens within the algorithm and therefore to detect bias and inaccuracy by employing traditional approaches. On the other hand, given the widespread adoption of AI in our society and the fact that in an increasing number of cases these applications rely on the same or similar pre-trained models (i.e. already processed input data, such as **BERT (https://en.wikipedia.org/wiki/BERT_(language_model))**, **GPT-2 (https://en.wikipedia.org/wiki/GPT-2)**, **GPT-3 (https://en.wikipedia.org/wiki/GPT-3)** in the **Natural Language Processing (NLP) (https://en.wikipedia.org/wiki/Natural_language_processing) secto (https://en.wikipedia.org/wiki/Natural_language_processing)**r), any bias present in these pre-processed datasets may be transferred to all implementations of that system giving rise not only to some sort of "bias carryover" but also "bias multiplication" effect.

Brad Spitz (http://www.realex.fr/)
REALEX (http://www.realex.fr/)

Jeremy Blum (http://www.bristows.com/o
people/jeremy_blum)
Bristows LLP (http://www.bristows.com/)

Gianluca Campus
University of Milan (http://www.unimi.it/ENG/)

P. Bernt Hugenholtz
(http://www.ivir.nl/medewerkerpagina/hug
Institute for Information Law (IViR) (http://www.ivir.nl/

Martin Husovec
(http://www.tilburguniversity.edu/webwijs/
London School of Economics (https://www.lse.ac.uk
staff/martin-husovec)

Bernd Justin Jütte
(http://www.nottingham.ac.uk)
University College Dublin (https://www.ucd.ie/)

Paul Keller
(https://www.uva.nl/en/profile/k/e/p.keller/
Institute for Information Law (IViR) (http://www.uva.n

Rita Matulionyte
Macquarie Law School

Jan Bernd Nordemann
(http://nordemann.de/team/#prof-dr-jan-bernd-nordemann)
NORDEMANN (http://nordemann.de/)

Giulia Priora
NOVA School of Law Lisbon (https://novalaw.unl.pt/

The focus of our analysis is on those elements of the EU copyright *acquis* that create conditions for access and reuse of non-personal data and that could therefore play a role in the propagation of bias. In doing so, it is important to highlight some key traits of the EU *acquis*. First, when we refer to *input* or *training data* (a technical category) we mean material that from a copyright law perspective could be either works of authorship, other protected subject matter or mere facts and data. Second, although mere facts and data as such are not protected by copyright, their extraction from protected subject matter (works, non-original databases, etc.) often requires authorization under EU copyright law. Third, exceptions and limitations for the extraction of information from works are narrower in the EU than in many other non-EU legal systems (see e.g., **here (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3578819)**). Fourth, as a consequence of the above, EU copyright law can be said to protect information needed as input data for AI applications to a greater extent than other legal systems (see e.g., **here (https://zenodo.org/record/5082012)**).

The mechanism by which these key traits of EU copyright law may influence bias is relatively simple. By regulating access to training data through the imposition of costs or other use conditions, property rights may create unanticipated incentives that drive AI developers towards "more available", "cheaper", "less risky" or as it has been called "low friction" data (**Levendowski 2018 (https://www.levendowski.net/copyright-ai-bias)**), which incidentally are more easily found outside the EU. However, whether these data represent the optimal choice in terms of quality, accuracy, and representativeness or whether they are simply chosen to reduce the economic, informational, or legal certainty costs, regardless of their suitability for the task, is far less clear.

Attempting a first categorisation of "data types", the following five scenarios may be logically derived from an application of the EU *acquis* to the specific field of input or training data.

Scenario (1): Public Domain

The public domain is arguably the "cheapest" source of data. The main problem with this category is that to enter the public domain underlying works are on average at least 70 years old, and often much older. Accordingly, there is a risk that data extracted from this source may likewise convey outdated, disproved or surpassed information.

Scenario (2): Open Licenses

Open Licenses are likely the closest scenario to the Public Domain in terms of "costs". Tools such as **Creative Commons (https://creativecommons.org/about/cclicenses/)** licenses (e.g., Wikipedia), GNU General Public License (e.g., Free and Open Source Software) and open government licenses (e.g., **reuse of Commission documents (https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF)**) play a major role here. However, **openly licensed information is not necessarily immune from bias (https://en.wikipedia.org/wiki/Ideological_bias_on_Wikipedia)** just for being openly available, even though openness certainly allows for greater transparency and thus closer scrutiny of the underlying dataset.

Scenario (3): Exceptions and Limitations

Next in terms of "costs" is information accessible thanks to exceptions and limitations (E&Ls) allowing data analytics or computational uses (e.g., **text and data mining or "TDM" (https://zenodo.org/record/5082012)**). As mentioned above, under EU law, exceptions for TDM purposes or use (in Articles 3 and 4 CDSM Directive) are arguably narrower than in other jurisdictions, such as the US, Canada or Japan. Accordingly, the training of AI systems in the EU is conceivably more "expensive" which, in turn, operates as an incentive for the training of AI systems in "cheaper" jurisdictions or to import from those jurisdictions "pre-trained" models. The deeper implications of these dynamics are far from clear. For instance, will a ML algorithm trained in the US learn the meaning of a given concept, e.g., "parody", within the US socio-cultural context and then apply that meaning in the EU when asked to perform a function such as "filter copyright infringing videos"? Is this technically plausible? If yes, what impact could this have in an online regulatory environment that is increasingly relying on privatized algorithmic enforcement?

Scenario (4): Third party content

When the above scenarios are not applicable or deemed not adequate by virtue for instance of a cost-benefit analysis, access to the information needed for training purposes may follow "traditional" contractual arrangements. Two main situations may be envisaged. Firstly, the required

Felix Reda (https://felixreda.eu/en/)
GFF (Society for Civil Rights)
(https://freiheitsrechte.org/team/)

(htt
red

Rainer Schultes (http://www.geistwert.at)
Geistwert (http://www.geistwert.at)

(htt
sch

Tatiana Synodinou
(http://www.ucy.ac.cy/~synodint.aspx)
University of Cyprus

(htt
syn

Alina Trapova
The University of Nottingham
(https://www.nottingham.ac.uk/law/people/alina.trap

(htt
trap

VIEW POSTS ON:

Australia (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/australia/) Austria (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/austria/) Authorship (http://copyrightblog.kluweriplaw.com/
Belgium (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/belgium/) Brexit (http://copyrightblog.kluweriplaw.com/category/brexit/
Case Law (http://copyrightblog.k law/) CJEU (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/eu/) Collective management (http://copyrightblog.kluweriplaw.com/category/collective-management/) Communication (right of) (http://copyrightblog.kluweriplaw.com/category/communication-right-of/) Conference (http://copyrightblog.kluweriplaw.com/category/conference/)
Copyright (http://copyrightblog.kluw
Copyright Authority/Board (http://copyrightblog.kluweriplaw.com/category/copyright-authorityboard/) Database right (http://copyrightblog.kluweriplaw.com/category/database-right/) Digital Single Market (http://copyrightblog.kluweriplaw.com/category/digital-single-market/) Distribution (right of) (http://copyrightblog.kluweriplaw.com/category/distribution-right-of/) Enforcement (http://copyrightblog.kluweripla

training information may already be hosted by the entity interested in the training, as it is arguably the case for all major Internet platforms hosting large amounts of third-party content which is licensed in a way that usually allows the platform to develop its own services. Secondly, the entity interested in the training needs to acquire access to third party content often hosted in large commercial databases, examples of which may be scientific commercial databases offering TDM licenses to commercial or academic users. This scenario seems to favor the strengthening of the dominant position of large platforms which will not need to pay the extra price to train on third party content they already host, a price that conversely other, usually smaller, players who do not own these large databases will need to pay thereby further reducing their competitiveness in this market.

A third case may be conceived where right holders are compelled to contribute "relevant and necessary information" about their content to qualifying platforms to develop dedicated filtering tools, as provided in Article 17(4) CDSM Directive. This is an interesting development, which however requires a dedicated analysis. Among among other aspect, such analysis will need to distinguish among the technologies adopted *i.e.,* fingerprinting and hashing, or ML algorithms, and assess the specific role of the provided data.

Scenario (5): A risk-benefit analysis leading to opacity

Finally, it is at least plausible that EU-based firms developing AI systems decide to perform training activities regardless of all the above considerations, knowing that it is unlikely they will be "discovered" engaging in potentially copyright-infringing activity. Once trained, especially employing modern algorithms that reach deeper levels of abstraction, it will be difficult to "reverse engineer" the models, i.e., to go back from the model to the training data and therefore to demonstrate infringement. This scenario – the plausibility of which should be further tested – may lead to an increased opacity in the training process: firms will have strong incentives not to disclose details about the training sources to avoid acknowledging their own infringement of third-party rights.

In conclusion, the identified "costs" associated with the uneasy case of property rights in data have the potential to favor the use of outdated and lower quality data sources, or to disincentivize transparency and accountability in the training process, all ideal conditions for bias and errors. This will push AI developers in high-cost data legal systems, such as the EU, to either outsource data analytics to non-EU legal systems – with the unexplored consequences on fundamental rights and cultural dynamic sketched above – or to be even more opaque in their data policies with similar negative consequences on fundamental rights and freedoms.

Following the above discussion of the concept of bias and the role of property rights in data and factual information, with a focus on copyright, the second part of this post will explore the potential for property rights to increase bias in content moderation. We will do so by looking at the topic from the perspective of Article 17 CDSM Directive.

Acknowledgments: This research is part of the following projects. All authors: the ***reCreating Europe*** **(https://www.recreating.eu/)***project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870626. João Pedro Quintais: VENI Project "Responsible Algorithms: How to Safeguard Freedom of Expression Online" funded by the Dutch Research Council (grant number: VI.Veni.201R.036).*

This blog post is based on the EPIP2021 roundtable organised in Madrid (September 8-10, 2021). The authors are grateful to Prof. Niva Elkin-Koren and to Dr. Irene Roche-Laguna for their participation and for their insightful perspectives and suggestions which have been helpful in developing this analysis. The blog post only reflects the view of the authors and any errors remain our own.

2/) Estonia (http://copyrightblog.kluweriplaw.com/category/ju 2/estonia/) European Union (http://copyrightblog.k 2/european-union/) Exhaus (http://copyrightblog.kluweriplaw.com/category/ex France (http://copyrightblog.kluweriplaw.c 2/france/) Germany (http://copyrightblog.kluweriplaw.c 2/germany/) Infringement (http://copyrightblog.klu Italy (http://copyrightblog.kluweriplaw.com/cate 2/italy/) Jurisdiction (http://copyrightblog.kluweri 2/) Landmark Cases (http://copyrightblog.kluweriplaw.com/category/ju 2/poland/) Remedies (http://copyrightblog.kluweripla Remuneration (equitable) (http://copyrightblog.kluweriplaw.co equitable/) Reproduction (right of) (http://copyrightblog.kluweriplaw.com/ right-of/) Software (http://copyrightblog.kluweriplaw.com/category/softwa Spain (http://copyrightblog.kluweriplaw.com/cate 2/spain/) Subject matter (copyrightable) (http://copyrightblog.kluweriplaw.com/ca matter-copyrightable/) Sweden (http://copyrightblog.kluweriplaw.com/category/ju 2/sweden/) Transfer (of right) (http://copyrightblog.kluweriplaw.com/category/transfer-of-

COPYRIGHT (HTTP://COPYRIGHTBLOG.KLUWERIPLAW.COM/CATEGORY/COPYRIGHT/), EUROPEAN UNION (HTTP://COPYRIGHTBLOG.KLUWERIPLAW.COM/CATEGORY/JURISDICTION-2/EUROPEAN-UNION/)

Algorithmic propagation: do property rights in data increase bias in content moderation? – Part II (http://copyrightblog.kluweriplaw.com/2022/06/09/algori…-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-ii/)

Thomas Margoni (http://copyrightblog.kluweriplaw.com/author/thomas-margoni/) (Centre for IT and IP Law (CiTiP), Faculty of Law, KU Leuven (https://www.kuleuven.be/english/kuleuven/index.html)), João Pedro Quintais (http://copyrightblog.kluweriplaw.com/author/joao-pedro-quintais/) (Institute for Information Law (IViR) (https://www.ivir.nl/)), and Sebastian Felix Schwemer (http://copyrightblog.kluweriplaw.com/author/sschwemer1/) (Københavns Universitet Centre for Information and Innovation Law (CIIR), University of Copenhagen (https://jura.ku.dk/ciir/english/staff/?pure=en/persons/389492)) / June 9, 2022 (http://copyrightblog.kluweriplaw.com/2022/06/09/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-ii/) / Leave a comment (http://copyrightblog.kluweriplaw.com/2022/06/09/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-ii/#respond)

This is the second installment of a reflection on the topic of content moderation and bias mitigation measures in copyright law. The **first part of this post (http://copyrightblog.kluweriplaw.com/?p=12796&preview_id=12796&preview_nonce=2eb12b137f&post_format=standard&_thumbnail_id=-1&preview=true)**

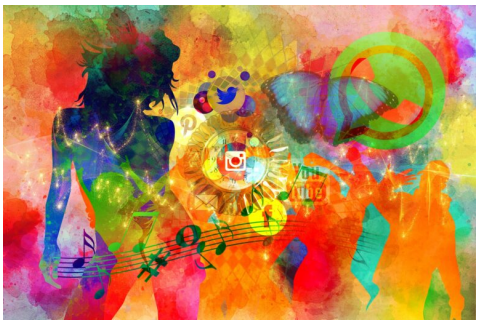


Image by **Gerd Altmann (https://pixabay.com/da/users/geralt-9301/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=3998128)** from **(https://pixabay.com/da/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=3998128)**

briefly discussed the concept of bias and examined the role of property rights in data and factual information, with a focus on copyright. This second part explores the potential of property rights to increase bias in content moderation by looking at the topic from the perspective of Article 17 CDSM Directive.

Article 17, content moderation tools, and the Commission's Guidance

Article 17 CDSM Directive regulates online content-sharing service providers (OCSSPs) through a complex set of rules. The provision states that OCSSPs carry out acts of communication to the public when they give access to works or subject matter uploaded by their users, making them directly liable for their users' uploads. At the same time, the provision includes a liability exemption mechanism in paragraph (4), as well as several mitigation measures and safeguards in paragraphs (5) to (9).

The liability exemption mechanism in Article 17(4) encompasses a series of *cumulative* obligations of "best efforts" on OCSSPs to: (a) obtain an authorisation; (b) ensure unavailability of specific protected content provided they have received "relevant and necessary information" from right holders; and (c) put in place notice and take down and notice and stay down mechanisms, provided they have received a "sufficiently substantiated notice" from right holders.

In interpreting these provisions, the **Commission's Guidance (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021DC0288)** (COM/2021/288 final) states that information is considered "relevant" if it is at least "accurate about the rights ownership of the particular work or subject matter in question". The consideration of whether it is "necessary" is trickier, and it will vary depending on the technical solutions deployed by OCSSPs; in any case, such information must allow for the effective application of the providers' solutions, where they are used (e.g., "fingerprinting" and "metadata-based solutions").

The Guidance further states that measures deployed by OCSSPs must follow "high industry standards of professional diligence", to be assessed especially against "available industry practices on the market" at the time, including technological solutions. When discussing current market practices that emerged from the **Stakeholder Dialogues (https://digital-strategy.ec.europa.eu/en/policies/stakeholder-dialogue-copyright)**, the Guidance highlights content recognition based on fingerprinting as the main example, although recognising that is not the market standard for smaller OCSSPs. Other technologies identified include hashing, watermarking, use of metadata and keyword search; these can also be used in combination. Such technologies are sometimes developed in-house (e.g., YouTube's **ContentID (https://support.google.com/youtube/answer/2797370?hl=en)** or Facebook's **Rights Manager (https://rightsmanager.fb.com/)**), and other times acquired from third parties (e.g., from **Audible Magic (https://www.audiblemagic.com/)** or **Pex (https://pex.com/)**). Crucially, the Guidance is a non-binding document published in the shadow of the action for annulment in Case **C-401/19 (https://curia.europa.eu/juris/liste.jsf?language=en&num=C-401/19)**, and which itself recognizes that it might need revising in light of that judgment.

The CJEU's interpretation of Article 17 and content filtering (C-401/19)

In its recent Grand Chamber judgment in **Case C-401/19 (https://curia.europa.eu/juris/document/document.jsf?text=&docid=258261&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=10742191)** (discussed **here (http://copyrightblog.kluweriplaw.com/2022/04/26/article-17-survives-but-freedom-of-expression-safeguards-are-key-c-401-19-poland-v-parliament-and-council/)**, **here (http://copyrightblog.kluweriplaw.com/2022/04/28/cjeu-upholds-article-17-but-not-in-the-form-most-member-states-imagined/)**, **here (http://copyrightblog.kluweriplaw.com/2022/06/01/the-meaning-of-additional-in-the-poland-ruling-of-the-court-of-justice-double-safeguards-ex-ante-flagging-and-ex-post-complaint-systems-are-indispensable/)**, and **here (http://copyrightblog.kluweriplaw.com/2022/06/06/constitutional-safeguards-in-the-freedom-of-expression-triangle-online-content-moderation-and-user-rights-after-the-cjeus-judgement-on-article-17-copyright-dsm-directive/)**), the Court clarified that Article 17(4)(b) does in fact require prior check of content by OCSSPs. In many cases, the only viable solution for

Brad Spitz (http://www.realex.fr/)
REALEX (http://www.realex.fr/)

Jeremy Blum (http://www.bristows.com/o
people/jeremy_blum)
Bristows LLP (http://www.bristows.com/)

Gianluca Campus
University of Milan (http://www.unimi.it/ENG/)

P. Bernt Hugenholtz
(http://www.ivir.nl/medewerkerpagina/hug
Institute for Information Law (IViR) (http://www.ivir.nl/

Martin Husovec
(http://www.tilburguniversity.edu/webwijs/
London School of Economics (https://www.lse.ac.uk
staff/martin-husovec)

Bernd Justin Jütte
(http://www.nottingham.ac.uk)
University College Dublin (https://www.ucd.ie/)

Paul Keller
(https://www.uva.nl/en/profile/k/e/p.keller/
Institute for Information Law (IViR) (http://www.uva.n

Rita Matulionyte
Macquarie Law School

Jan Bernd Nordemann
(http://nordemann.de/team/#prof-dr-jan-bernd-nordemann)
NORDEMANN (http://nordemann.de/)

Giulia Priora
NOVA School of Law Lisbon (https://novalaw.unl.pt/

platforms to moderate content is to deploy automated recognition and filtering tools, i.e., "upload filters" which constitute a justified restriction of users' freedom of expression (see **here (https://verfassungsblog.de/filters-poland/)**).

The Court advances a number of arguments why Article 17 constitutes a proportionate restriction of OCSSP's users' freedom of expression. In essence, such arguments relate to the legislative design of Article 17 and how the provision should be interpreted and implemented in light of fundamental rights. It is also noteworthy among these arguments that the Court outlines the scope of permissible filtering. For our present purposes, the key points are the following.

- First, only filtering/blocking systems that can distinguish lawful from unlawful content are compatible with the requirements of Article 17 and strike a fair balance between competing rights and interests. Despite this important statement, establishing exactly what kind of thresholds or error rates are admissible in practice remains unclear and arguably one of the key issues in this new system.

- Second, Member States must ensure that filtering measures do not prevent the exercise of user rights to upload content that consists of quotation, criticism, review, caricature, parody or pastiche under Article 17(7). It is here where we argue that AI/ML tools will probably become essential given their superiority over fingerprinting and hashing in determining contextual uses (e.g., parody). Whether this superiority is sufficient to protect users' fundamental rights remains an open question, but it is clear that the role played by bias and errors in this assessment is decisive.

Potential bias in copyright content moderation and error rates

So, what does the Court's judgment mean for our discussion on bias propagation?

First, it is clear from Article 17 and the CJEU's judgment that the only acceptable point of reference to deploy (re)upload filters is the information and/or notice provided by right holders under Article 17(4)(b) and (c). Depending on the technology employed, the information provided will play a critical role in the correct identification of the allegedly infringing material. A dedicated analysis should be carried out for each of the **two main groups (https://journals.sagepub.com/doi/full/10.1177/2053951719897945)** of content moderation approaches adopted in this field: matching (fingerprinting, hashing, metadata, etc.) and predictive analysis (AI/ML). It suffices here to say that the process of identification of the input data and its use to classify users' uploaded content is liable to embed the potential bias and errors identified in the first part of this blog. This bias may very well influence the ability to correctly identify content (false positives and/or negatives), the ability to establish acceptable thresholds (e.g., a 20% match, a 90% match), and, most importantly for the analysis here developed, the (in)ability to correctly determine contextual uses in order to safeguard users' fundamental rights as provided for in Article 17(7).

Second, such filters must be able to distinguish lawful from unlawful content without the need for an independent assessment by the providers. As major platforms have admitted during the stakeholder dialogues, current moderation tools are not capable of assessing these contextual uses. That suggests both a future push towards a more AI/ML intensive approach but also explains the current dependence of existing filters on *error thresholds* to distinguish between what content is blocked and what content stays up.

According to the above, the core question may be reformulated as a question about what type of errors or bias are legally acceptable. The Advocate General (AG) in his **Opinion (https://curia.europa.eu/juris/document/document.jsf;jsessionid=9A2E213F410CBACEB5D5E2F8EA49FA08?text=&docid=244201&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=164140)** considered it crucial to ensure that the error rate of "false positives" resulting from the deployment of content recognition tools by OCSSPs "should be as low as possible". Hence, where "it is not possible, in the current state of technology… to use an automatic filtering tool without resulting in a 'false positive' rate that is significant, the use of such a tool should… be precluded" [para 214]. In the AG's view, allowing *ex ante* filtering in such cases of 'transformative' content would "risk

Felix Reda (https://felixreda.eu/en/)
GFF (Society for Civil Rights)
(https://freiheitsrechte.org/team/)


(htt
reda

Rainer Schultes (http://www.geistwert.at)
Geistwert (http://www.geistwert.at)


(htt
sch

Tatiana Synodinou
(http://www.ucy.ac.cy/~synodint.aspx)
University of Cyprus


(htt
syn

Alina Trapova
The University of Nottingham
(https://www.nottingham.ac.uk/law/people/alina.trap


(htt
trap

VIEW POSTS ON:

Australia (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/australia/) Austria (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/austria/) Authorship (http://copyrightblog.kluweriplaw.com/
Belgium (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/belgium/) Brexit (http://copyrightblog.kluweriplaw.com/category/brexit/
Case Law (http://copyrightblog.k law/) CJEU (http://copyrightblog.kluweriplaw.com/category/jurisdiction-2/eu/) Collective management (http://copyrightblog.kluweriplaw.com/category/collective-management/) Communication (right of) (http://copyrightblog.kluweriplaw.com/category/communication-right-of/) Conference (http://copyrightblog.kluweriplaw.com/category/conference/)
Copyright (http://copyrightblog.kluw
Copyright Authority/Board (http://copyrightblog.kluweriplaw.com/category/copyright-authorityboard/) Database right (http://copyrightblog.kluweriplaw.com/category/database-right/) Digital Single Market (http://copyrightblog.kluweriplaw.com/category/digital-single-market/) Distribution (right of) (http://copyrightblog.kluweriplaw.com/category/distribution-right-of/) Enforcement (http://copyrightblog.kluweripla

causing 'irreparable' damage to freedom of expression" [para 216]. Although the Court was less detailed than the AG, it generally endorsed this approach of preventing overblocking and the risk of chilling freedom of expression by focusing on avoiding false positives.

In spite of the above, the question as to what degree of error is acceptable and how to ensure unbiased results remains largely answered. The most difficult step may well be the design and training of algorithms able to assess potentially complex copyright law questions, e.g., under what conditions a certain use should be classified as parody or criticism. On the one hand, it seems almost impossible to encapsulate the concept of parody or criticism in an error-rate percentage. On the other hand, when ML algorithms are deployed instead, it appears similarly dubious that the priority followed in selecting training data was to allow filters to replicate legally educated answers. As we have seen in Part I, the incentives currently driving this market, especially in the EU, point towards other priorities, such as cost reductions, legal certainty and in-house confidential development. Consequently, if the scenarios introduced in Part I are plausible, parties involved or affected by algorithmic content moderation will need to familiarise themselves with questions in cluding how a US model trained for parody detection may influence error rates in the EU market; or how algorithms trained on large language *corpi* or pop music would perform on smaller languages, or niche repertoires. In the light of the CJEU's judgment, questions like these remain open. Delegating market or industry dynamics to offer answers carries the risk of denying effective protection to user rights in Article 17(7).

Conclusions: looking forward

It seems that –at least *prima facie*– property rights in data may possess the unexpected effect of favoring errors and bias propagation into the trained AI. This effect seems noticeable in areas where AI is already deployed such as **in the context of content recommendation (https://www.jipitec.eu/issues/jipitec-13-1-2022/5515)**, but it may be expected to also become more relevant in algorithmic content *moderation*, especially with the increasing adoption of ML approaches.

As argued, bias propagation does not follow a unique pattern, but may develop along different lines depending on the type of technology employed and how this technology relates to input data. The general principles and the arrival point, however, appear to be similar. A reduced availability and transparency of input/training data has a negative effect on access, verification and replication of results, which, in turn, are ideal conditions for bias and errors to the detriment of users' rights.

What mechanisms or remedies can we envision to mitigate the propagation of bias from data to AI? Can they be found within the field of property rights? Several approaches may be conceived:

- Internationally, a broad call for user rights to research in AI has been proposed, which would enhance **data retrieval from protected works (https://digitalcommons.wcl.american.edu/research/48/)**. This would enhance access to relevant data (especially training data) and consequently favor a more open, transparent and verifiable data ecosystem.

- In legal systems relying on open standards such as the US, fair use has been identified by some authors as a powerful **bias mitigation device (https://www.levendowski.net/copyright-ai-bias)** (Levendowski 2018). At the same time, other authors have pointed out the risk that algorithmic enforcement systems deployed by large-scale platforms like YouTube (ContentID) or Meta (Rights Manager) may ultimately "become embedded in public behaviour and consciousness", and thus progressively shape the legal standard itself, by habituating the "participants to its own biases and so progressively altering the fair use standard it attempts to embody" (**Burke 2017 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3076139)**).

- This same risk is clear for the algorithmic content moderation systems of large-scale OCSSPs under Article 17 CDSM (some of which will qualify as "very large online platforms" under the Digital Services Act). Such large platforms have the power through their algorithms

2/) Estonia (http://copyrightblog.kluweriplaw.com/category/ju 2/estonia/) European Union (http://copyrightblog.k 2/european-union/) Exhaus (http://copyrightblog.kluweriplaw.com/category/ex France (http://copyrightblog.kluweriplaw.c 2/france/) Germany (http://copyrightblog.kluweriplaw.c 2/germany/) Infringement (http://copyrightblog.klu Italy (http://copyrightblog.kluweriplaw.com/cate 2/italy/) Jurisdiction (http://copyrightblog.kluweri 2/) Landmark Cases (http://copyrightblog.kluweriplaw.com/category/ copyright-cases/) Legislative process (http://copyrightblog.kluv process/) Liability (http://copyrightblog.kluweriplaw.com/category/ Limitations (http://copyrightblog.kluweri Making available (right of) (http://copyrightblog.kluweriplaw.co available-right-of/) Moral rights (http://copyrightblog.kluweriplaw.com/category/moral- rights/) Neighbouring rights (http://copyrightblog.kluweriplaw.com/category/ne rights/) Netherlands (http://copyrightblog.kluweriplaw.co 2/netherlands/) Originality (http://copyrightblog.kluweriplaw.c Ownership (http://copyrightblog.kluweriplaw.com/ Poland (http://copyrightblog.kluweriplaw.com/category/ju 2/poland/) Remedies (http://copyrightblog.kluweripla Remuneration (equitable) (http://copyrightblog.kluweriplaw.co equitable/) Reproduction (right of) (http://copyrightblog.kluweriplaw.com/ right-of/) Software (http://copyrightblog.kluweriplaw.com/category/softwa Spain (http://copyrightblog.kluweriplaw.com/cate 2/spain/) Subject matter (copyrightable) (http://copyrightblog.kluweriplaw.com/ca matter-copyrightable/) Sweden (http://copyrightblog.kluweriplaw.com/category/ju 2/sweden/) Transfer (of right) (http://copyrightblog.kluweriplaw.com/category/transfer-of-

to crystallize the cultural and ultimately legal meaning in the online environment of the concepts underlying the E&Ls for quotation, criticism, review, caricature, parody or pastiche.

- From a legal design perspective, an adversarial procedure has been proposed, i.e. the ability to **contest algorithm (https://journals.sagepub.com/doi/full/10.1177/2053951720932296)** determinations by introducing a public adversarial AI which embeds counterbalancing values (Elkin-Koren, 2020). In this way, users and the public at large may be empowered in their challenges against algorithms that are designed by platforms in collaborations with right holders.

- In the EU, it could be interesting to explore to what extent AI applications employed for content moderation could or should be considered as high-risk AI systems in the sense proposed by the **AI Act (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206)**. As stated in the AI Act proposal, "for high-risk AI systems, the requirements of high-quality data, documentation and traceability, transparency, human oversight, accuracy and robustness, are strictly necessary to mitigate the risks to fundamental rights". Whereas this solution certainly needs further consideration, it possesses several of the elements necessary to mitigate the discussed perils of bias propagation in content moderation.

Algorithmic content moderation is a powerful tool that may contribute to a fairer use of copyright material online. However, it may also embed most of the bias, errors and inaccuracies that characterize the information it has been trained on. Therefore, if the user rights contained in Article 17(7) CDSM Directive are to be given an effective protection, simply indicating the expected results omitting *how* to reach them, may not be sufficient. The problem of over-blocking is not simply a technical or technological issue. It is a cultural, social and economic issue, as well and, perhaps more than anything, it is a power dynamic issue. It is unrealistic to put on equal footing the threat of a (primary) copyright infringement action brought by right holders due to under-blocking on the one hand, with that of individual users experiencing the removal of their parody or criticisms on the other, especially considering that users normally agree in the terms of service of platforms that blocking of their content is at the sole discretion of the platform. Recognizing parody, criticisms and review as "user rights", as the CJEU does in C-401/19, may be a first step towards the strengthening of users' prerogatives. But the road to reach a situation of power symmetry with platforms and right holders seems a long one. Ensuring that bias and errors concealed in technological opacity do not circumvent such recognition and render Article 17(7) ineffective in practice would be a logical second step.

*Acknowledgments: This research is part of the following projects. All authors: the **reCreating Europe (https://www.recreating.eu/)** project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870626. João Pedro Quintais: VENI Project "Responsible Algorithms: How to Safeguard Freedom of Expression Online" funded by the Dutch Research Council (grant number: VI.Veni.201R.036).*

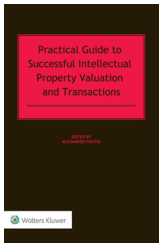
This blog post is based on the EPIP2021 roundtable organised in Madrid (September 8-10, 2021). The authors are grateful to Prof. Niva Elkin-Koren and to Dr. Irene Roche-Laguna for their participation and for their insightful perspectives and suggestions which have been helpful in developing this analysis. The blog post only reflects the view of the authors and any errors remain our own.

United Kingdom (http://copyrightblog.kluweriplaw 2/united-kingdom/) USA (http://copyrightblog.kluwerip 2/usa/)

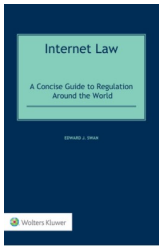
KLUWER IP LAW NEWS ALERT

Stay informed on IP law.

SUBSCRIBE LATEST NEWSLET



Practical Guide to Successf Intellectual Property Valuati Transactions (https://law-store.wolterskluwer.com/s/ the-deal-practical-guide-to-intellectual-property-transac/01t4R00000NqdgyQ
Alexander Puutio
€ 135



Internet Law: A Concise Gui Regulation Around the Worl (https://law-store.wolterskluwer.com/s/ law/01t4R00000OjzqUQAR)
Edward J. Swan
€ 112

BROWSE CATEGORIES

by Jurisdiction...
by Category...
by Contributor...
by Affiliate...
by Date...

RELATED SITES

Kluwer IP Law (http://www.kluweriplaw.com/)
Kluwer Patent Blog (http://www.kluwerpatentblog.com
Kluwer Trademark Blog (http://kluwertrademarkblog.co
Author Portal (https://www.wolterskluwer.com/en/solutions/kluwerla

RSS FEEDS

Summary Feed

Article Feed (http://feeds.feedburner.com/KluwerC

(http://feeds.feedburner.com/KluwerC

AFFILIATES

Kluwer Copyright Cases (http://www.kluweriplaw.com) Members

(htt gID

FINALLY OPENING UP?
THE EVOLUTION OF TRANSPARENCY REPORTING PRACTICES OF SOCIAL MEDIA PLATFORMS

Christian Katzenbach, Selim Basoglu, Dennis Redeker
University of Bremen

Submitted to ICA 2023

November 2022

ABSTRACT

Social media platforms' rise is accompanied by longstanding calls for more transparency on their internal processes, decisions, and practices of content moderation and governance. Platforms have reluctantly responded to this call with increasing release of information. In this context, Transparency Reports have become a common practice and platforms' key instrument to demonstrate legitimacy and accountability. This paper investigates the evolution and status quo of transparency reporting across seven major platforms in a 10Y timespan from 2012 to 2021. We firstly investigate what kind of information and data is being shared in the reports, and secondly we analyse and compare across platforms the actual numbers that are reported for copyright-related content moderation. Our results indicate the establishment of reporting standards across platforms, driven by increasing political and public pressure as well as mutual observation and mimicry. The substantive numbers on copyright content moderation do not align, however, indicating that both practices and rules are still unstable in this context.

KEYWORDS

Platform Governance; Content Moderation; Transparency; Accountability; Copyright

INTRODUCTION

Social media platforms have established themselves as key intermediaries of our societies and political systems. In recent years, their roles and responsibilities have become increasingly the object of public and political debate. While they were striving to position themselves as neutral intermediaries in their

early years (Haupt, 2021), there is now broad consensus that they need to take responsibility for communication and interaction on their sites (Katzenbach, 2021). With hate speech and misinformation becoming major issues of concern in democratic societies, platforms are increasingly under pressure to take action and to regulate and moderate the content on their sites more actively. This includes the development of elaborate platform policies (Katzenbach et al., 2022), the expansion of content moderation teams (Roberts 2019), and the deployment of automated systems of content moderation (Gorwa et al., 2020).

This development of platforms' increasing power is accompanied by longstanding calls for more transparency on their processes, decisions, and the actual content being removed. The more content platforms take down the more they become the "new governors" or "custodians of the Internet" (Gillespie, 2018; Klonick, 2017). This sizable degree of power requires both accountability and transparency to be legitimate. In this context, platforms' transparency reports are a response to both the deployment of content moderation at scale and the increasing demand for more transparency. Since 2012, when Twitter published the first transparency report of a major social media platform, the practice has become a virtual necessity. The "techlash" of the 2010s has put platforms under pressure from political stakeholders to improve operations and to increase transparency. Increasingly as well, transparency reporting is being demanded by governments to make platforms more accountable (Nonnecke & Carlton, 2022). Against this background, it seems that social media platforms have strongly aligned in transparency reporting in a way that might be similar to the "isomorphism through algorithms" that Robyn Caplan and Danay Boyd (2018) have observed for the development of social media and the larger data-driven industry in the early 2010s. In any case, transparency reports have become a key tool "to cultivate legitimacy with users and civil society organizations" (Suzor et al., 2018, p. 393).

Copyright is an interesting case for the study of content moderation and transparency in this field. It has been the most prominent early subject for moderation on platforms. This is due to the fact that most early platforms were set up in the United States, where speech rules are relatively liberal – apart from user content that infringes upon economic interests such as those related to intellectual property. The currently predominant model of copyright rules is based on the US Digital Millennium Copyright Act (DMCA) of 1998, which was born in a pre-platform digital age. Under this regime, copyright infringements are usually brought to the attention of an intermediary platform by the copyright holder, instead of the latter having to go to court immediately. Consequently, platforms "offer a natural point of control for monitoring access to illegitimate content, which makes them ideal

partners for performing civil and criminal enforcement" (Perel & Elkin-Koren, 2015, p. 473) in this area. Outfitted with this power, platforms have long been compelled to shed light into their moderation practices in the field of copyright notices and subsequent take-downs. However, differences persist in how content removal is structured and reported upon across platforms.

This paper investigates the evolution and status quo of transparency reporting with a focus on copyright-based content removal. It describes changes over time and analyses how these changes surface competing pictures of convergence and divergence in social media platforms' content moderation practices. The paper situates these two fields of practices in the general competition between platforms to win market shares, regulatory approval and gain legitimacy. We show that a competition to set new standards for transparency reporting (if not for content moderations as a whole) leads to increasing isomorphism of these practices, as – arguably – no platform can afford to fall behind and be scolded by politicians, advertisers and civil society. At the same time, we observe that although platforms' transparency regarding the data for copyright content moderation is a rising trend, the type of data that platforms disclose in this area does not converge as much and the limits of transparency for each platform remains.

The paper first outlines the background conditions under which transparency reporting and specifically copyright-based content removal occurs. To that end, we discuss the general trend toward greater accountability of non-state actors and the relationship between accountability and legitimacy. We also discuss some of the existing public and civil society demands for increased transparency of social media platform content moderation. Thereafter, we outline our empirical approach, first focusing on our research design and then on definitions of social media platform reporting criteria, including specifically those for copyright-based content moderation.

MAKING CONTENT REMOVALS TRANSPARENT: COPYRIGHT ENFORCEMENT AND BEYOND

Accountability of non-state actors is an important topic in the governance literature due to the increased power of both corporations and other non-state actors in transnational affairs. While political theory and practice have long debated the ways in which public organizations can be held accountable (e.g., through elections, referenda or the right to strike and protest), transnational non-state actors have not had the same attention until the late twentieth century (Redeker & Martens, 2018). Due to the immense power of Internet-based corporations, particularly social media platforms e.g., with their influence on the outcomes of democratic elections or on the things that can be

communicated online, the discussion of transparency of transnational corporations has further increased. The goal of platforms when engaging in transparency-increasing measures - such as the creation of transparency reports, transparency microsites ("centers") that bring together various metrics and by engaging researchers and others - is to gain legitimacy. Legitimacy relates to the "right to govern" in the eyes of the users (the governed) but also, as a response or pre-emptive measure to public regulation, in the eyes of politically powerful stakeholders.

Legitimacy is the currency attained and retained by platforms that they require to respond to political and societal stakeholders such as the press and non-profit pressure groups (Suzor, 2018). Platforms have to prove that their governance over content lives up to standards of democratic legitimacy, specifically "throughput legitimacy"; transparency also, at least in principle, can lead to better results for users and societies ("output legitimacy"), but may not be a replacement for a debate on who should be at the table when rules are made ("input legitimacy") (Haggart & Keller, 2013). When platform content governance rules are being perceived as legitimate, intrusive regulation such as the recent Texas law on political standpoint neutrality (Brodkin, 2022) or the blocking of social media services such as the Twitter ban in Nigeria could be prevented (Akinwotu, 2022). Greater transparency about the type of content and the reasons for removal might help. But what specifically is demanded by various stakeholders with regard to transparency of content moderation in general?

Civil society organizations have put out a number of documents over recent years that enumerate transparency principles, which they demand to be recognized and implemented by the platforms. A recent study analyzed 40 such documents, finding that around half of them explicitly include "good governance principles" such as transparency and accountability among the demands civil society has for content moderation (Celeste et al., 2022). One of the most well-known advocacy documents pertaining to transparency in content moderation are the 2018 Santa Clara Principles on Transparency and Accountability in Content Moderation. The Santa Clara Principles, which exist in a revised second iteration since 2021, include ten principles signed jointly by fourteen civil society organizations such as the Electronic Frontier Foundation, Access Now and Ranking Digital Rights. A detailed principle on transparency demands from platforms to report detailed key indicators on their content actions, with a specific set of information demanded in cases where platforms follow state actions that prescribe content or account takedowns. Platforms "should report information that reflects the whole suite of actions the company may take against user content and accounts due to violations of company rules and policies, so that users and researchers understand and trust the systems in place" (Santa Clara Principles, 2021). A number of platforms have signed up to the content of the Santa Clara Principles,

among them Facebook and Instagram (Meta), YouTube (Google), Reddit and Twitter (Crocker et al., 2019).

Increasingly, regulators prescribe how platforms are required to report about their content moderation practices. India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules of 2021 requires larger platforms to produce monthly reports about complaints and actions taken (Tewari, 2022). The EU's Digital Services Act (DSA) and the Platform Accountability and Transparency Act (PATA) are two legislative proposals that would increase transparency requirements for platforms significantly. Transparency can further be enhanced by providing data on content moderation to academic researchers. Consequently, regulators increasingly perceive "access to data for empirical research (...) as a necessary step in ensuring transparency and accountability" (Nonnecke & Carlton, 2022, p. 610). The DSA specifically, "seeks a new level of granularity in transparency surrounding content moderation practices" surpassing previous national transparency reporting requirements such as the bi-annual requirement of the German NextDG and India's transparency rules (Tewari, 2022). Less in the focus of public attention yet already codified are transparency reporting standards for platforms toward their business partners as part of the EU's Platform-to-business Regulation of 2019 (European Union, 2019). Based on this, Meta now regularly reports to their advertisers not only the number of complaints lodged against decisions and the type of complaint, but also the average time to process such appeals.

With regard to copyright-related notice-and-takedowns, additional voluntary transparency practices exist. For instance, the Lumen project at Harvard's Berkman Klein Center for Internet & Society collects and makes available DMCA takedown notices from those who receive them. This allows researchers and others to gain an understanding of individual practices and overarching trends. As of late 2021, Lumen included more than eighteen million notices, most of them copyright-related, from companies such as Wikipedia, Google (including YouTube) and Twitter (Lumen, 2022).

Pending the passage of some of the more stringent legislative proposals, what is the level of transparency if platforms are being compared? Until 2019, the Electronic Frontier Foundation produced an annual report in which content moderation practices of 16 online platforms were compared based on six overarching categories such as transparency about government takedown requests, transparency about content removal based on the platform's policies, transparency about appeals and even endorsement of the Santa Clara Principles (Crocker et al., 2019). In 2019, for the last iteration of the report, Reddit was able to receive a star in all six categories with Facebook, Instagram, Vimeo and Dailymotion performing particularly poorly.

Ranking Digital Rights produces an annual Big Tech Scorecard, which evaluates the corporate accountability of 14 (2022) large digital platforms from the US, China, South Korea and Russia, subdivided by offered services (such as WeChat and QQ for Tencent Holdings Ltd. or Facebook and Instagram for Meta Inc.) (Ranking Digital Tech, 2022). The report includes indicators on content moderation transparency reporting in its section on freedom of expression, such as the reporting of "data about government demands to restrict content or accounts" and data about platform policy enforcement. Overall, in that section, the report finds that Twitter Inc. “took the top spot, for its detailed content policies and public data about moderation of user-generated content” (Ranking Digital Tech, 2022b).

New America’s Transparency Report Tracking Tool is a continuously updated project that currates data from transparency reports of six services of five platform companies (Singh & Doty, 2021). The tracking tool allows readers to find in one place the *categories of transparency reporting* included in transparency reports of Facebook, Instagram, Reddit, TikTok, Twitter, and YouTube. The tracking tool also allows an over-time view of when certain reporting categories have been added or dropped by the services. What is not included is any attempt to find common categories of transparency reporting that would allow to compare changes over time between the different platforms.

In this report, we aim to provide a new way of looking into platform transparency reporting through comparison of what is being reported on between platforms and over time. In addition to this meta-perspective, we investigate the reporting of the same set of platforms with regard to actual numbers of copyright-related content moderation, in order to understand both trends and the quality of information that is available through this reporting.

METHODS

RESEARCH DESIGN AND SAMPLE

The research design for this paper is a longitudinal comparative approach, operationalized with a qualitative content analysis of published transparency reports of major platforms and the information given in their transparency centers. We compare reporting practices and substantial copyright-related decisions of seven of the largest social media platforms: Facebook, Instagram, Pornhub, TikTok, Tumblr, Twitter and YouTube. These were selected based on content production, user traffic and economic impact.

To create a baseline for reporting, we surveyed a larger set of 20 social media platforms to arrive at a broad understanding of how platforms report[1]. From this broader sample, we created a set of generic operational definitions of transparency categories (Table 1) and operational definitions of *copyright-related* transparency categories (Table 2). Based on these categories, a qualitative content analysis was performed on a narrower sample of seven platforms' transparency reports. For the first analysis, we coded whether certain transparency categories are reported on. For the second analysis with a focus on copyright, we performed a substantive coding in line with the categories set for substantive copyright-based moderation reporting. An important aspect of the study is its focus on longitudinal changes. The chosen platforms' transparency reports were investigated since the year 2012, which is the year the earliest transparency data are available. The data collection extends until 2021, which is the year the latest complete data for transparency are available (as of mid-2022). This allows us to identify trends on transparency of platforms and changes in how copyright-based content moderation changed over time. In addition, for the substantial copyright-related data, this approach allows us to observe the impacts of new regulations on content moderation.

DIMENSIONS OF PLATFORM TRANSPARENCY REPORTING

For the criteria comparison part of the study, the data collection for disclosed data is made through an investigation of platforms' transparency reports and newly established "transparency centers" including, if available, revealed datasets. In addition, platform policies were also scrutinized to better understand the reporting and to be able to construct the transparency categories.

Following an examination of the 20 platforms' transparency practices, every reporting measure regarding content moderation is categorized in Table 1. We excluded data that is unrelated to platforms' own content moderation, for example transparency about transfer of user data in the context of governments' information requests. Thus, the collection of transparency categories in Table 1 was created by taking similarities and differences into account considering the following three aspects: (1) the data which are directly relevant to content moderation, (2) information about the platforms' way of presenting data, (3) information about how platforms moderate the content.

(1)	Reporter ID	Whether the platform reveals the identity of the copyright reporters

[1] In addition to those mentioned previously, we also examined transparency actions of Audius, Diaspora, Dribble, DTube, FanFiction, Mastodon, Periscope, Pixelfed, Soundcloud, Twitch, Vimeo, Vine and WordPress.

	Effects on Users	Whether the platform shares data about how many users were affected due to the content removals
	Exemplary Cases	Whether the platform shares exemplary cases or detailed information about content removal requests
	Invalid Requests	Whether the platform shares how many of the copyright infringement reports were invalid
	Disputes	Whether the platform shares how many content removals due to copyright infringement were disputed by the creators
(2)	Since	Since when the platform shares transparency reports
	Dataset	Whether the platform shares an accessible dataset in addition to the report
	Regional Data	Whether the platform shares separate data for different countries
	Language	Whether the platform provides transparency reports in multiple languages
	Percentage	Whether the platform provides percentages about the removed content apart from the absolute numbers
	Sub-annual reporting	Whether the platform shares data for content moderation on a monthly or quarterly level
	Visualization	Whether the platform provides visual materials related to content moderation data, including the processes involved in platform governance
(3)	Legal Requests	Whether the platform shares data on legal requests for content removal
	Guideline Enforcement	Whether the platform shares data for content removal due to the platform's own guideline
	Management Tools	Whether the platform reveals any data about the algorithmic tools or the way the staff manage the content removal process

Investment and Goal	Whether the platform shares how much they invest in the field of content moderation based on algorithmic decision-making systems.

Table 1: Operational definitions of transparency categories

If information in the transparency categories can be accessed through any of the abovementioned sources (reports, transparency centers), we take this as disclosed data. Moreover, if the data is accessible indirectly, it is also accepted as disclosed data. For instance, if a platform reveals the number for the copyright notices and only the valid requests, the invalid requests are considered to be disclosed data too, since the actual number can be calculated. However, if the data provided by the platform is too vague, then it is understood to be undisclosed data. For example, Facebook, until 2019, reported the removal rate for the copyright requests as the number of notices sustained rather than the number of individual posts deleted (notices may relate to several posts). Due to this, although Facebook provides data for reports submitted and removed content, it does not indicate Invalid Requests (one of our transparency categories) per se before 2019.

INDICATORS FOR COPYRIGHT CONTENT MODERATION

For the substantive comparison of copyright-related content moderation, we have used the same set of sources. Where platforms release numbers bi-annually or quarterly, these numbers were aggregated up and annualized. When data for the whole year is not available, it is indicated in the footnotes. Table 2 shows two kinds of indicators or copyright-related categories, (1) absolute numbers, (2) percentages of copyright related data. The aim of this separation is to allow for useful comparison across platforms.

(1)	Number of Copyright Removals	The total number of pieces of content removed due to copyright infringement in a given year. Trademark and counterfeit data are not included.
	Number of Copyright Reports	The total number of copyright reports (notices) provided to a platform in a given year. If notices may contain reports for multiple posts, it is indicated in the footnotes.
	Number of Disputed Claims	The total number of disputed claims related to copyright infringement

	Number of Removed Content due to Guideline Enforcement	The total number of pieces of content removed due to the enforcement of guideline of platform
	Number of Removed Content due to Legal Request	The total number of removed content due to the demands of governments and courts in a given year.
(2)	% of Copyright Removal as Share of Overall Reports (Notices)	The percentage of removed content as share of the total number of reports in a given year
	% of Disputed Claims Decided in Favor of Uploaders	The percentage of disputed claims concerning content removal due to copyright infringement resulting in favor of uploaders as share of the total number of disputed claims
	% of Disputed Claims Decided in Favor of Reporters	The percentage of disputed claims concerning content removal due to copyright infringement resulted in favor of reporters as share of the total number of disputed claims

Table 2: Operational definitions of copyright-related transparency categories

Some platforms do not disclose absolute numbers but only percentages for copyright reports, disputed claims, or copyright removals. To gain a more complete and comparable view, we conducted reverse calculations. For instance, the approximate number of the removed content is accessed through a reverse calculation from the disclosed percentages.

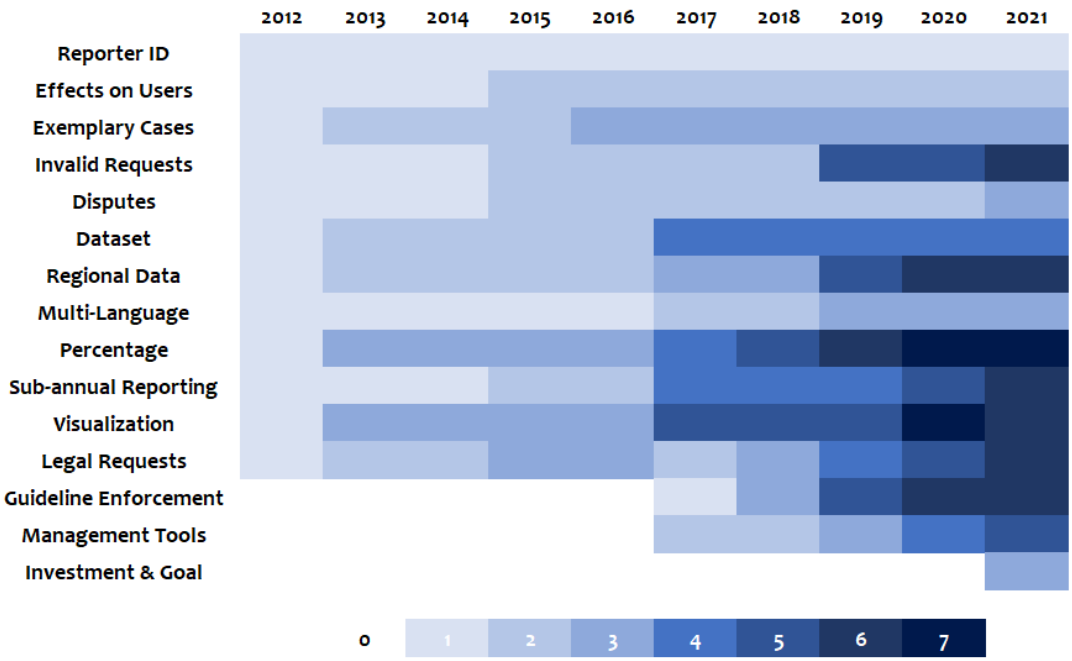
RESULTS: PLATFORM TRANSPARENCY REPORTING TRENDS

This section outlines the results of the longitudinal comparative study of seven social media platforms investigating the scale of transparency reporting and aiming to reconstruct the evolution of these practices.

TRANSPARENCY REPORTING ON COPYRIGHT CONTENT MODERATION BY INDIVIDUAL PLATFORMS

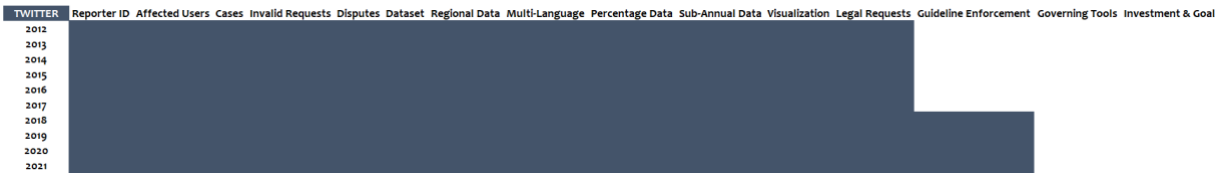
We start with portraying the inclusion and chronological occurrence of transparency categories (see Table 1) in transparency reports of the seven major social media platforms.

Figure 1: Inclusion of reporting criteria in platform transparency reports over time (cumulative)



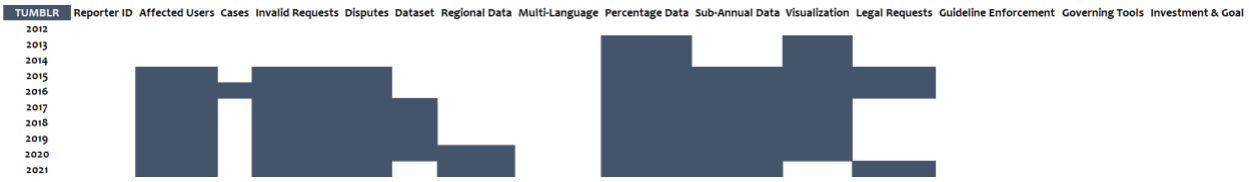
Twitter has published detailed transparency data since 2012, including the identity of the top reporters. Besides, the monthly numbers can be found in the graphs and tables on its transparency website. Except for information concerning the tools and investments for algorithmic detection of copyright, Twitter is quite transparent for all criteria. Twitter consistently sustained the disclosure of data over the years since the first transparency report.

Figure 2: Inclusion of reporting criteria in Twitter's platform transparency documents



Tumblr has published data regarding copyright content moderation since 2015 but it does not produce a dataset nor does it reveal information about the tools and investments for algorithmic detection of copyright. Additionally, Tumblr only revealed data about the total number of disputed claims and valid counter-notices by the uploaders until 2018. Since then only the number of valid disputed claims was revealed. Thus, the actual number of disputed claims and the actual rate for the disputes resulting in favor of uploaders or reporters remain undisclosed.

Figure 3: Inclusion of reporting criteria in Tumblr's platform transparency documents



In 2017, **Facebook** and **Instagram** started to disclose data about copyright-based moderation, first ones among the investigated platforms. Since then, they reveal the number of removed content due to copyright infringement. Facebook started to calculate the copyright removal rate differently in 2019, taking the actual number of items which are subject to a copyright report. Before that, the company reported removal rate as the share of reports causing (partial) removal actions. Considering that a report can contain files for a vague amount of content, it is impossible to calculate the actual rate of the removed content as a share of the reported content, producing challenges for transparency. Facebook and Instagram also have never disclosed any data concerning disputed claims unlike platforms such as Twitter, Tumblr, YouTube.

Figure 4: Inclusion of reporting criteria in Facebook's platform transparency documents

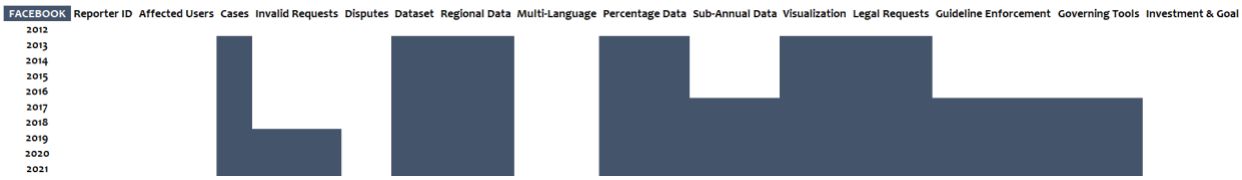
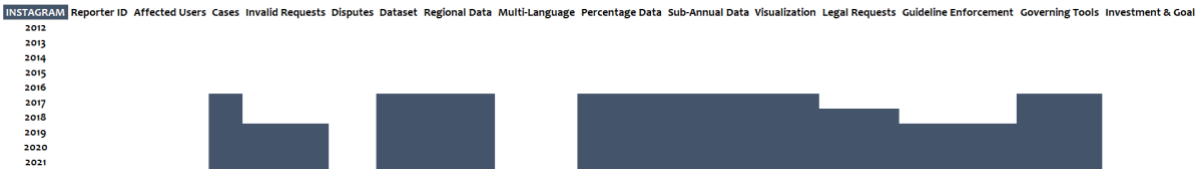
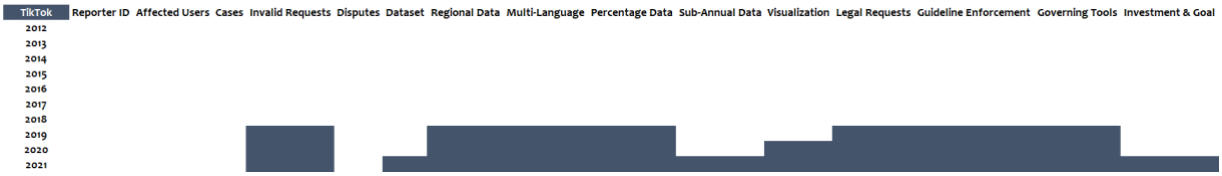


Figure 5: Inclusion of reporting criteria in Instagram's platform transparency documents



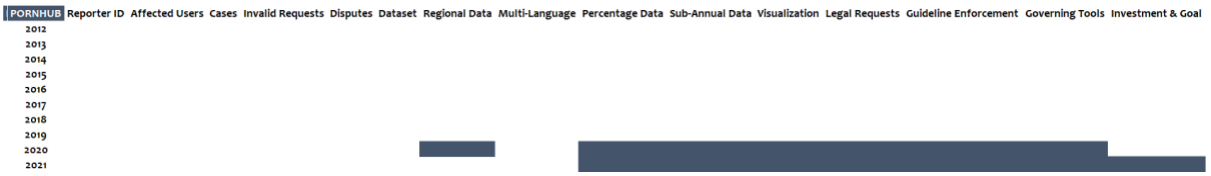
In 2019, **TikTok** started to publish data about copyright content moderation, albeit in a very limited way. An essential information regarding content moderation, the number of removed posts was never revealed. The platform reveals only the number of notices and the removal rate. However, it is not certain whether this removal rate is only for the number of contents or notices which are subject to partial action-taking. Since a copyright-notice might include multiple reports, the actual amount and rate of valid requests are not transparent. Moreover, TikTok does not reveal any data about disputed claims and their results. Thus, not much information is available about TikTok's moderation of copyright content.

Figure 6: Inclusion of reporting criteria in TikTok's platform transparency documents



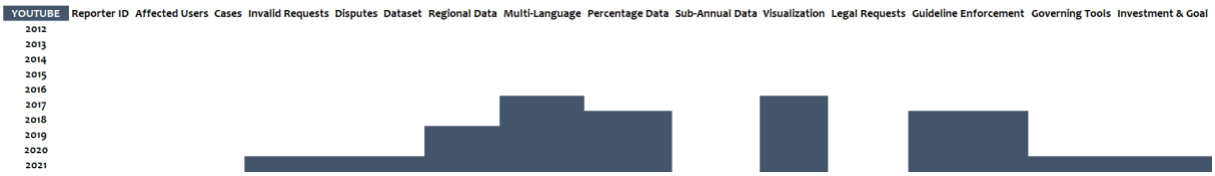
In 2020, **Pornhub** began to reveal data about copyright removal. For the year 2020, there is only the total number of removed content. For 2021, additionally they revealed the data of the number of received notices. However, Pornhub does not publish the number of videos reported, but the number of notices which one may contain claims for several videos. Hence, it is impossible to calculate the removal rate as a share of videos included in notices.

Figure 7: Inclusion of reporting criteria in YouTube's platform transparency documents



Lastly, **YouTube** published numbers for copyright content moderation for the first time with 2021 a quite comprehensive transparency report. The number of removed content, the number of copyright notices, the removal rate, the number of disputed claims and the results of the disputes were revealed. Besides, Youtube gave information about their tools, investments and goals for further development of algorithmic systems.

Figure 8: Inclusion of reporting criteria in Pornhub's platform transparency documents



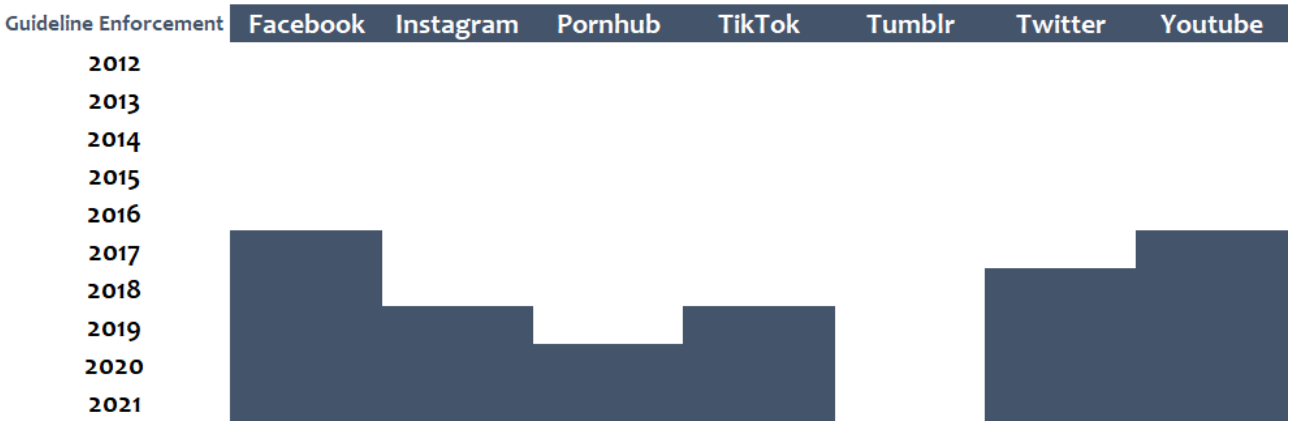
COMPARATIVE ANALYSIS OF TRANSPARENY REPORTING ON COPYRIGHT CONTENT MODERATION

When looking comparatively at transparency reporting with a focus on copyright content moderation, disclosure of *numbers for invalid requests* is an important parameter. It allows to assess the excessive use of copyright laws by copyright owners against user-generated content, indicating strategy by rightsholders to systematically overblock content. Twitter shares data about invalid requests since

their first report in 2012. Tumblr joins Twitter on disclosing the number of invalid requests in 2014. Facebook, Instagram and TikTok started to reveal the data for invalid requests simultaneously in 2019. This represented the first time that TikTok disclosed any data publicly. Tumblr has disclosed such data since 2015. Facebook and Instagram started to be relatively transparent regarding copyright-related issues in 2017. YouTube disclosed the data for invalid requests on its first transparency platform in 2021. Pornhub, on the other hand, never revealed such data in its transparency reports. Twitter is the only platform revealing the IDs of most frequent copyright reporters. Although Twitter discloses such data since the first transparency report, none of the other platforms have revealed such data. Summing up, platforms have adopted a common transparency practice with regard to the core issue of invalid requests, only with Pornhub as an exception.

Also with regard to *content removal based on conflict with platform content policies* (also community guidelines) there is isomorphism across platforms, only Tumblr is not disclosing data on this. The evolution of this common practice is interesting, though. It is actually one of the latest category that was introduced into transparency reporting. The number of moderated contents was first shared by Facebook and YouTube in 2017. Only subsequently in 2018, Twitter started to disclose the data for content removed due to its "Twitter Rules" content. Instagram, Pornhub, and TikTok started to reveal the data for such platform-policy based moderation of content then in 2019 whereas Tumblr, considering that it is one of the platforms publishing transparency reports relatively early, surprisingly, never disclosed such data.

Figure 9: Comparative chronological inclusion of policy-based removal data in platform reporting

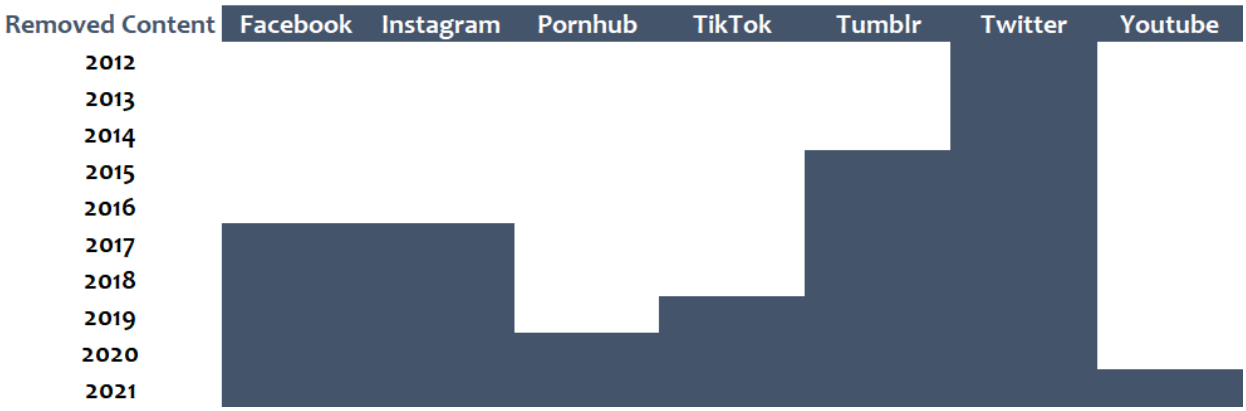


Since *algorithmic tools* are involved in various ways in many elements of the content moderation processes of large platforms, and due to their significant societal and political implications,

transparency regarding the tools is also important (Gorwa et al., 2020). Considering that their first disclosed data for content taken down due to platform policies or copyright actions is from 2017, we can assume that Instagram and Facebook have been transparent about their algorithmic detection tools since 2017. Pornhub and TikTok provided information about the algorithmic detection tools they deploy in 2019, and YouTube introduced its algorithmic tools for content moderation in 2021.

When it comes to *copyright content moderation*, the most essential data relates to the scale of removed content. Twitter leads the other platforms in this category, too, by starting to reveal the data for the number of removed content items already in 2012. Tumblr joined Twitter in 2015, followed by Facebook and Instagram two years later, in 2017. Since 2020, Pornhub declares the number of removed content items. YouTube, perhaps the most discussed platform with respect to copyright enforcement, disclosed relevant data only in 2021, in its first transparency report. Tiktok reveals only the number of copyright notices received, rather than the number of removed content for copyright reasons.

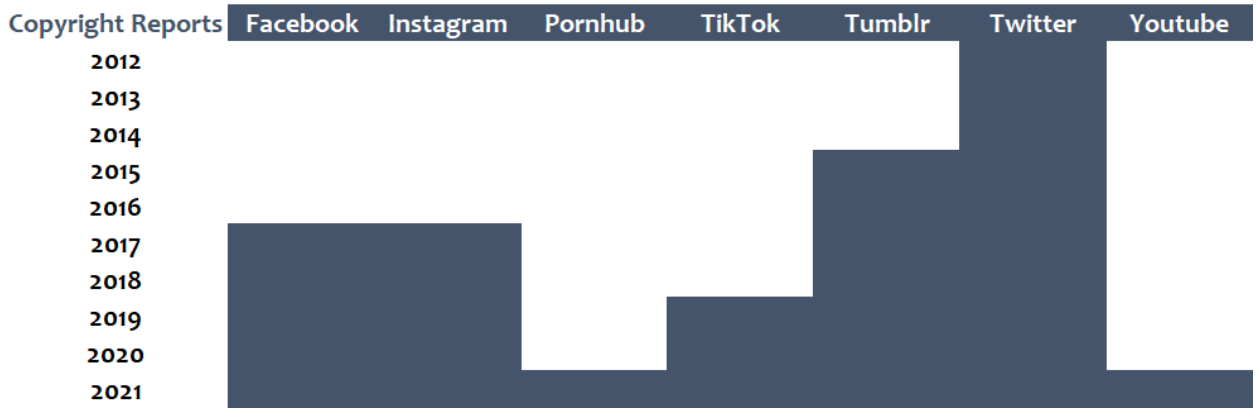
Figure 10: Comparative chronological inclusion of copyright-based removal data in platform reporting



.

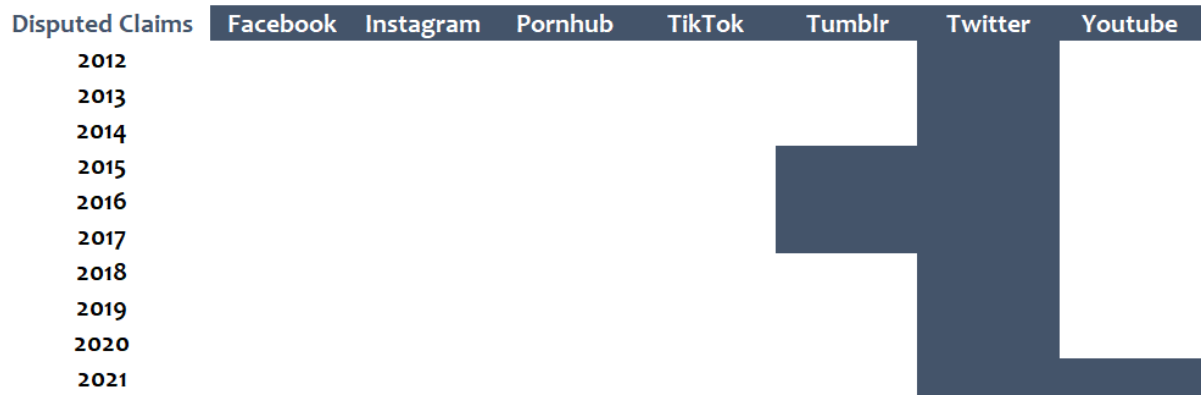
The number of *copyright notices* is a transparency category all platforms agreed to reveal at some point. Many of them share these notices with Lumen as discussed above. In 2012, Twitter began to publish data about copyright notices sent to the platform. Tumblr followed in 2015. Facebook and Instagram publish data on notices since 2017, while TikTok does so since 2019. Pornhub and YouTube reported the numbers only in 2021.

Figure 11: Comparative chronological inclusion of data on copyright notices in platform reporting



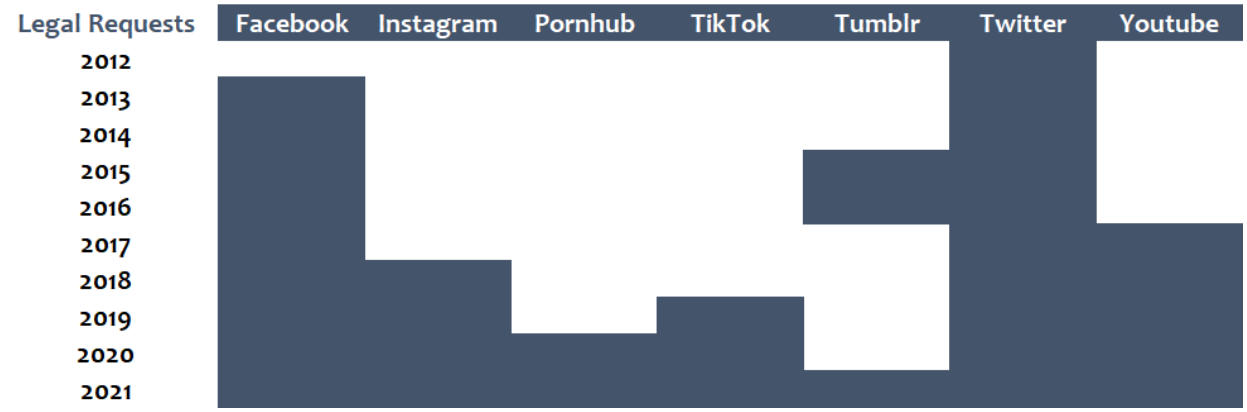
The *number of disputes* is another category, which is relevant to the merit of requests. Disputes arise when the platform and copyright owner agree while the user who posted the content counter-noticed the platform arguing that the post was lawful. This category is key to understanding copyright practices, as it reflects a platform's sensitivity towards any kind of unrightful removal of user-generated content, and platform users' tendency to protect their rights when they feel their rights are violated. However, interestingly, the data for disputed claims remains one of the most obscure transparency categories in transparency reports. Twitter discloses the number of disputed claims since its first report. Although Tumblr starts to reveal data for disputed claims in 2015, it abandoned this practice after three years, and now only reveals absolute numbers of disputed claims that resulted in favor of users. This does not allow us to understand the overall number of disputed copyright notices and the share of those disputes with a favorable result for the user. YouTube revealed data for the disputed claims only in 2021. The other platforms do not reveal any data for the number of disputed claims. Unlike invalid requests, disputed claims reflect the conflict between not only copyright owners and users, but also users and platforms. Hence, it seems quite interesting that Facebook, Instagram and TikTok do not reveal such data at all, and Tumblr changed reporting to decisively less relevant data point on this matter.

Figure 12: Comparative chronological inclusion of data on disputed copyright claims in platform reporting



Last but not least, transparency concerning the number of *removals based on government requests or injunctions by courts* is also a noteworthy category. Strikingly, Facebook started to disclose much earlier than many other platforms in our sample when it began to do so in 2013, still one year later than Twitter. In 2015, Tumblr started to share the number of content removals due to government demands. However, this practice only lasted for two years. From 2017 to 2020, it did not publish such data, publishing the data again in 2021. Instagram shares data on this matter since 2017. TikTok, Pornhub, and YouTube made such data available in their first transparency reports in 2019, 2020 and 2021 respectively.

Figure 13: Comparative chronological inclusion of state-requested removal data in platform reporting



RESULTS: SUBSTANTIAL ANALYSIS OF COPYRIGHT CONTENT MODERATION

After review the kind of data shared by platforms in transparency reports, we now turn to the actual numbers in the reporting with a focus on copyright content moderation. This will provide us with a

quite comprehensive picture of copyright-based moderation practices, comparing several platforms across the several dimensions of content moderation – but also helps to understand the opportunities and limits of transparency reporting as a means to understanding content moderation.

CHRONOLOGICAL OVERVIEW ON COPYRIGHT CONTENT MODERATION NUMBERS

We start by giving a descriptive chronological overview of *content removal due to copyright demands*. Twitter is the only platform in our sample providing numbers for 2012-14, and Tumblr is joining in 2015 (cf. Table 3). As mentioned above, there is no data for the policy-enforcement based content moderation for Tumblr at all and for Twitter until 2018, therefore such data is lacking in this table.

	2012	**2013**	**2014**	**2015**		**2016**	
	Twitter	**Twitter**	**Twitter**	**Twitter**	**Tumblr**	**Twitter**	**Tumblr**
Copyright Removal	13,079	54,759	96,229	146,849	362,242	290,771	284,863
Copyright Reports	6,646	12,433	25,847	33,733	25,882	59,371	23,626
Removal Rate	46	61	69	70	84	70	84
Disputed Claims	5*	16	31	148	115	516	205
Disputes resulted in favor of Uploaders	100	100	96	100	40	100	70
Disputes resulted in favor of Reporters	0	0	4	0	60	0	30
Legal Requests Removal	44	264	2,233	5,707	103	3,712	255

*: The data is only for the second half of the year.

Table 3: Twitter's and Tumblr's content moderation data for 2012 - 2016.

For 2012, the scale of moderation due to copyright notices and juridical demands the former is clearly more significant quantitatively. And in 2013, possibly because of the rising popularity of the platform, Twitter was subject to a significant increase both for the number of copyright notices and the amount of removed content. Unsurprisingly, legal demands were increasing in hand with the platform's popularity. Yet, the rate for the number of disputed claims resulting in favor of uploaders remained the same. Also, in 2013 Facebook reported the content removal data caused by legal demands for the second half of the year which was 7,371. In 2014, numbers almost doubled. And in the same year, one and only disputed claim resulted in favor of reporter has occurred. And Facebook removed 18,481 content due to legal demands.

In 2015, Tumblr joined in disclosing data. The numbers for Twitter kept increasing and the number of removed contents whereas the removal rate did not show any significant change. On the other hand, at Tumblr the number of content removals due to copyright was even greater than the twice the size of Twitter's despite the lower number of notices. And, unlike Twitter, there was no advantage for uploaders in the case of disputes. Yet the number of government-censored contents was drastically lower than the copyright censored contents consistently over the years, just like Twitter. Also, in 2015 Facebook reported 76,395 removed content because of the legal demands.

In 2016, after four years of transparency data reporting, the number of removed contents due to government demands decreased for the first time, on Twitter. Tumblr's data shows the opposite pattern. The government demands caused 255 content removals, almost 2.5-times greater than the previous year. Facebook, with a significant decrease, removed 16,610 pieces of content due to jurisdictional requests in 2016.

2017 represents the first year for which Facebook and Instagram disclose copyright removal data (cf. Table 4). There is no disclosed data for the number of disputed claims but the number for platform guideline-based removals of content amounted to more than 1.4 million items - for the final quarter of 2017 only. In 2017, Tumblr did not disclose any data regarding judicial and government demands for post removal. Data about Youtube's guideline enforcement for the final quarter of 2017 can be reached through Google's transparency center. Data shows that 7,762,431 videos were removed because they were contradicting the platform's terms and policies.

	Facebook	Instagram	Twitter	Tumblr

Copyright Removal	3,7 M	805,065	441,702	201,510
Copyright Reports	479,329	189,077	87,828	20,053
Removal Rate	68.1*!	67.6*!	78	86
Disputed Claims			924	104
Disputes resulted in favor of Uploaders			100	97
Disputes resulted in favor of Reporters			0	3
Legal Requests Removal	42,330		2,784	

*!: Facebook's reported removal rate.

Table 4: Facebook, Instagram, Tumblr and Twitter's transparency data for 2017.

In 2018, there was a considerable setback in reporting numbers on content governance (cf. Table 5). Alas, Tumblr stopped disclosing the numbers of disputed claims in 2018 after it had done that for 3 years. 2018 was also the only year in which Instagram shared data for the legal requests causing content removal, albeit only for the second half of the year. Twitter disclosed the data for the platform's content removal due to its policies for the first time, yet only for the second half of the year. Lastly, according to Google's data, Youtube removed 32,923,038 videos because they did not suit the platform's rules.

	Facebook	Instagram	Twitter	Tumblr
Copyright Removal	3,7 M	2,8 M	734,573	222,084
Copyright Reports	664,311	501,743	111,149	20,873
Removal Rate	73.6*!	71.4*!	66	93

Disputed Claims			2,214	27
Disputes resulted in favor of Uploaders			100	
Disputes resulted in favor of Reporters			0	
Guideline Enforcement Removal	8,5 M		1,2 M*	
Legal Requests Removal	49,500	1,800*	5,307	

*!: Facebook's reported removal rate.

*: only for the second half of the year

Table 5: Facebook, Instagram, Tumblr and Twitter's transparency data for 2018.

2019 was a year in which on the Meta side, there was not a significant increase in copyright removals (cf. Table 6). Also, Youtube removed 31,935,763 videos, slightly less than 2018. A fortunate point for the sake of transparency is that both Facebook and Instagram began to calculate the removal rate upon the exact number of reported contents rather than taking even the reports which caused partial actions as valid ones. Also, 2019 is the year when Instagram began to reveal the numbers for content removal due to guideline enforcement, besides being the first year for Twitter's first report on the same case for the complete year. In 2019, TikTok also reported transparency data for the first time. (cf. Table 6).

	Facebook	Instagram	Twitter	TikTok	Tumblr
Copyright Removal	3,6 M	2,7 M	1,3 M		92,599
Copyright Reports	752,358	392,901	229,709	4,683	7,775
Removal Rate	76.5	76.2	47	84	88
Disputed Claims			10,463		15

Disputes resulted in favor of Uploaders			100		
Disputes resulted in favor of Reporters			0		
Guideline Enforcement Removal	15,6 B	20,2 M	4,7 M	49,2 M*	
Legal Requests Removal	32,400	1,186	5,726	53	

*: Only for the second half of the year

Table 6: Facebook, Instagram, Tumblr, TikTok and Twitter's transparency data for 2019.

In 2020, a general increase in copyright-related numbers is observed (cf. Table 7). Although Instagram close to being identical to 2019, Facebook and Twitter showed a significant increase in these terms while Tumblr repeated the significant decrease in 2020. In terms of guideline enforcement, Facebook and Instagram showed the opposite patterns while Facebook removed almost 1/3 less than the previous year, and Instagram removed almost five times greater than 2019. In 2020, presumably due to its increased popularity, legal requests caused 15,763 content removal at TikTok while it was only 53 in 2019. A significant increase can be observed on Instagram too, as it removed 17,200 contents, almost 17 times greater than the previous year. And 34,707,336 videos were removed by Youtube in 2020 due to guideline enforcement. Also, Pornhub published its first transparency report in 2020.

	Facebook	Instagram	Twitter	Pornhub	TikTok	Tumblr
Copyright Removal	5,3 M	2,7 M	4,4 M	650,862		45,656
Copyright Reports	1 M	501,748	283,205		51,547	6,700
Removal Rate	83.9	88.5	57		48	77
Disputed Claims			6,477			14
Disputes resulted in favor of Uploaders			100			

Disputes resulted in favor of Reporters			0			
Guideline Enforcement Removal	11,1 B	97,8 M	6,3 M	653,465	193,6 M	
Legal Requests Removal	47,500	17,200	5,688	1,081	15,673	

Table 7: Facebook, Instagram, Tumblr, TikTok, Pornhub and Twitter's transparency data for 2020.

2021 can be considered as the most suggestive year not only because it is the most recent one, but also due to the inclusion of the most platforms and criteria. We present the numbers of platforms in our sample at the Table 8.

	Facebook	Instagram	Twitter	Youtube	Pornhub	Tiktok	Tumblr
Copyright Removal	5,7 M	3,4 M	2,5 M	1,5 B	8,547*		49,545
Copyright Reports	1,6 M	709,651	318,653	1,5 M	6,585	139,607	7,581
Removal Rate	81.6	88.4	32.3	99.7		49	88
Disputed Claims			20,057	7,5 M			16
Disputes resulted in favor of Uploaders			100	61.2			
Disputes resulted in favor of Reporters			0	38.8			
Guideline Enforcement Removal	10,7 B	162,2 M	11 M	25,8 M	245865	320,7 M	
Legal Requests Removal	123,100	14,600	15,135	528,025	388	23,033	9,201

*: Pornhub does not publish the number of videos reported, but the number of reports which one

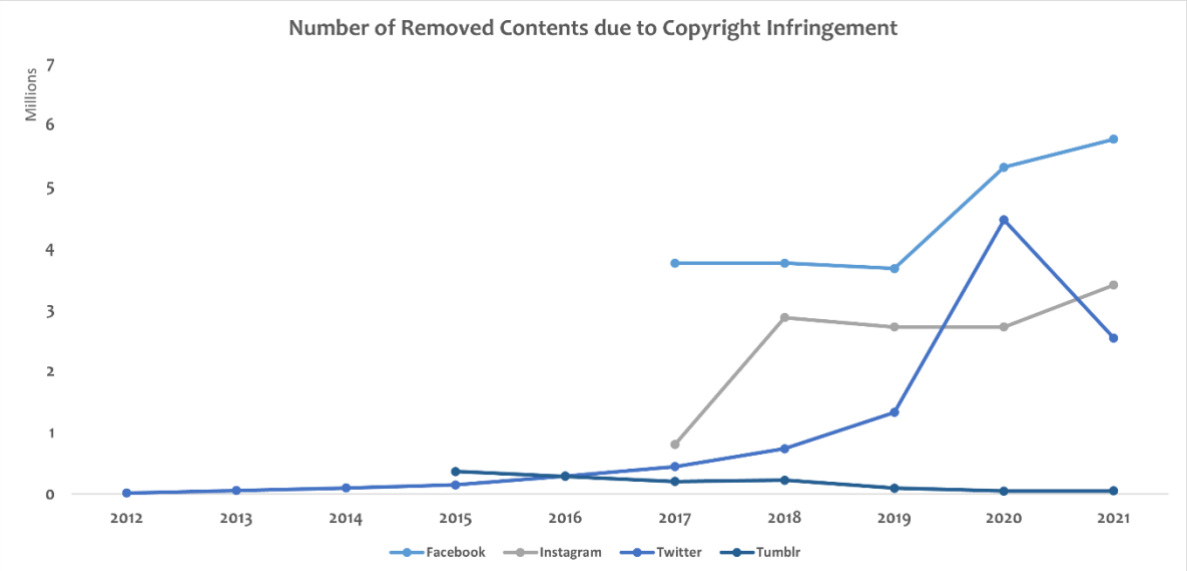
may contain claims for several videos. Therefore, the removed content number is higher than the report and it is impossible to calculate the removal rate in percentage.

Table 8: Platforms' transparency numbers for 2021.

COMPARATIVE ANALYSIS OF COPYRIGHT CONTENT MODERATION NUMBERS

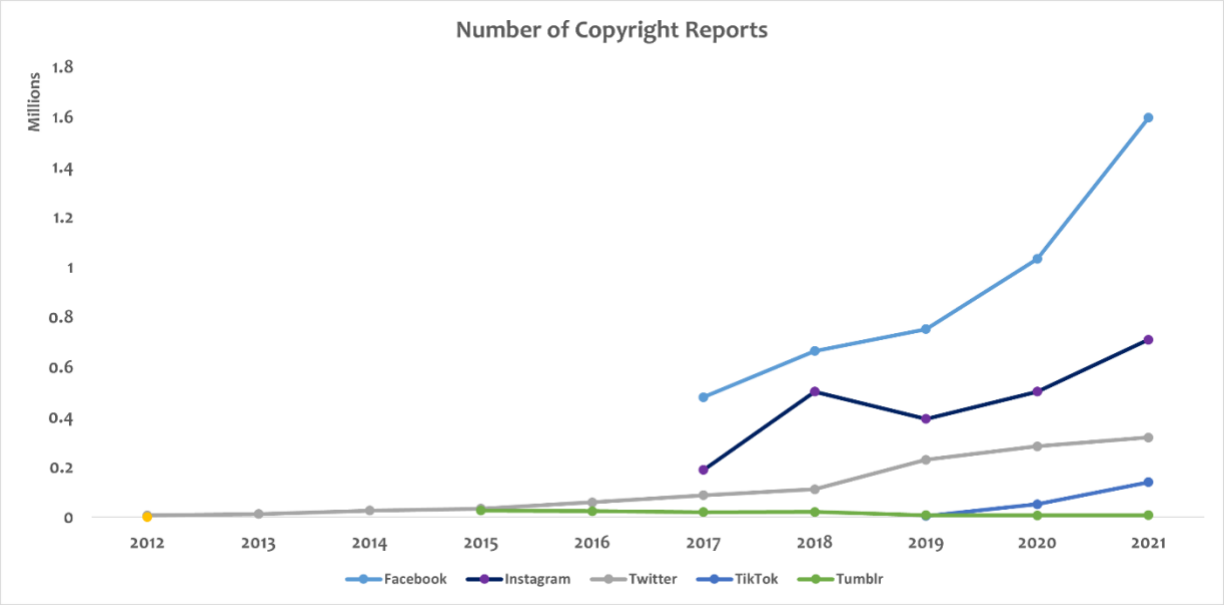
Next, we analyse the data described above across the platforms under study and years for a selected set of categories. Figure 14 shows the *amount of content removed for copyright reasons*. Over the years, it is possible to observe a general increase in the amount of removed content overall on the social media platforms, driven by the growth of the user base and impact of Facebook, Instagram and Twitter, specifically.

Figure 14: Amount of content removed due to copyright claims per platform



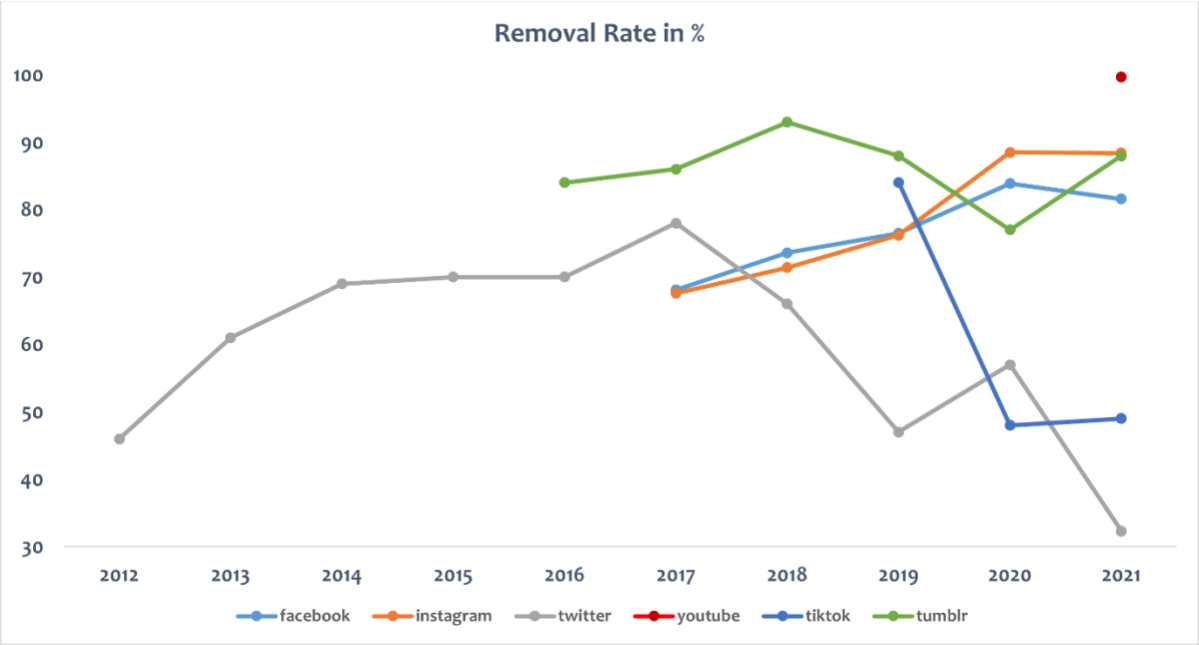
A similar pattern can also be observed concerning the number of *copyright notices* reported to the platforms. Except for Tumblr, almost every year platforms receive more such takedown notices compared to the previous year. For Twitter the increase is relatively mild whereas for Instagram, numbers increase significantly. As can be seen in Figure 15, there has been a considerable acceleration in the number of notices received by Facebook, particularly: from 2019 to 2021, the number of notices has almost doubled.

Figure 15: Number of takedown notices received per platform



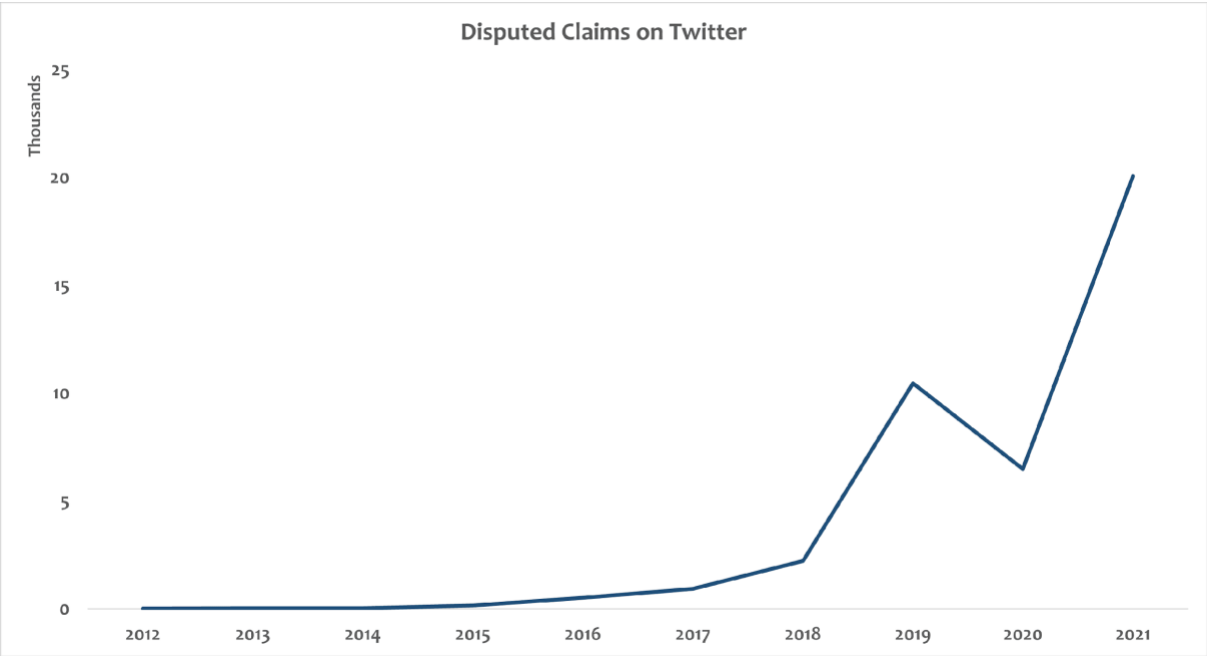
With regard to the *removal rate*, though, there are interesting and far more complex trends. Twitter, displays a remarkable downturn starting in 2017 after consistent increase, pointing at some structural change, either in the nature and merit of the takedown notices or in the (political) attitudes toward these notices and to the platform's users. A similar decrease is also visible for TikTok. In 2019, the first time TikTok revealed numbers for its removal rate, it was around 80 percent, while the next year it decreased to around 40 percent, staying steady in 2021, too. For Facebook and Instagram, a general increase is significant. As seen in Figure 16, no particular time period was subject to an overall increase or decrease, so there does not seem to be a common force to affect removal rates or a sense of convergence of copyright moderation practices across platforms. YouTube published a transparency report only for 2021, so it does not allow for an over-time comparison yet; the existing data point indicates a far greater removal rate than other platforms.

Figure 16: Removal rate for copyright-based moderation per platform



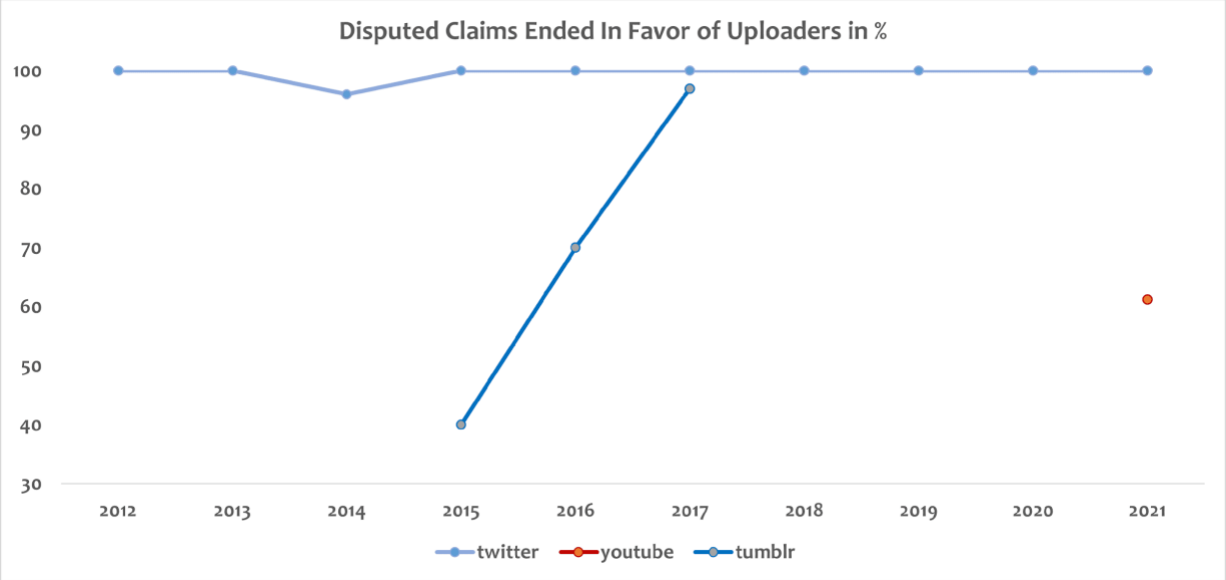
Data for *disputed claims* is one of the least disclosed data by the platforms. Only Twitter reveals the numbers consistently since 2012. In Figure 17, a remarkable increase is visible: 2019 was subject to a significantly higher number of disputed claims. Although there is a decrease in 2020 compared to 2019, the number of claims significantly increased as it almost doubled the number of 2019 in 2021.

Figure 17: Number of disputes about copyright-based moderation, data for Twitter



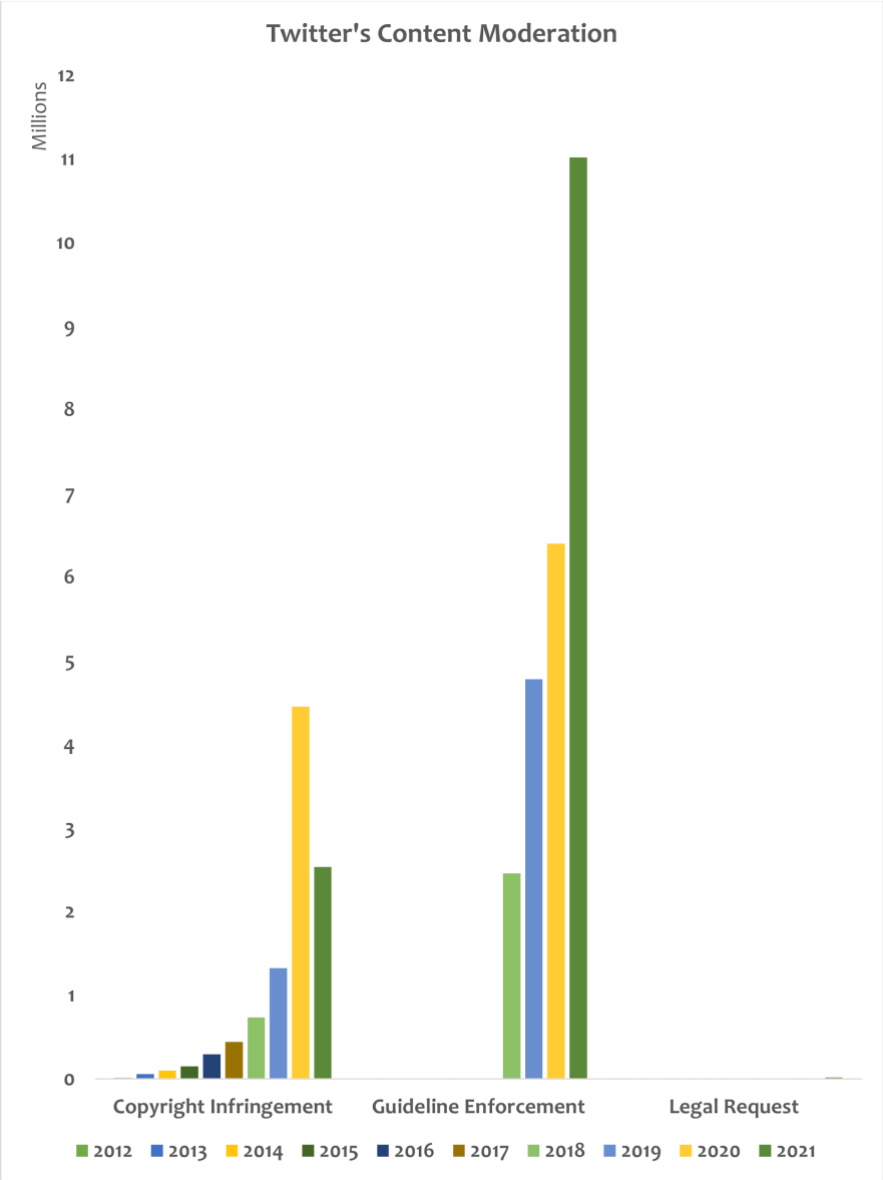
The *share of disputed claims resulting in favor of uploaders* is surprisingly high and consistent for Twitter. Except for 2014, in which one case resulted in favor of the reporter, uploaders were found to be right in all claims every year. A sharply increasing trend is visible with Tumblr. From 2015 to 2017, the rate more than doubled, increasing from 40 percent to over 90 percent. However, it is not possible to track the trend ever since, because Tumblr is not transparent about disputed claims since 2017. YouTube declared the rate as 60 percent in its transparency report in 2021.

Figure 18: Share of disputes resolved in favor of uploaders per platform



What are the most-important *reasons for a platform to remove content*? Looking at Twitter as an example - since it reports data for the longest period of time and most consistently, we can observe that content removal has generally increased in recent years (see Figure 19). We can also see that the majority of deletions occurred due to guideline enforcement, followed by copyright-related reasons. Demands by governments and courts play a smaller role, at least in terms of overall numbers (not necessarily in terms of political importance).

Figure 19: Reasons for content removal, data for Twitter



**: Data for guideline enforcement in 2018 was only for the second half of the year. To create the consistency in the graph, the number was doubled as a representative figure of 2018.*

DISCUSSION AND CONCLUSION

In this paper, we examined two different questions: How do large social media platforms practice transparency reporting? And, secondly, what does the reporting reveal about the scale and nature of copyright-related moderation of these platforms?

With regard to the first part, in which we analyze which transparency categories can be found in the transparency reports and transparency centers, and to what extent seven large social media platforms report in these categories, we argue that a significant trend toward convergence can be observed. Over the years, transparency reports have added additional categories for reporting - most recently data about algorithmic tools deployed and about investments aimed to improve these tools. But it also seems as if a regulatory competition between platforms is in full swing. We observe a strong tendency toward converging transparency reporting practices in the field of content moderation and we suggest that platforms - through extended reporting - try to outcompete each other for positive attention from important social and political stakeholders. They by no means react to the same political and judicial pressures that have increased in recent years. But they also, no doubt, observe the practices of their competitors in order to match increased transparency. From these results the race for more comprehensive transparency reporting as we have seen it when examining the reports of different platforms over time. Interestingly, it appears that - while no doubt imperfect in its transparency reporting - Twitter is the innovator the competitors follow, to not be left behind.

In the field of copyright-based content moderation, which we investigated based on substantive reporting data from the same seven social media platforms, we find that notable divergences characterize the developments in recent years. Rather than mirroring the trend toward convergence seen in the first part of the analysis, here platform practices of reacting to takedown notices became entrenched or changed in idiosyncratic ways that no common practice can be identified. Perhaps, because copyright has been a highly political topic since the start of the digital revolution, platforms find their own ideological niche that is expressed by the degree to which they aim to empower users (vis-a-vis alleged copyright holders) or protect the intellectual rights of creators (or how they balance the two). Internal mechanisms of content moderation, management processes and ideology might all contribute to diverging outcomes for copyright holders and those who upload content. Platforms can stand at the side of uploaders (like in the case of Twitter) or they can have a less protective stance (such as the one data point we know from YouTube or Tumblr in 2015 and 2016. A low removal rate might also mean that a variety of actors use takedown notices and their chilling effects to pursue economic or political interests (Seltzer, 2010).

Any transparency reporting, and this report subsequently, have a number of important limitations that potentially jeopardize platforms' perceived accountability and positive effects of the reporting on their legitimacy in the eyes of external stakeholders. First of all, "aggregated data in transparency reports only shows the platforms' own assessments, and not the merits of the underlying cases (and)

researchers cannot evaluate the accuracy of takedown decisions or spot any trends of inconsistent enforcement" (Keller & Leerssen, 2020, p. 228). Additional limitations of transparency reports in their current form are that they largely focus on the removal of content (and accounts) rather than other (often called "softer") forms of moderation. More recently, practices described as "shadow banning" have taken hold on platforms; users' content is not outright deleted but instead merely not shown to wider audiences, effectively stymying free expression (Savolainen, 2022). Due to the lack of notice of users and their resulting inability to dispute such a moderation measure, shadow banning or the related downranking of content are controversial issues. Even the extent of such "softer" practices is still relatively opaque as "platforms like Instagram, Twitter and TikTok vehemently deny the existence of the practice" (Savolainen, 2022, p. 1092). Shadow banning is likely less relevant for copyright-based moderation, since there are more categorical issues when intellectual property is being reproduced without permission. In general, the lack of information on how moderation algorithms work is a shortcoming for platform transparency; platforms often engage in "black box gaslighting" to deflect critique (Cotter, 2021). All in all, there is still room for improvement of platform transparency practices, as there is for their moderation practices.

Bibliography

Akinwotu, E. (2022). *Nigeria lifts Twitter ban seven months after site deleted president's post*. The Guardian. Retrieved on October 28, 2022 from: https://www.theguardian.com/world/2022/jan/13/nigeria-lifts-twitter-ban-seven-months-after-site-deleted-presidents-post

Brodkin, J. (2022). *Sorry, Texas: Supreme Court blocks law banning "censorship" on social media*. Ars Technica (1 June 2022). Retrieved on October 28, 2022 from: https://arstechnica.com/tech-policy/2022/05/sorry-texas-supreme-court-blocks-law-banning-censorship-on-social-media/

Caplan, R., & Boyd, D. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society, 5*(1), 205395171875725. https://doi.org/10.1177/2053951718757253

Celeste, E., Palladino, N., Redeker, D., & Yilma, K. (2022). Digital Constitutionalism: In Search of a Content Governance Standard. In: Edoardo Celeste, Amélie Heldt, Clara Iglesias Keller (Eds.). *Constitutionalising Social Media*, 267–288. Oxford: Hart Publishing.

Cotter, K. (2021). "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, 1-18.

Crocker, A., Gebhart, G., Mackey, A., Opsahl, K., Tsukayama, H., Williams, J. L., and Jillian C. York (2019). *Who has your back? Electronic Frontier Foundation*. Retrieved on October 28, 2022 from: https://www.eff.org/wp/who-has-your-back-2019

European Union (2019). *Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services* (Text with EEA relevance). Retrieved on October 28, 2022 from https://eur-lex.europa.eu/eli/reg/2019/1150/oj

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance: *Big Data & Society, 7*(1), 1–15. https://doi.org/10.1177/2053951719897945

Haggart, B., & Keller, C. I. (2021). Democratic legitimacy in global platform governance. *Telecommunications Policy, 45*(6), 102152.

Haupt, J. (2021). Facebook futures: Mark Zuckerberg's discursive construction of a better world. *New Media & Society, 23*(2), 237–257. https://doi.org/10.1177/1461444820929315

Katzenbach, C. (2021). "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society, 8*(2). https://doi.org/10.1177/20539517211046182

Katzenbach, C., Kopps, A., Magalhaes, J.C., Redeker, D., Sühr, T., & Wunderlich, L. (2022). The Platform Governance Archive. A longitudinal dataset to study the governance of communication and interactions by platforms. Unpublished Manuscript. Alexander von Humboldt Institute for Internet and Society,

Keller, D., & Leerssen, P. (2020). Facts and where to find them: empirical research on internet platforms and content moderation. *Social Media and Democracy: The State of the Field and Prospects for Reform, 220*, 224.

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review, 131*, 1598.

Lumen (2022). *About us*. Retrieved on October 28, 2022 from: https://lumendatabase.org/pages/about

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society, 20*(11), 4366-4383.

Nonnecke, B., & Carlton, C. (2022). EU and US legislation seek to open up digital platform data. *Science, 375*(6581), 610-612.

Perel, M., & Elkin-Koren, N. (2015). Accountability in algorithmic copyright enforcement. *Stan. Tech. L. Rev., 19*, 473.

Ranking Digital Tech (2022). *The Big Tech Scorecard 2022*. Retrieved on October 28, 2022 from: https://rankingdigitalrights.org/index2022/

Ranking Digital Tech (2022b). *Key Findings from the 2022 RDR Big Tech Scorecard*. Retrieved on August 5, 2022 from: https://rankingdigitalrights.org/mini-report/key-findings-2022/regen

Redeker, D., & Martens, K. (2018). NGOs and accountability. In: Aynsley Kellow, Hannah Murphy-Gregory (Eds.). *Handbook of Research on NGOs*, 303-324. Cheltenham: Edgar Elgar.

Roberts, S. T. (2019). *Behind the screen*. Yale University Press.

Santa Clara Principles. (2021). *The Santa Clara Principles: On Transparency and Accountability in Content Moderation*. Retrieved October 28, 2022, from https://santaclaraprinciples.org/

Savolainen, L. (2022). The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society, 44*(6), 1091–1109. https://doi.org/10.1177/01634437221077174

Seltzer, W. (2010). Free speech unmoored in copyright's safe harbor: Chilling effects of the DMCA on the first amendment. *Harv. JL & Tech., 24*, 171.

Singh, S. & Doty, L. (2021). *The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules*. New America. Retrieved on October 28, 2022 from: https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media & Society, 4*(3), 2056305118787812.

Suzor, N., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette, 80*(4), 385–400. https://doi.org/10.1177/1748048518757142

Tewari, S. (2022). *Transparency initiatives in the DSA: An Exciting Step Forward in Transparency Reporting*. Lumen Project. Retrieved on October 28, 2022 from: https://www.lumendatabase.org/blog_entries/transparency-initiatives-in-the-dsa-an-exciting-step-forward-in-transparency-reporting

Mandate to Overblock?
Understanding the impact of EU's Art. 17 on automated copyright content moderation on YouTube

Daria Dergacheva, Christian Katzenbach
University of Bremen

Submitted to ICA 2023

November 2022

Abstract

This article presents the results of measuring possible over-blocking and changes in diversity of cultural products supply on YouTube in the EU member states. We used a comparison of the YouTube mapping by Reider et al (2020) and the current platforms' snapshot of selected channels and videos collected in 2022 to answer the research questions. We assessed how the increased automated copyright moderation related to the Copyright in the Digital Single Market (CDSM) Directive (2019/790) passed in the EU might have influenced content removal. We also studied diversity of cultural supply according to Stirling's model of diversity using dual-concept diversity index (McDonald & Dimmic, 2003). To our knowledge, this is the first attempt to measure changes that were predicted to happen in cultural diversity supply on YouTube after the new rules on copyright came into force in the EU. Over-blocking and possible impact on cultural diversity are notoriously difficult to study and we hope that this research would lay down some ground for further explorations.

Introduction

Spring of 2019 saw thousands of people marching through the capitals of EU countries (Martin, March 2rd, 2019) protesting the future EU Copyright directive which would have made social media platforms liable for user-generated content. Critics were concerned that when forced to automatically filter and remove content, social media platforms would be characterized by less

diversity and more censorship of users (DW, February 14th 2019). The regulation "would mandate Internet platforms to embed an automated infrastructure for monitoring and censorship deep into their network", as prominent Internet founding figures and spokesmen said in their open letter (EFF, December 6th 2018).

The move to pass the European Copyright in the Digital Single Market (CDSM) Directive (2019/790) has been happening while social media platforms had clearly become key players in contemporary societies (van Dijk, 2013), and AI technologies are increasingly presented as solutions to the major societal problems (Katzenbach, 2021). Under increasing public and political pressure, social media platforms have massively increased their efforts to monitor and moderate content on their sites. Platforms have invested strongly in both quickly growing their teams of content moderators (Roberts, 2019) as well as in algorithmic systems to automatically govern contested content (Gorwa et al., 2020),

The CSDM Directive now is highly relevant for the future role of platforms as intermediaries and their impact on cultural diversity and access to culture. The directive distinctly raises the level of platforms' liability for the content that they host in cases of copyright infringement. For large platforms now "automated content filtering is required to comply with the best efforts obligations in Article 17(4) CDSMD" (Quintais, April 26, 2022). Critics' concern of mandatory upload filters and potential structural overblocking thus seems to be real. YouTube's latest Transparency Reportin December 2021 seems to confirm this prospect: The aggregated data show thatin the first half of 2021 YouTube needed to roll back over 2.2 million content take-downs based on users' disputes and appeals(The YouTube Team, December 6th, 2021). Thus, YouTube's automated regulatory Content ID system has generated at least 2.2 million unjustified copyright actions against its users on behalf of rightsholders. (Keller, 2021). Yet, possibly the number of unnoticed take-downs that are unqueally unjustified is even much higher. The issue of private platforms algorithmic moderation systems remains opaque if not completely untransparent, often compared to a "black box" (Perel & Elkin-Koren, 2017). This opacity is multiple, not only technical in nature, as Gray & Suzor (2020, p. 7) note importantly, but includes institutional and legal issues. This exacerbates the lack of accountability of intermediaries, and

concerns that they fail to respect fundamental rights compared to in cases where the judiciary would be involved in the decision-making process." (Jaques, 2018, p.2)

The European Copyright in the Digital Single Market (CDSM) Directive (2019/790) has been adapted and came into force in June 2019. Countries have had two years to implement the CDSM into national law but almost one year after the deadline, on May 19th, 2022, the EU Commission sent out a press-release saying that Belgium, Bulgaria, Cyprus Denmark, Greece, France, Latvia, Poland, Portugal, Slovenia, Slovakia, Finland and Sweden have not yet notified the Commission on changes to national law (Press Release, EU Commission, May 19th 2022).

Against this background, this empirical study investigates the changes and influences in access and cultural diversity on social media and streaming platforms, specifically YouTube, in the context of the CDSM Directive and its implementation in the countries of the European Union. The Research Questions this study is addressing are:

1. What is the scope of automated blocking and moderation of social media content in the EU with regard to copyright? How much content is being blocked and moderated? How does this change over time ?

2. What are the characteristics of content that is being blocked and moderated? Are there structural differences, e.g. with regard to categories?

3. How do changes (or continuities) in automated copyright content moderation relate to the national implementation(s) of Article 17 of the CDSM Directive?

4. Does (the increasing usage of) *automated* content moderation lead to a decrease of the diversity of content supplied on the YouTube in the EU (as predicted by critics of Article 17 CDSM Directive)?

The paper is structured as follows. In the next section we overview previous research on cultural diversity within cultural production, issues of copyright moderation, over-blocking and access in connection to social media platforms, specifically YouTube. Methodology and results section describes our YouTube data collection strategy, methodology, and presents our findings. The first part presents general findings on the copyright takedowns on YouTube in the EU after 2019.

The second part measures the platforms' diversity index in the four countries of the EU in 2022 and compares it with the data from 2019 to see if there were any changes in content supply diversity during that time and whether is varies by the countries in the sample. The discussion and conclusion part elaborates on the results and introduces the debate whether public policy regulations, in the case of European Copyright in the Digital Single Market (CDSM) Directive (2019/790), among other factors, could have a negative impact on cultural diversity in social media and streaming platforms as suppliers of cultural goods. We conclude by pointing out limitations of this research and possible directions for further studies.

Cultural diversity in cultural production research

We situate our research on potential effects of platform regulation and automated content moderation in the concept of cultural diversity. The concept is very complex and broad, and media studies applies it to both source diversity as well as content diversity (Voakes et al., 1996; Deacon et al., 2021). Given the policy interest of our study, the UNESCO definition of cultural diversity offers an instructive starting point. The concept of cultural diversity as a cornerstone for policy developments in cultural field and international negotiations around it, was developed within the UNESCO Universal Declaration on Cultural Diversity (2001) and adopted with the approval of the UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions (2005). Article 8 of the Declaration states, "Cultural goods and services are commodities of a unique kind… particular attention must be paid to the diversity of supply of creative work". In line with this definition, we are focusing on the question of diversity of supply of creative work on social media platforms, taking YouTube as a case study. We further investigate if the diversity supplied on YouTube corresponds to the diversity consumed (Moreau & Petlier, 2004) and how they interact with each other on the biggest video sharing platform in four selected countries of the EU.

Previously, diversity in supply of creative work in providing cultural goods and services has been the focus of media economics studies, sociology and communication sciences. Cultural diversity in the movie industry (Moreau & Peltier, 2004; Lévy-Hartmann, 2011), TV networks Hellman, 2001; McDonnald & Lynn, 2004), recording companies (Ranaivoson, 2010), publishing

(Benhamou and Peltier, 2007), and broadcasting (Farchy and Ranaivoson, 2011) were investigated before. Recent studies of platforms which explored exposure diversity in terms of algorithmic news curation (Woisjezak et al, 2021; Jungers & Starrk, 2022) is one example of the relatively new field of studying platforms as intermediates.

While studying cultural diversity in supply of cultural goods and services, researchers previously used the Stirling model of diversity (1999) which is derived from the models in economics, ecology and information theory, and consists of the measurement of three components: variety, balance, and disparity.

According to the model, the greater variety, balance, and disparity, the greater is diversity. As Ranaivoson (2011) points out, in order to assess, for example, the diversity of the music industry (or any system), it must be first divided into different types of categories, such as for instance geographic origin, title or genre, etc.

Other researchers highlight that it is not always possible to define such component as a disparity in cultural production. For example, Benhamou & Peltier (2007) state that the question of measuring disparity remains unanswered since it must rely on certain assumptions of a distance between for example one geographic origin and another or one movie genre to another. Thus, studies have been considered only two components of cultural diversity measurement, balance and variety (McDonnald & Lynn: 2004, Benhamou & Petlier: 2011). This corresponds to the conceptual paper of McDonald & Dimmic, 2003 who insist that in order to measure cultural diversity, at least two components need to be taken into account. When all the three components are taken into account, it is always done at the expense of the huge assumptions (Ranaivoson, 2011), as in the cases of Benhamou and Peltier (2007) and Farchy and Ranaivoson (2011). It is important to note that many papers consider the impact of policies and regulations on diversity (for example protectionists legislation for movie industry). Jacques (et al, 2018) empirical research sought to assess the diversity in cultural expressions featured on YouTube prior to the adaptation of European Copyright in the Digital Single Market (CDSM) Directive. The study found that parodies were disproportionally removed by YouTube's Content ID, but at the

same time videos from the US and the UK, which initially were over-represented in the sample, were more taken down, thus balancing the countries representation a bit.

This study is using a methodological approach which allows international comparison of YouTube channels and videos by country. We apply the mentioned Stirling model of diversity to assess cultural diversity of YouTube channels in the four countries of the EU, using two components of the model: variety and balance, and track how it has changed in time from 2019 to 2022. To our knowledge, this is the first attempt to use this framework from media economics and sociology in platform studies while focusing on cultural goods supply.

We also evaluate the impact of recent EU policy regulation, European Copyright in the Digital Single Market (CDSM) Directive (2019/790), on cultural production diversity within YouTube, tracing the videos blocked and deleted in the EU on YouTube (using a sample of videos). This approach follows Jacques et al (2018) and Gray & Suzor (2020) studies.

Mapping the YouTube and copyright moderation

The large-scale study of channels on YouTube, was done by Rieder, Coromina, and Matamorez-Fernandez in 2020. This was the first, and due to further inaccessibility of platform's data, perhaps the last large-scale description of the YouTube's platform media system. The mentioned research relied on a sample of channels of over 36M and explored the platform's media system in three main directions: stratification and hierarchization in broadly quantitative terms; channel categories, their relationships, and their proportions, and finally, channels according to country affiliation (Rieder et al, 2020). Our research relies on the same method of channel sampling through network crawling as the one by Rieder et al, thus according to these authors, it differs from other YouTube studies which used individual users or issue samples. Due to YouTube's API v.3 restrictions which appeared in 2021, we could not gather a comparable sample but by using the same method of network crawling, we ensured that our smaller sample was comparable to the previous study. Other attempts of mapping YouTube, prior to Rieder et al, were also done by Burgess and Green (2009 and 2018, second edition), who described YouTube as one of the world's most powerful digital platforms, which combines the logic of community and commerciality. They used content analysis of the most popular YouTube videos for the first

edition, and elaborated through analyzing change in YouTube's development historically, in the second. YouTube's core structure and network was also examined by Paolillo (2008) through tags attached to the published videos. Bärtl(2018) attempted to present an overall characterization of YouTube over the course of the past 10 years, based on a random sample of channel and video data.

Studies of copyright takedowns had previously been multidisciplinary, utilized a variety of methodologies and relied on data from different platforms. Notice and takedown procedures as done by Google (Bar-Ziv & & Elkin-Koren,2017; Urbat et al: 2017); YouTube's Content ID and takedown practices (Tushnet: 2014; Edwards, 2018; Erickson and Kretschmer, 2018), Amazon's Kindle world (Tushnet: 2014).

Perhaps the largest in terms of the data study of YouTube, aimed at understanding copyright moderation by the platform, has been done recently by Gray and Suzor (2020). They were able to use a random sample of the text metadata of 76.7 million YouTube videos that included information about whether and why each video was removed or blocked. Using this data set they applied machine learning to understand which videos were mostly taken down by the copyright moderation and why it might have been problematic for some content (the borderline between fair use and unfair use of copyright material being a very thin one in case of, for example, gaming videos or sports-related topics). Due to latest restrictions of the access to data in YouTube's API v.3, this study is impossible to repeat today.

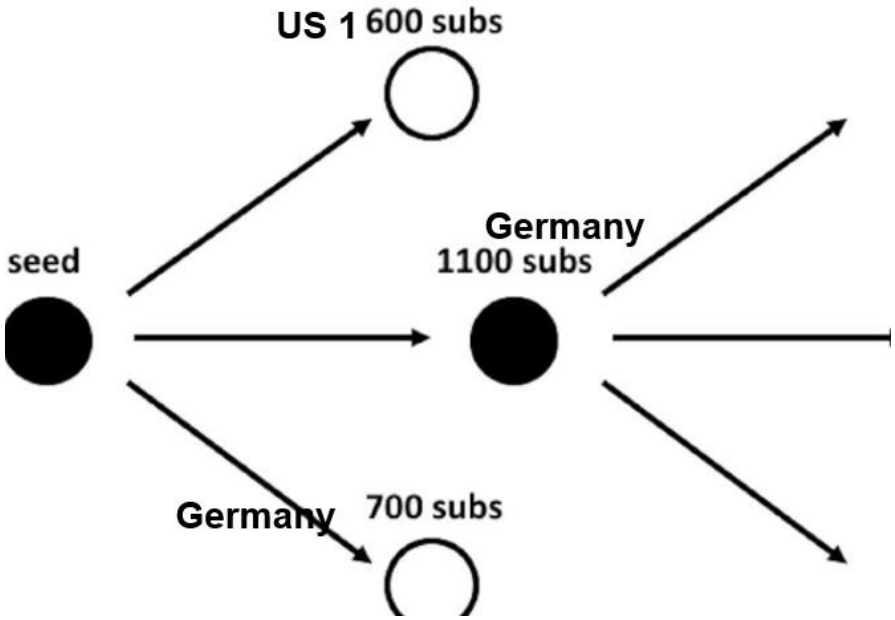
Methodology and Results

Data collection

In order to collect data on diversity of the creative goods supply in the EU, we built a Breadth depth first YouTube crawler in Python by Channels, country, number of subscribers, which works with YouTube API. Following the data collection methodology of Rieder et al (2020) but with more limited capacity due to reduced data access through YouTube API, we built a sample of channels in four EU countries: Germany, France, Ireland and Estonia. Countries in the sample represent both large and smaller member states of the European Union, and their adaptation of European Copyright in the Digital Single Market (CDSM) Directive was different in time: while

Germany was the first country to implement it into the national law in August 2021, France had only done it in november 2021 Estonia was also rather late in implementation (December 2021), and Ireland, although implementing the Directive at the same time as France, has a common law system, which could also may show. The cut-off in the number of subscribers that we used was set at 1000, following both Rieder et al (2020) methodology as well as the logic that YouTube uses in monetization program for its creators: the borderline stands at 1000 subscribers. The crawlers' principles of work are shown at the Graph 1:

Graph 1: principles of crawling the YouTube for channels sampling

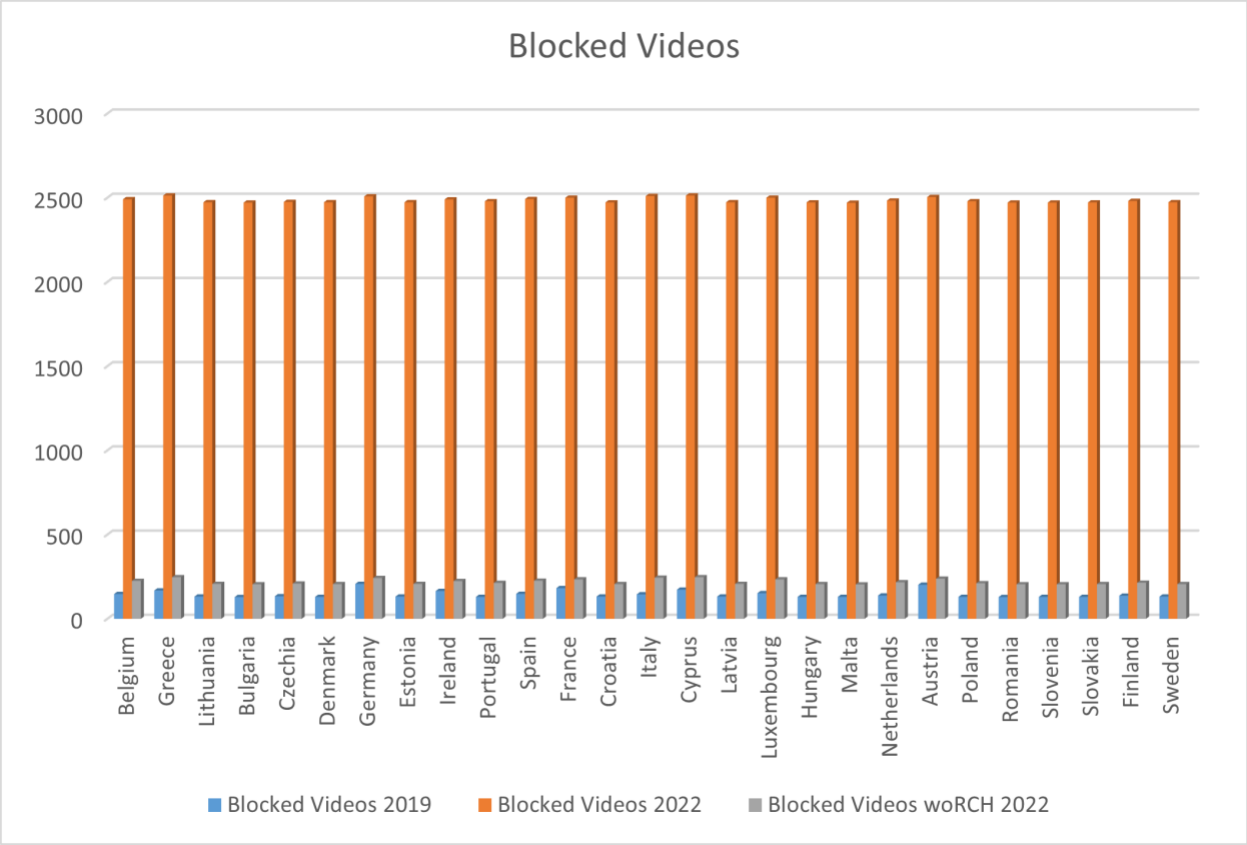


Our first step in studying over-blocking, was to compare the data base of individual videos in the EU . In order to compare the videos that had been blocked or deleted by the time period from June 2019 till June 2022, we gathered through YouTube API v.3 a 2.3 % sample of the metadata of the 4 000 000 data set of videos which were previously collected by the research team of Bernhard Rieder in 2019, or 94 000 videos meta data.

Methodology and results: blocked and deleted videos

As Gray and Suzor noted in their research, only approximately 1% of all videos uploaded are removed due to an apparent copyright violation (Gray & Suzor, 2020). We have found that from 2019 till 2022, in total 2733 videos were blocked in the 27 member states of the EU in our sample of 91 000 videos. This number was surprisingly high: around 3% of all the content, while previous research indicated 1% (Gray & Suzor, 2020).

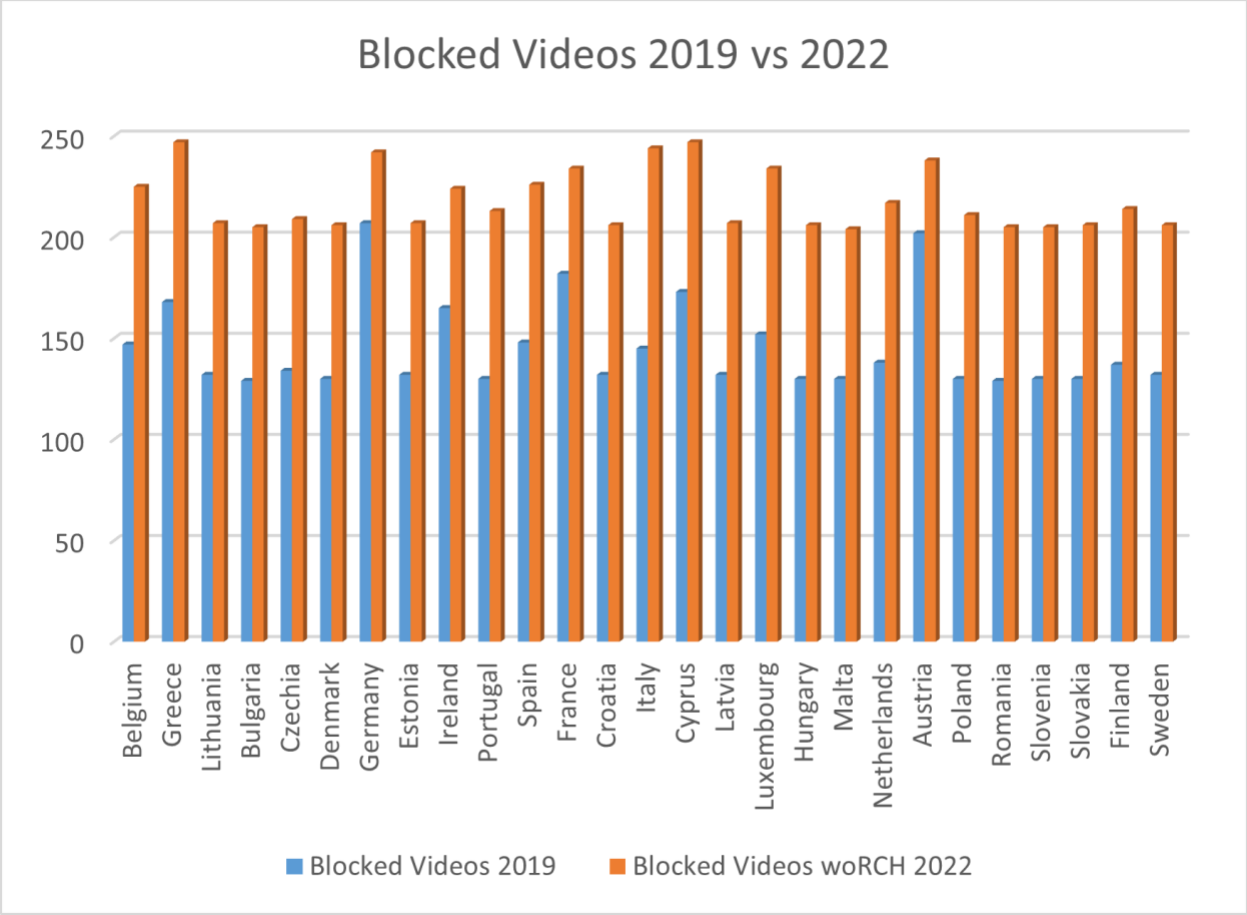
Graph 2: Videos blocked in the 27 member states of the EU (comparison of 2019 and 2022 samples)



Looking at the channel categories for the videos blocked, we have noticed that once again, contrary to previous research, most of these videos were in the "News and politics" category. We manually investigated which channels did these videos belong to, and it turned out that most of the videos from "News and politics" category that were blocked, belonged to the two blocked channels: Russia Today France and Ruptly videos by Russia channels (another product of Russia

Today Channel). Both were blocked by the YouTube after the EU suspended the TV channels broadcasting (March 2022) in lieu of the Russian invasion of Ukraine. While the finding might have been interesting for other research, these channels were outliners for our copyright moderation research questions. Thus, we have decided to exclude them from our data set, and 2513 videos belonging to these channels were excluded. The number of videos blocked without the Russian YouTube channels, was 220 in this case, or 0.25% of the data set.

Graph 3: Videos blocked in the 27 member states of the EU (comparison of 2019 and 2022 samples), without videos from two Russian channels

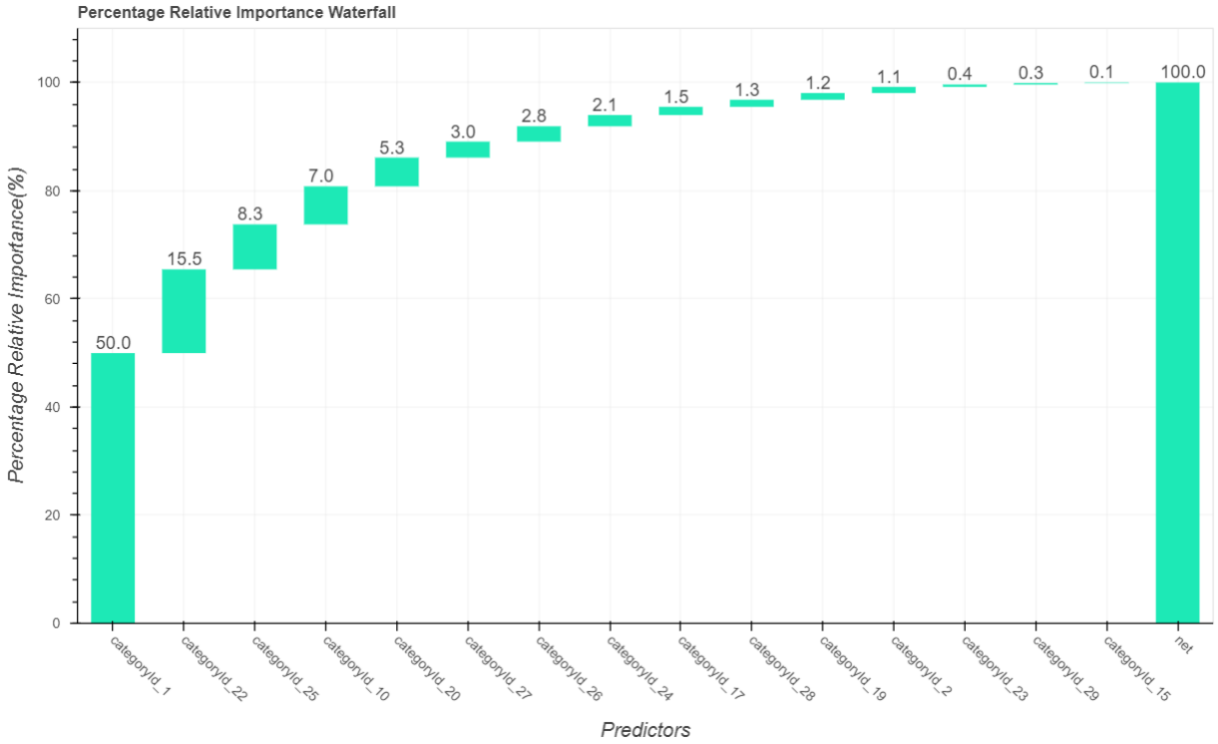


In order to understand the relative importance of categories as predictor of a video being blocked, we used Dominance-analysis approach for comparing predictors in multiple regression (Azen & Budescu, 2003). The determination of relative importance depends on how one defines importance; Budescu (1993) and Azen and Budescu (2003) proposed using dominance analysis (DA) because it invokes a general and intuitive definition of "relative importance" that is based

on the additional contribution of a predictor in all subset models. The purpose of determining predictor importance in the context of DA is not model selection but rather uncovering the individual contributions of the predictors.

In case the target is a continuous variable, as in our case, the method determines the dominance of one predictor over another by comparing their incremental R-squared contribution across all subset models. The results of multimodal regression and dominance analysis done in Python 3.7 package “dominance analysis”, are presented in Graph 4.

Graph 4. Relative importance of categories as predictor of a video being blocked in the EU from 2019 till 2022



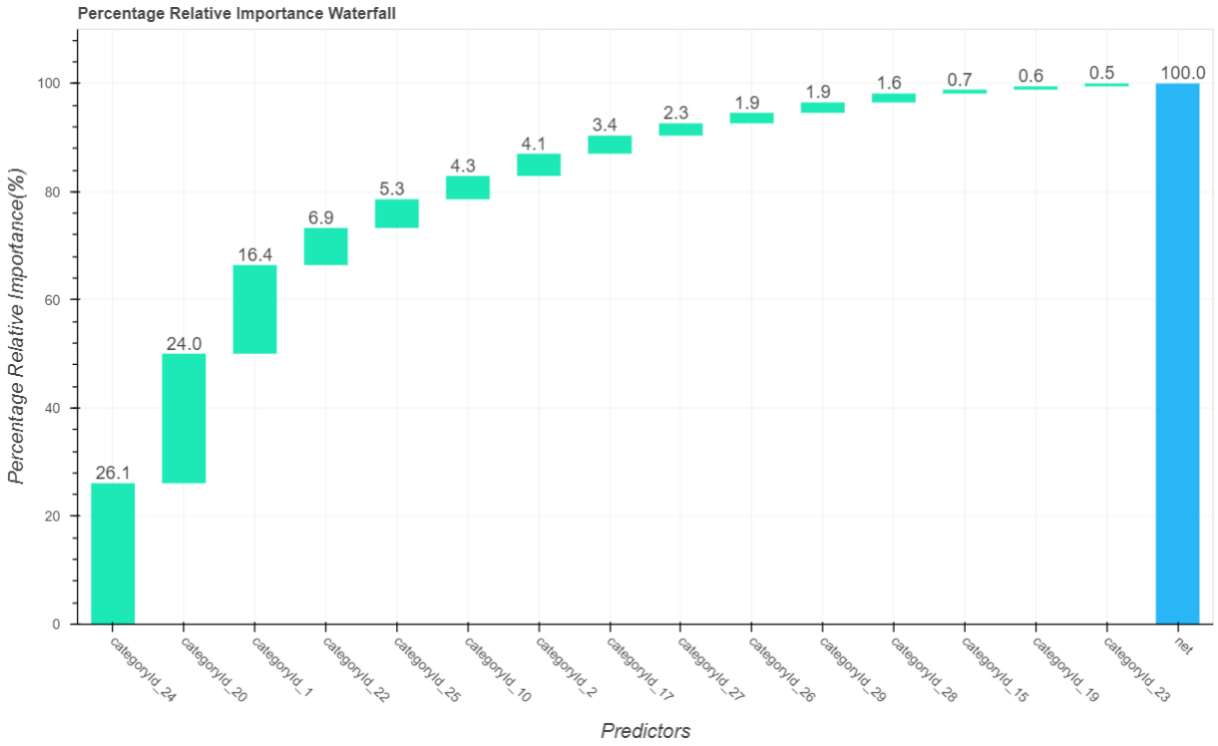
Legend (in order descending in importance)	Description
Categoryid_1	Film & animation
Categoryid_22	People and blogs
Categoryid_25	News and politics
Categoryid_10	Music
Categoryid_20	Gaming
Categoryid_27	Education
Categoryid_26	How to & style
Categoryid_24	Entertainment

Categoryid_17	Sports
Categoryid_28	Science and technology
Categoryid_19	Travel & events
Categoryid_2	Autos & vehicles
Categoryid_23	Comedy
Categoryid_29	Non-profits & activism
Categoryid_15	Pets

We can conclude from the results of this analysis that the five most important predictors for videos being blocked in the EU from 2019 till 2022, were belonging to the categories Film & Animation, People & Blogs, News & Politics, Music & Gaming. Three of these predictors are consistent with previous research on copyright moderation on the YouTube (Gray & Suzor, 2020; Erickson and Kretschmer, 2018). Thus, it makes it possible to suggest that these videos were blocked due to the copyright and belonging to these categories remains a strong predictor for being subjected to copyright moderation on the platform.

Further analyses included comparing the sample of videos meta data (91 000) from two data sets from 2019 and 2022 and finding deleted videos. Using R statistical programming language, we found that 6524 videos were deleted. We could get this finding by comparing the data sets and seeing that the meta data on these videos, although present in the original data set of 2019, was no longer there in 2022. Unfortunately, contrary to previous research (Gray & Suzor, 2020), YouTube API does not longer allow to gather meta data of deleted videos including the reason for deletion. Thus, we had the meta data of deleted videos (from 2019) but could not collect the reason for videos' deletion via API due to restrictions imposed by YouTube. The only way to at least partially find out why the video was deleted, was to manually open the deleted videos pages (via video IDs) and read the reason that YouTube provides on these pages. We found three types of reasons: account disactivated; video made private, and video unavailable. We further manually identified unavailable videos – 459 - in a random sample of 1000 deleted videos, and once again ran Dominance-analysis approach for comparing predictors in multiple regression (Azen & Budescu, 2003) in Python 3.7. The results are presented in Graph 5.

Graph 5. Relative importance of categories as predictor of a video being deleted in the EU from 2019 till 2022



Legend (in order descending in importance)	Description
Categoryid_24	Entertainment
Categoryid_20	Gaming
Categoryid_1	Film & animation
Categoryid_22	People & blogs
Categoryid_25	News & politics
Categoryid_10	Music
Categoryid_2	Autos & vehicles
Categoryid_17	Sports
Categoryid_27	Education
Categoryid_26	How to & Style
Categoryid_29	Nonprofits & activism
Categoryid_28	Science & technology
Categoryid_15	Pets & animals
Categoryid_19	Travel & events
Categoryid_23	Comedy

As it is visible from the analysis, three out of five main predictors of video deletion on YouTube during the period 2019 – 2022 belonged to categories Entertainment, Gaming, and Film & animation. Once again, this is consistent with the previous research on copyright moderation on YouTube. While copyright moderation may not be specific reason for unavailability of video, and

this data could only be obtained from the platform itself (which is currently not possible), we can see that some of these videos belong to the categories mostly deleted via copyright moderation according to previous research (Gray & Suzor, 2020).

Applying the results to our overall sample, we can suggest that 3.6 % of videos were deleted on the YouTube (with a specific reason of video being 'unavailable', not 'private' or 'account deleted'). Summing this number with the videos blocked (0.25%), we can say that almost 3.8% of videos were deleted on YouTube in the EU member states between 2019 and 2022. This is a much bigger number than the usual number for copyright moderation deletion (1% (Gray & Suzor, 2020)). While we can not confirm they were deleted due to increased copyright moderation by YouTube only, this possibility exists and requires both further transparencies, as well as availability of data from the platform.

Methodology and results: impact on cultural diversity of channels in four EU countries

Our sample of channels, gathered in 2022, equals to at least 5% of channels with 1000 and more subscribers that was shared with us by the research team of Bernhard Rieder. It consists of channels in Germany, France, Ireland and Estonia, gathered by the team of Reider et el in 2019 and filtered by country and the number of subscribers. Unfortunately, YouTube quota for retrieving data though API is five million of times less in 2022 than in 2019 (10 000 requests versus 50 000 000 requests per day), thus we were able to compare samples rather than the full data sets from 2019 and 2022. The number of channels from each country are presented in the Table 1.

Table 1: number of YouTube in the data set, by country

Country	**Number of channels**
Germany	1555
France	1555
Ireland	335
Estonia	224

In this section we are assessing whether the diversity of cultural goods has changed in the four countries of the EU with different timeline of the Directive implementation and answer the Research Question 4: Does (the increasing usage of) *automated* content moderation lead to a decrease of the diversity of content supplied on the YouTube in the EU (as predicted by critics of Article 17 CDSM Directive)?

We use Stirling model of diversity (Stirling, 2008) measures of cultural diversity, using two out of the three indicators: variety and balance. In out data, variety is represented by the number of categories the YouTube channels of each country belong to, and balance is represented by a number of channels each of the categories has. We identified categories according to the wiki page which YouTube assigns and creators use as their "primary" category of a channel (__).

Our diversity index, counting both variety and balance, for each country and the years 2019 and 2022, was the Hirschman-Herfindahl index [Σi (pi) 2], which is widely used in social science, economics and media studies (Moreau & Petler, 2004; ____) and takes into account two out of three indicators of the Stirling model. While McDonald & Dimmic (2003) had doubts in using the Hirschman-Herfindahl index in studying diversity, research on film and TV networks shows that that the HHI is, "in reality, an indicator that simultaneously measures variety and balance" (Moreau & Petlier, 2004, p. 129), provided there is data on both (which is the case of this research).If there is no variety, but complete concentration into a single variant, the index is equal to one. Where there is more variety and more balance (distribution of channels by categories), the index is closer to zero. Alternatively, if whole percentages are used, the index ranges from 0 to 10,000 "points". For example, an index of .25 is the same as 2,500 points. We measured the index using R statistical language's package "diverse".

As the first step, we counted all the categories present in each year sample by each country, and attributed all the channels to one primary category, using wiki pages definition of the primary category (as the channel's owner chose it). Further on, we discarded channels where this category was unavailable (NA) and counted the diversity index. Finally, we compared the indexes between years and calculated the percentage of channels in top ten categories for each year and each of the four countries.

Diversity Index. Germany

Germany was among the few Member States to have met the 7 June deadline of the Article 17 CDSMD implementation: it entered into force on 1 August 2021 (Brieske & Peukert,2022).

In Germany, 35 primary categories according to wiki pages were used for identifying channels in 2019, and 39 were used in 2022.

Table 1

Distribution of channels in YouTube by category, Germany 2019		
Germany	**0 - 1**	**0 – 10 000**
2019	**HHI 0.09585823**	**HHI 958**
Action game	16.3	**%**
Lifestyle	15.7	**%**
Entertainment	13.4	**%**
Hobby	11.1	**%**
Gaming	6.3	**%**
Hip hop music	4.7	**%**
Sports	3.7	**%**
Music	3.6	**%**
Pop music	3.2	**%**
Electronic music	3.1	**%**

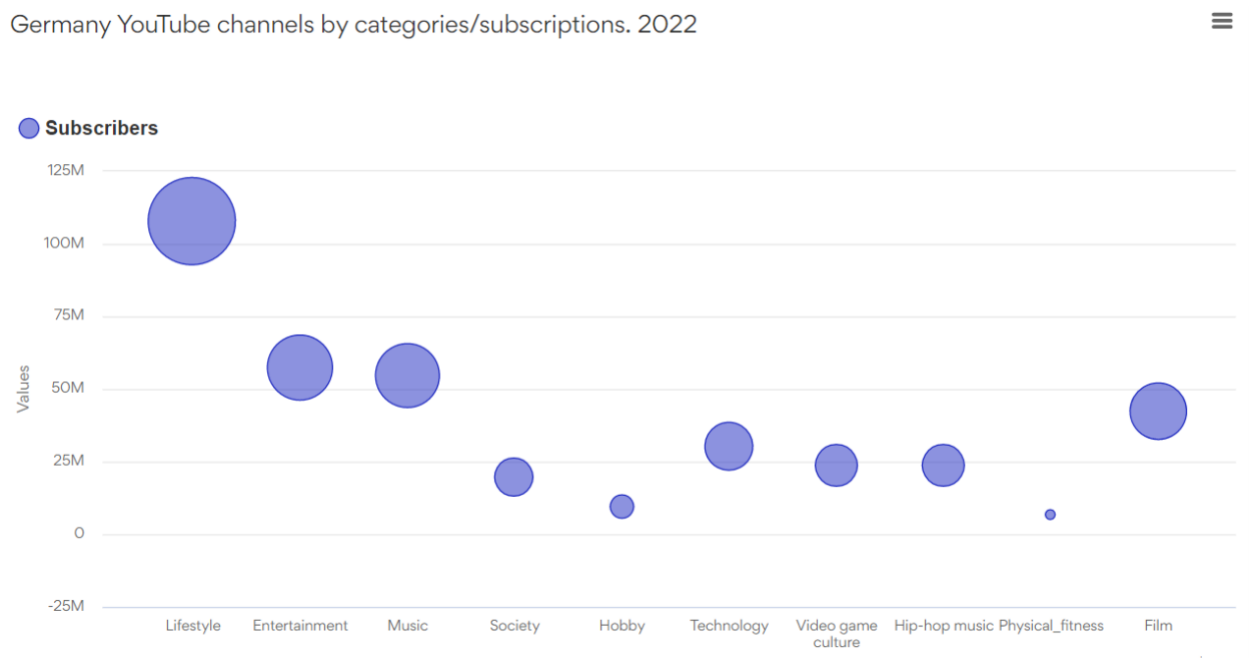
Table 2

Distribution of channels in YouTube by category, Germany 2019		
Germany	**0 - 1**	**0 – 10 000**
Categories 2022	**HHI 0.1225402**	**HHI 1225**
Lifestyle (sociology)	**30**	**%**
Entertainment	**10.2**	**%**
Music	**6.6**	**%**
Society	**5.6**	**%**

Hobby	**5.5**	**%**
Technology	**4.8**	**%**
Video_game_culture	**4.7**	**%**
Hip_hop_music	**4.5**	**%**
Physical_fitness	**3.3**	**%**
Film	**3.2**	**%**

The Herfindhal-Hirschmann index (HHI)HH in Germany's YouTube channels has increased from 958 in 2019 to 1225 in 2022. This means that there is less diversity in Germany's YouTube channels production now than it had been then. It increased by 225 points or 0.225, which is the second highest result in four countries of our sample

Graph 6: most popular YouTube categories by subscribers, France, 2022



YouTube channel categories by subscribers in Germany in 2022		
Categories	Subscribers	Channels
Lifestyle	107677630	395
Entertainment	57540502	135
Music	54653556	87

Film	42252420	42
Technology	30158572	63
Hip-hop music	23919126	60
Video game culture	23880500	62
Society	19797120	74
Hobby	9496260	73
Physical fitness	6866540	44

The most popular categories in Germany (those that had the greatest number of subscribers in our sample), were Lifestyle, Entertainment, Music, Film and Technology, which also looks different from Rieder et al (2020) findings where Technology did not make it into YouTube's top five categories by subscribers.

Diversity index . France

France has implemented the Directive on Copyright in the Digital Single Market ((EU) 2019/790) (CDSM Directive) on 24th November 2021.

In France, 36 primary categories according to wiki pages were used for identifying channels in 2019, and 45 were used in 2022.

Table 3: Distribution of channels in YouTube by category, France 2019

YouTube channels France 2019		
France	**0 - 1**	**0 – 10 000**
2019	**HHI 0.09725204**	**HHI 975**
1) Action game	18.7	%
2) Lifestyle	14.3	%
3) Entertainment	14.2	%
4) Gaming	8.8	%
5) Hobby	8.0	%
6) Hip-hop music	7.3	%
7) Sports	5.2	%

8) Society	3.4	%
9) Music	3.1	%
10) Technology	3.1	%

Table 4: Distribution of channels in YouTube by category , France 2022

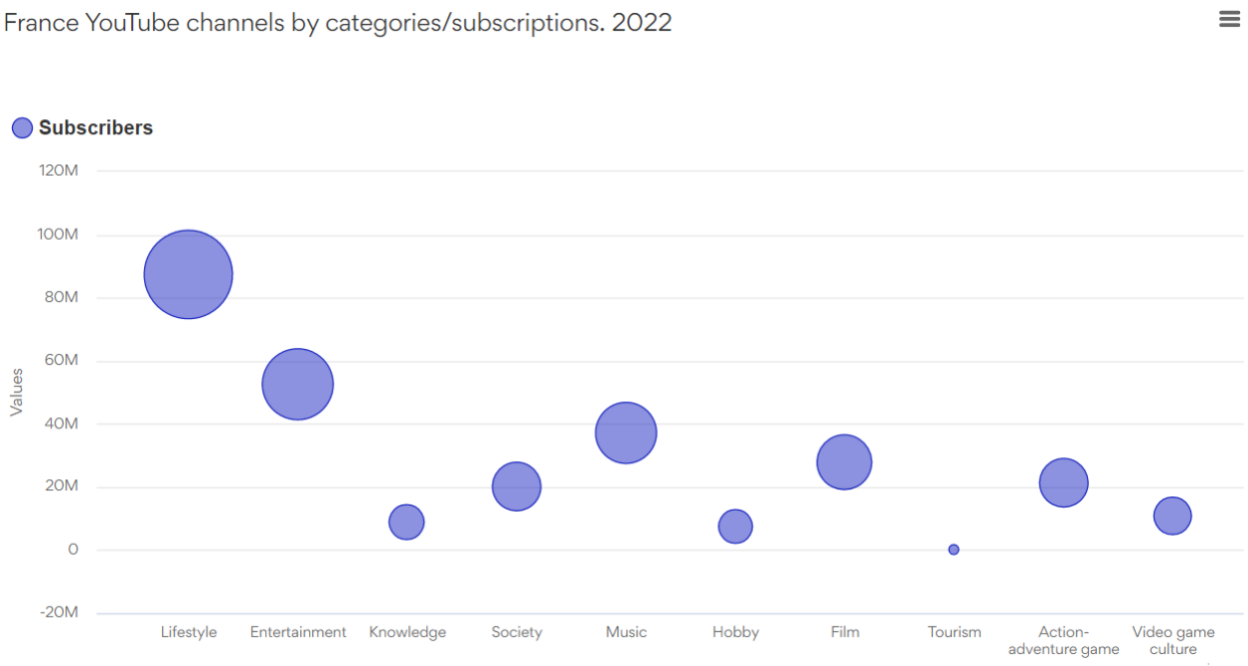
YouTube channels France 2022		
France	**0 - 1**	**0 – 10 000**
2022	HHI **0.1055009**	HHI **1055**
1) Lifestyle	25.9	%
2) Entertainment	11.2	%
3) Knowledge	8.3	%
4) Society	7	%
5) Music	6.1	%
6) Hobby	4.6	%
7) Film	3.9	%
8) Tourism	3.6	%
9) Action adventure game	3.2	%
10) Video game culture	2.9	%

The diversity index in France's YouTube channels has slightly increased from 2019 to 2022. This means that there is less diversity in France's YouTube channels production now than it had been then. However, it only decreased by 80 points (0.008) which is the lowest result of the three countries where it had increased.

Most popular categories. France

Contrary to previous research (Rieder et al: 2020; Paolillo, 2008; Paolillo, et al. , 2019; Bärtl , 2018); in 2022 YouTube's API v3 did not allow us to access channel categories (those of YouTube) but did allow to access "content description" which also had a primary definition of channel's content on wiki page. Thus, we analysed those categories and found differences with the previous analysis. This graph represents the diversity consumed in France's YouTube in 2022.

Graph 6: most popular YouTube categories by subscribers, France, 2022



The most popular categories in France (those that had the greatest number of subscribers in our sample), were Lifestyle, Entertainment, Music, Film and one of the gaming categories, Action-adventure game. Contrary to Rieder et al (2020), Society category was lagging behind these most popular ones (in terms of numbers of subscribers) but only marginally, while during the previous research society had not had many subscribers. It is also possible that the wiki categories split overall Gaming category to sub-categories thus making gaming channels numbers look less prominent. For instance, there are two gaming-related channels in France's sample: Action-adventure game and Video game culture. Both YouTube's own categories (which are not available to be collected through YouTube API now) and wiki categories in channel's 'descriptions' are assigned automatically by the platform.

Table 5: YouTube channel categories by subscribers, France 2022

YouTube channel categories by subscribers in France in 2022		
Categories	Subscribers	Channels
Lifestyle_(sociology)	87332154	317
Entertainment	52525122	137
Music	37205191	73
Film	27669678	47
Action-adventure_game	21257835	39
Society	20248181	85
Knowledge	9023268	101
Video_game_culture	10783255	35
Hobby	7570111	56
Tourism	350772	44

Diversity Index. Ireland

Ireland has become the seventh EU Member State to implement the Directive on Copyright in the Digital Single Market ((EU) 2019/790) (CDSM Directive) on 19 November 2021

In Ireland, 33 primary categories were used for identifying channels through wiki pages in 2022 and 27 categories were identified in 2019.

Table 6: Distribution of channels in YouTube by category, Ireland

YouTube channels Ireland 2022		
Ireland	**0 - 1**	**0 – 10 000**
Categories 2022	**HHI 0.08327914**	**HHI 832**
Lifestyle (sociology)	19.7	**%**
Music	15.3	**%**
Technology	5.6	**%**
Entertainment	5.2	**%**
Knowledge	4.8	**%**
Pop_music	4.4	**%**
Video_game_culture	4.4	**%**
Film	3.6	

Independent music	3.2	**%**
Action game	2.8	**%**

Table 7: Distribution of channels in YouTube by category, Ireland 2019

YouTube channels Ireland 2019		
Ireland	**0 - 1**	**0 – 10 000**
Categories 2019	**HHI 0.0984664**	**HHI 984**
Entertainment	**16**	**%**
Lifestyle	**12.4**	**%**
Action game	**4.5**	**%**
Hobby	**4.2**	**%**
Sports	**3.9**	**%**
Gaming	**3.6**	**%**
Music	**3.6**	**%**
Pop music	**2.9**	**%**

The Herfindhal-Hirschmann index (HHI)in Ireland's YouTube channels has slightly decreased from 984 in 2019 to 832 in 2022. It equals a decrease in 152 points. It means that there is more cultural diversity in Ireland's YouTube channels production now than it had been in 2019. This is the only country among our sample where the diversity supply has increased.

Graph 6: most popular YouTube categories by subscribers, Ireland, 2022

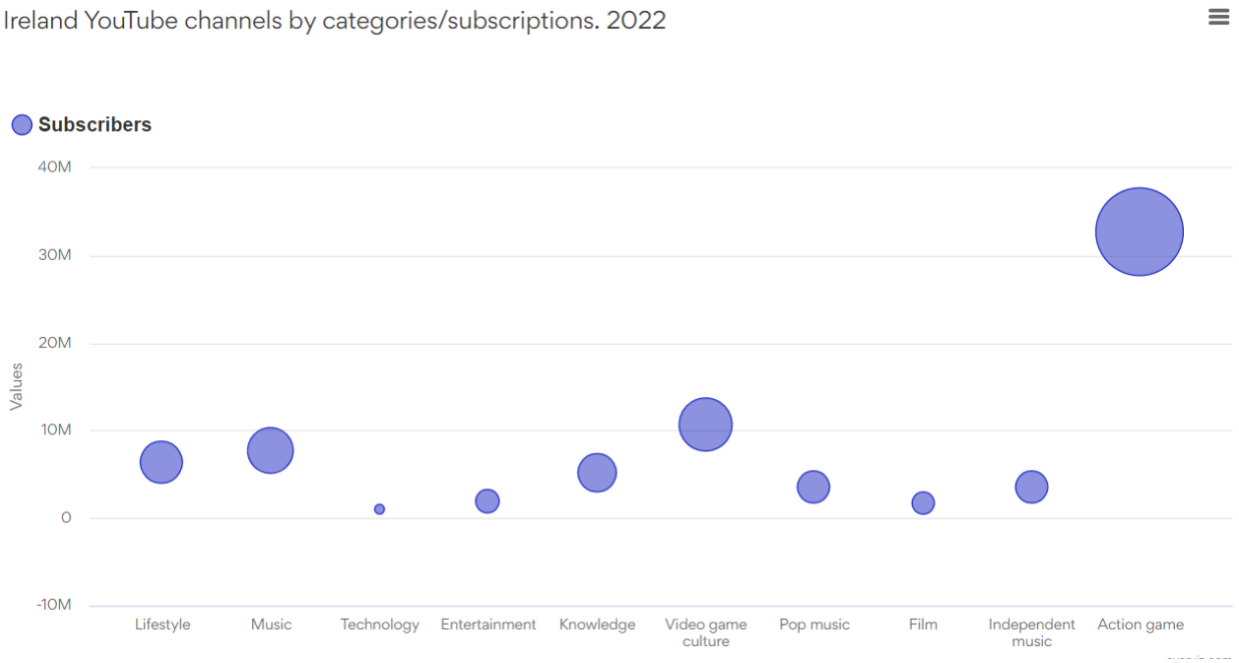


Table 8: YouTube channel categories by subscribers

YouTube channel categories by subscribers in Ireland in 2022		
Categories	Subscribers	Channels
Action game	32747060	7
Video game culture	10733910	11
Music	7690270	38
Lifestyle	6389240	49
Knowledge	5214750	12
Independent music	3599850	8
Pop music	3565410	11
Entertainment	1948700	13
Film	1697680	9
Technology	1023480	14

The most popular category in Ireland was Action game, which is different from Germany and France. The other four were Video game culture, Music, Lifestyle and Knowledge. However, there were more Gaming categories (Video game culture) and music categories (Independent music, Pop music) that would make these two categories' subscribers' numbers even bigger.

Contrary tour findings for France and Germany, the 'Film' category was not featured that much among Irish channels according to the number of subscribers.

Diversity Index. Estonia

In Estonia, the act amending the Copyright Act to include the provisions of the DSM Directive signed into law on the 20th of December 20In Estonia, 29 primary categories were used for identifying channels through wiki pages in 2022, and 24 in 2019.

Table 9. Distribution of channels in YouTube by category, Estonia, 2022

YouTube channels Estonia 2022		
Estonia	**0 - 1**	**0 – 10 000**
Categories 2022	**HHI index 0.142095**	**HHI index 1420**
Lifestyle (sociology)	19.7	**%**
Music	15.3	**%**
Vehicle	5.6	**%**
Technology	5.2	**%**
Video game culture	4.8	**%**
Pop_music	4.4	**%**
Entertainment	4.4	**%**
Electronic music	3.6	
Action adventure game	3.2	**%**
Sports	2.8	**%**

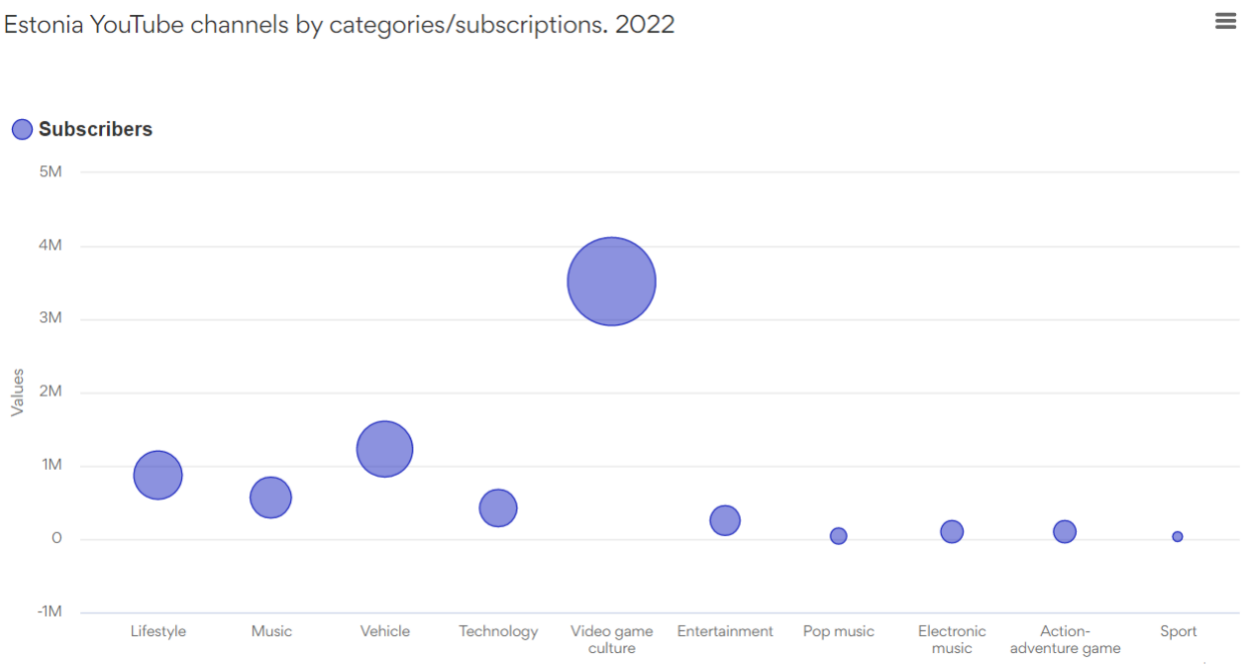
Table 10. Distribution of channels in YouTube by category, Estonia 2019

YouTube channels Estonia 2019		
Estonia	**0 - 1**	**0 – 10 000**
Categories 2019	**HHI 0.09612193**	**HHI 961**

Action game	15.3	%
Lifestyle	13.2	%
Entertainment	11.2	%
Gaming	8.3	%
Hobby	7	%
Pop music	5.4	%
Hip hop music	3.7	%
TV shows	3.3	
Technology	3.3	%
Society	3.3	%

The Herfindhal-Hirschmann index (HHI)in Estonia's YouTube channels has increased from 2019 to 2022. Its' index had an increase of 459 points, which is the highest among the four countries sample. This means that there is less cultural diversity in Estonia's YouTube channels production now than it had been then.

Graph 7: most popular YouTube categories by subscribers, Ireland, 2022



YouTube channel categories by subscribers in Estonia in 2022		
Categories	Subscribers	Channels
Video game culture	3519340	9
Vehicle	1229590	14
Lifestyle (sociology)	870470	53
Music	573200	27
Technology	424950	12
Entertainment	251190	6
Electronic music	106560	5
Action-adventure game	106030	5
Pop music	48990	6
Sport	27110	3

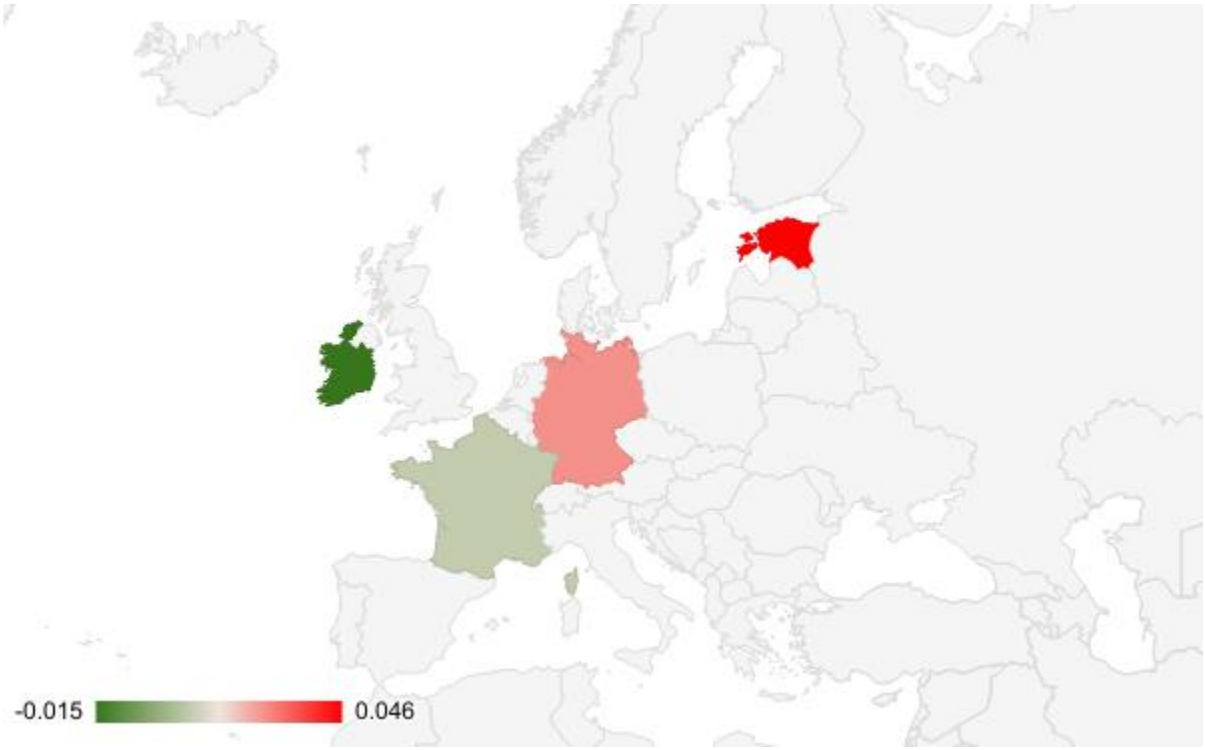
In Estonia, much like in Ireland, gaming category – Video game culture – was the leading category according to the number of subscribers. The other top four were Lifestyle, Music, Vehicle (which has appeared for the first time among all four samples) and Technology.

As in previous YouTube studies, (Bartl, 2017) we have also found a mismatch of supply and demand on YouTube in two out of four countries. In contrast to gaming, categories such as "society" and "knowledge" are characterized by a high number of channels and small numbers of subscribers in Ireland and Estonia. Thus we could say that the diversity consumed in Estonia and Ireland are similar and lower that that of Germany and France, which had a more normal distribution of channels and subscribers.

Four countries in a sample comparison of diversity index

According to the changes in Herfindhal-Hirschmann index (HHI) among the four countries in or sample, only Ireland had the diversity increased, France had it decrease very slightly, Germany more so, and Estonia the most. See Graph 5 which depicts the difference in diversity of YouTube channels (cultural products supply) of 2019 and 2022 years among the four countries.

Graph 5: differences in diversity 2019 – 2022: Ireland, France, Germany, Estonia



Country	Difference in Herfindhal-Hirschmann index (HHI) index
Estonia	0.04597307
Germany	0.02668197
France	0.00824886
Ireland	-0.01518726

Discussion and conclusion

One of the key observations presented in this research is the decline of diversity in the supply of cultural goods in YouTube during the recent years (2019 – 2022) in three out of four countries of the sample. In Estonia, the HHI index displayed the greatest change, and saw a decline of 459

points (out of 10 000) from the year 2019. In Germany, the index decreasedby 266 points in 2022 compared to 2019, in France by 82 points. Only in Ireland we saw a slight increase of diversity in the supply of YouTube channels (by 151 points). These change do not correlate directly with the implementationof Article 17 of the CDSM Directive and its timing, so the national legislation of the four countries under question seems to have had no influence onto the changes oberserved in diversity of YouTube channels. While Germany was the first to implement it, it was not Germany but Estonia, the last of the countries in question to implement the Directive, that saw the biggest decline in channels diversity. It also saw, along with Ireland, lower diversity consumed: i.e. more concentrated audience in few YouTube channel categories. Because the diversity of YouTube channels may be subject to a lot of variables, including not only policies such as the Article 17 CDSM implementation, but also YouTube's algorithmic and content policies, further research and availability of data from the platform is needed to address this change.

A second takeaway from this study and its comparison with the existing work is that there was an increase in blocked and deleted videos in all 27 countries of the European Union during the two years (2019 – 2022). Given the available data, it is hard to evaluate possible explanations for this raise, but the adaptation of the Article 17 of the CDSM might have played a role, either as a direct influence or even more convincing as an anticipation by the platform of more stricter regulation in the future. Running a dominance-analysis approach for comparing predictors of a video being blocked or deleted, we found that in both cases three of these predictors were consistent with previous research on copyright moderation on the YouTube (Gray & Suzor, 2020; Erickson and Kretschmer, 2018). These categories are film & animation, music, entertainment and gaming: all of these categories remain strong predictors for videos being blocked or deleted on the platform. We have also found that the scope of deleted and blocked videos has increased to 3.6% compared to previous studies which indicated 1% (Gray & Suzor: 2020). However, due to the new YouTube API restrictions on information concerning the reason for video removal, it is difficult to distinguish copyright and non-copyright related deletion.

Some of the research limitations concern the timeline of the study: while the policy was adopted officially by the European union in June 2019, actual implementation of Article 17 of the CDSM Directive is not yet fully in place in the countries under study, and it is possible that we could not

yet see its full-scale influence. Future and continuing research is needed to assess these questions, when the policy implementations become effectivevisible at full scale. It would also be of value to study and compare other platforms, either the largest platforms in the EU that are also subject to the stricter regulations of Art. 17 (e.g. Meta, TikTok) or smaller ones that cater to more niche audiences and that do not fall under Art. 17 obligations (e.g. Vimeo, Twitch). It will be interesting to see how these different types of platform develop in terms of their automated copyright moderation practices and its effect on cultural diversity once the CDSM Directive is fully implemented. Policy might actually have a performative effect here, in shaping platforms' characteristics by obliging them to report and monitor content much more closely. Another limitation of the study is the nature of YouTube's Content ID copyright moderation. It has been noted earlier that YouTube has had a robust and through system of copyright moderation for years (e.g. Venture, 2018). Its Content ID system has already been criticized for over blocking, and it may be difficult to say which of the automated copyright moderation evidence that we have presented had been in place for years without the willingness of the platform to give access to this kind of data.

The main takeaways from this research offer important implications both for research as well as for policymaker on platforms regulation.. Yet, current research options are highly limited and dependent on internal decisions of platforms. The main obstacle to our study was the limited number of daily API requests that we could make, thus, even small sampling of YouTube videos and channels made the data collection tedious and time consuming. In addition, the platform constantly changes its rules of what kind of data can and cannot be accessed via its API. Due to these arbitrary and opaque restrictions, researchers face uncertainty in what type of data they could and could not get, and it changes over time, so the previous research can not be continued and the new one has to start from scratch. As Jurgens & Stark highlight, "In addition to the theoretical challenges, the difficulty to empirically isolate individual curation mechanisms remains a frustratingly persistent problem. Without explicit assistance by platforms, researchers can only resort to creative designs make use of the occasional availability of insightful data" (Jurgens & Stark, p. 17).

Hence there is urgent t need for more robust rules on data access for researchers. Mandatory data access clauses such as those included in the German NetzDG, the German CDSM implementation as well as in the Digital Services Act pave an important avenue in this regard. Yet it remains to be seen how robust and effective these clauses are, since they demand highest levels of data security and infrastructure facilities on the side the researchers. Finding practical and fair solutions and best practices for data access that are not only accessible to researchers at elite and perfectly-equipped institutions is a key challenge for policy and research in the next few years. Only if we are successful with this, we can hope to fully understand the real costs and perils of platform governance and content moderation.

Bibliography

Access for all, a balanced ecosystem, and powerful tools. (n.d.). Blog.Youtube. Retrieved July 11, 2022, from https://blog.youtube/news-and-events/access-all-balanced-ecosystem-and-powerful-tools/

Alan, F. (2022). YouTube and Political Ideologies: Technology, Populism and Rhetorical Form. *Political Studies*, *70*(1), 62–80. https://doi.org/10.1177/0032321720934630

Article 17 survives, but freedom of expression safeguards are key: C-401/19 - Poland v Parliament and Council. (2022, April 26). Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2022/04/26/article-17-survives-but-freedom-of-expression-safeguards-are-key-c-401-19-poland-v-parliament-and-council/

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*(2), 129–148. https://doi.org/10.1037/1082-989X.8.2.129

Azen, R., & Budescu, D. V. (2006). Comparing Predictors in Multivariate Regression Models: An Extension of Dominance Analysis. *Journal of Educational and Behavioral Statistics*, *31*(2), 157–180.

Bärtl, M. (2018). YouTube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence*, *24*(1), 16–32. https://doi.org/10.1177/1354856517736979

Bar-Ziv, S., & Elkin-Koren, N. (2018). *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown* [SSRN Scholarly Paper]. https://papers.ssrn.com/abstract=3214214

Benhamou, F., & Peltier, S. (2007). How should cultural diversity be measured? An application using the French publishing industry. *Journal of Cultural Economics*, *31*(2), 85–107. https://doi.org/10.1007/s10824-007-9037-8

Brady, W. F.-C., Kirton, D., Sugrue, J., & Cullen, S. (2021, December 17). *Copy That! Ireland Implements EU Copyright Directive*. Lexology. https://www.lexology.com/library/detail.aspx?g=b36a11e2-cab6-425b-855c-9c53608819bc

Buccafurri, F., Lax, G., Nocera, A., & Ursino, D. (2014). Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences*, *256*, 126–137. https://doi.org/10.1016/j.ins.2013.08.046

Burk, D. L., & Cohen, J. E. (2000). *Fair Use Infrastructure for Copyright Management Systems* [SSRN Scholarly Paper]. https://papers.ssrn.com/abstract=239731

CDSM Implementation resource page – CREATe. (n.d.). Retrieved July 29, 2022, from https://www.create.ac.uk/cdsm-implementation-resource-page/

CJEU upholds Article 17, but not in the form (most) Member States imagined. (2022, April 28). Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2022/04/28/cjeu-upholds-article-17-but-not-in-the-form-most-member-states-imagined/

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, (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1625142238402&uri=CELEX%3A52021DC0288

Du, Y., Zhou, Q., Luo, J., Li, X., & Hu, J. (2021). Detection of key figures in social networks by combining harmonic modularity with community structure-regulated network embedding. *Information Sciences*, *570*, 722–743. https://doi.org/10.1016/j.ins.2021.04.081

Edwards, D. W. (2018). Circulation Gatekeepers: Unbundling the Platform Politics of YouTube's Content ID. *Computers and Composition*, *47*, 61–74. https://doi.org/10.1016/j.compcom.2017.12.001

Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/j.procs.2017.08.129

Erickson and Kretschmer (2018)—Copyright EVIDENCE. (n.d.). Retrieved July 11, 2022, from https://www.copyrightevidence.org/wiki/index.php/Erickson_and_Kretschmer_(2018)

Erickson, K., & Kretschmer, M. (2018). *"This Video is Unavailable": Analyzing Copyright Takedown of User-Generated Content on YouTube* [SSRN Scholarly Paper]. https://papers.ssrn.com/abstract=3144329

European Commission ignores civil society concerns and sides with entertainment industries. (2021, June 4). https://en.epicenter.works/content/european-commission-ignores-civil-society-concerns-and-sides-with-entertainment-industries

Farchy, J., & Ranaivoson, H. (2011). Measuring the Diversity of Cultural Expressions: Applying the Stirling Model of Diversity in Culture. *Undefined*. https://www.semanticscholar.org/paper/Measuring-the-Diversity-of-Cultural-Expressions%3A-of-Farchy-Ranaivoson/7514314a6d462408b4be77a44621123c79b9d089

Ferri, F. (2021). The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market. *China-EU Law Journal*, *7*(1), 21–38. https://doi.org/10.1007/s12689-020-00089-5

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945

Gorwa, R. (2021). Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG. *Telecommunications Policy*, *45*(6), 102145. https://doi.org/10.1016/j.telpol.2021.102145

Gray, J. E., & Suzor, N. P. (2020). Playing with machines: Using machine learning to understand automated copyright enforcement at scale. *Big Data & Society*, *7*(1), 2053951720919963. https://doi.org/10.1177/2053951720919963

Himelboim, I., Golbeck, J., & Trude, B. M. (2020). Chapter 13 - YouTube: Exploring video networks. In D. L. Hansen, B. Shneiderman, M. A. Smith, & I. Himelboim (Eds.), *Analyzing Social Media Networks with NodeXL (Second Edition)* (pp. 187–203). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-817756-3.00013-3

Jacques, S., Garstka, K., Hviid, M., & Street, J. (2017). *The Impact on Cultural Diversity of Automated Anti-Piracy Systems As Copyright Enforcement Mechanisms: An Empirical Study of YouTube's Content ID Digital Fingerprinting Technology* [SSRN Scholarly Paper]. https://doi.org/10.2139/ssrn.2902714

Keller, P. (2021, December 9). *YouTube Copyright Transparency Report: Overblocking is real*. Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2021/12/09/youtube-copyright-transparency-report-overblocking-is-real/

Kurant, M., Markopoulou, A., & Thiran, P. (2010). On the bias of BFS (Breadth First Search). *2010 22nd International Teletraffic Congress (LTC 22)*, 1–8. https://doi.org/10.1109/ITC.2010.5608727

Kurdi, M., Albadi, N., & Mishra, S. (2020). "Video Unavailable": Analysis and Prediction of Deleted and Moderated YouTube Videos. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 166–173. https://doi.org/10.1109/ASONAM49781.2020.9381310

Lévy-Hartmann, F. (2011). Une mesure de la diversité des marchés du film en salles et en vidéogrammes en France et en Europe. *Culture Méthodes*, *1*, 1. https://doi.org/10.3917/culm.111.0001

Matamoros-Fernandez, A., Gray, J. E., Bartolo, L., Burgess, J., & Suzor, N. (2021). WHAT'S 'UP NEXT'? INVESTIGATING ALGORITHMIC RECOMMENDATIONS ON YOUTUBE ACROSS ISSUES AND OVER TIME. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2021i0.12208

McDonald, D. G., & Lin, S.-F. (2004). The Effect of New Networks on U.S. Television Diversity. *Journal of Media Economics*, *17*(2), 105–121. https://doi.org/10.1207/s15327736me1702_3

McGrady, R. (n.d.). *Research Note: YouTube Is Infrastructure YouTube may have started as a platform, but it has become infrastructure.* https://www.mediaecosystems.org/explorations/research-note-youtube-is-infrastructure

Mihelj, S., Leguina, A., & Downey, J. (2019). Culture is digital: Cultural participation, diversity and the digital divide. *New Media & Society*, *21*(7), 1465–1485. https://doi.org/10.1177/1461444818822816

Moreau, F., & Peltier, S. (2004). Cultural Diversity in the Movie Industry: A Cross-National Study. *Journal of Media Economics*, *17*(2), 123–143. https://doi.org/10.1207/s15327736me1702_4

Moving from social networks to social internetworking scenarios: The crawling perspective | Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/j.ins.2013.08.046

Paolillo, J. C. (2008). Structure and Network in the YouTube Core. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. https://doi.org/10.1109/HICSS.2008.415

Perel, M., & Elkin-Koren, N. (2018a). Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. *Florida Law Review*, *69*(1), 181.

Perel, M., & Elkin-Koren, N. (2018b). Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. *Florida Law Review*, *69*(1), 181.

Article 17 survives, but freedom of expression safeguards are key: C-401/19 - Poland v Parliament and Council. (2022, April 26). Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2022/04/26/article-17-survives-but-freedom-of-expression-safeguards-are-key-c-401-19-poland-v-parliament-and-council/

Ranaivoson, H. (2010). *The Determinants of the Diversity of Cultural Expressions—An International Quantitative Analysis of Diversity of Production in the Recording Industry*.

Rankin, J. (2018, June 20). EU votes for copyright law that would make internet a “tool for control.” *The Guardian*. https://www.theguardian.com/technology/2018/jun/20/eu-votes-for-copyright-law-that-would-make-internet-a-tool-for-control

Roberts, S. T. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Rieder, B., Coromina, Ò., & Matamoros-Fernández, A. (2020). Mapping YouTube: A quantitative exploration of a platformed media system. *First Monday*. https://doi.org/10.5210/fm.v25i8.10667

Suen, W., Anderson, C., & Swimmer, G. (1997). An empirical analysis of viewer demand for U.S. programming and the effect of Canadian broadcasting regulations. *Journal of Policy Analysis and Management*, *16*, 525–540. https://doi.org/10.1002/(SICI)1520-6688(199723)16:43.0.CO;2-I

The DSM Directive two years on: Is it in force? | Simmons & Simmons. (n.d.). Retrieved July 29, 2022, from https://www.simmons-simmons.com/en/publications/cktct3gfi1tls0b02r8s9vgfb/the-dsm-directive-two-years-on-is-it-in-force-

Tushnet, R. (2014). All of This Has Happened before and All of This Will Happen Again: Innovation in Copyright Licensing. *Georgetown Law Faculty Publications and Other Works*. https://scholarship.law.georgetown.edu/facpub/1459

Urban, J. M., Karaganis, J., & Schofield, B. (2017). *Notice and Takedown in Everyday Practice* [SSRN Scholarly Paper]. https://doi.org/10.2139/ssrn.2755628

van Dijck, J. (2020). Governing digital societies: Private platforms, public values. *Computer Law & Security Review*, *36*, 105377. https://doi.org/10.1016/j.clsr.2019.105377

Welle (www.dw.com), D. (n.d.-a). *EU targets tech giants with new copyright deal for digital age | DW | 14.02.2019*. DW.COM. Retrieved July 9, 2022, from https://www.dw.com/en/eu-targets-tech-giants-with-new-copyright-deal-for-digital-age/a-47511311

Welle (www.dw.com), D. (n.d.-b). *Thousands in Berlin protest EU's online copyright plans | DW | 02.03.2019*. DW.COM. Retrieved July 9, 2022, from https://www.dw.com/en/thousands-in-berlin-protest-eus-online-copyright-plans/a-47753399

YouTube: We've invested $100 million in Content ID and paid over $3 billion to rightsholders. (2018, November 7). *VentureBeat*. https://venturebeat.com/2018/11/07/youtube-weve-invested-100-million-in-content-id-and-paid-over-3-billion-to-rightsholders/

(N.d.).

Losing authenticity: social media creators' perspective on copyright restrictions in the EU

Daria Dergacheva, Christian Katzenbach, Paloma Viejo Otero
University of Bremen

Submitted to ICA 2023

November 2022

Abstract

Artists who try to promote their cultural production on social media platforms today have to navigate the algorithmic environment which relates not only to the visibility and content distribution but also the increasing activity of algorithmic copyright moderation. Social media creators and users in the EU saw a rise in algorithmic copyright moderation when the Article 17th of CSDSMD was passed and came into force in the nation states (2021). The focus of this research is on creators' understanding and experiences of copyright moderation in relation to their creative work and the labor of media production on social media platforms. The study draws upon semi-structured interviews with 14 artists from various EU countries. The sample was drawn from those artists who participated in the survey on digitization and digital access to cultural content. We found that the anticipation of platform punishments directly influenced the cultural products that they produced. Indeed, most of them used self-censorship, avoidance and engaged in content adjustment in their creative work before posting it on social media platforms. The research draws on a multimodal framework to analyze the copyright governance of creative practices and products, focusing on regulative dimension (in our case – adaptation of the Article 17(4) CDSMD)), normative dimension (prevalent assumptions about legitimate and illegitimate behavior in a specific community) and the influence of technological affordances relevant to creative work.

Introduction

"YouTube and its users face an existential threat from the EU's new copyright directive" was one of the dozens similarly apocalyptic headlines which appeared in the spring of 2019 in European and US media (Feiner, 2019, October 14th). Indeed, worries about the Article 17 CDCS Directive implementation took a lot of people to the streets in Germany (Martin, 2019, March 2rd), bothered platforms to the point of saying they would 'put up a fight' (Feiner, 2019, October 14th) and created fears that the users would flee elsewhere (ibid). The new directive, which in simple worlds made the tech giants such as YouTube and Meta liable for copyright infringement by their users (in the EU), had a lot of people worried that it would literally kill the diversity of users and contents.

The step that was taken by the EU legislators has drastically changed the regulatory aspects of platform governance. In the previous years, the rhetoric used by platform stakeholders was one claiming that they are not media companies, thus allowing them to evade liability for user generated content (Gorwa, 2019) . This rhetoric was supported by regulation as well: since most of the social media platforms are private companies based in the US, they were regulated by § 230 of the Communications Decency Act18 (CDA), which gives online intermediaries broad immunity (Klonick, 2018).

In the EU, they were regulated under the "safe harbour" rules of the Directive on electronic commerce (E-Commerce Directive), thus were not liable for acts of their users committed through their services, in case they did not have knowledge about them (Geiger & Jütte, 2021).

In spite of this, social media platforms have created their own robust system of moderation prior to Article 17 implementation: YouTube alone has invested $100 plus million in its content filtering system since its inception with its Content ID, whose purpose is precisely to identify copyrighted content (YouTube: We've invested $100 million in Content ID and paid over $3 billion to rightsholders, 2018, November 7). Meta's Facebook was reported stating to the EU Commission that they believed there were two sets of laws: private law (Facebook community standards) and public law (legislations) (Politico, 2017).

Thus, producers of cultural content on social media platforms today have to navigate the algorithmic environment which relates not only to the visibility that its content creates (Bucher, 2017; Bishop, 2019) and content distribution (Hallinan and Striphas, 2016; Wilson, 2017), but also the increasing activity of algorithmic copyright moderation (Gray & Suzor, 2020).

This study investigates algorithmic copyright moderation in post-Article 17th Directive environment through experiences of producers of cultural content on social media platforms in the EU Member States.

One year into the Directive's implementation, platforms did not see the mass exodus of users as of yet. But what about the experiences of users themselves, especially those that create the valuable content of the platform economy? Being regulated by platforms is not a new experience for them: YouTube itself has imposed various rules on monetization and content (Caplan & Gillespie, 2020); both YouTube and Meta had already had in place robust systems of automated copyright detection. Now, it seems, that policymakers in the EU have added an additional layer to the surveillance of the creative work. Indeed, how did the top-down approach to platform governance, efforts of the states to improve it for those creating the content (Cunningham and Craig (2019), has manifested itself in creators' experiences?

Creators on social media platforms have to constantly be involved in pursue of algorithmic visibility measured in 'platformed indices' of visibility (i.e. likes, views, shares) (Duffy & Meisner, 2022; Bucher, 2017). Platforms use 'whims' of content moderation, regulations set by platforms themselves (and those under government policy influence), evoking the threat of 'invisibility', and this was described before as being 'dangerous' for creators (Cunningham and Craig (2019))

How much authenticity does this leave in the creative work? Is it still the work that we would have seen without regulations (by platforms and of platforms) or is it something completely different? The focus of this article is thus on creators understanding and experiences of copyright moderation in relation to their creative work and the labor of media production on social media platforms. To what extent does copyright moderation on the former influence the creations that are posted there? What about the changes to one's creative process? In order to answer these questions, the article uses creator's experiences and descriptions of their interaction with

copyright moderation and algorithms. The goal is to provide a ground of understanding the changes and influences that automated copyright moderation brings to creative artistic work. "What may have begun as a "partner" revenue sharing arrangement, a bonus offered to already motivated and prolific creators, has in practice set the terms for the labor of media production at YouTube, imposing specific expectations for users who count on that revenue." (Caplan, 2022, p.). Social media users that we discuss in this paper are also sometimes relying on platforms for their income; however, they are not producing media that would bring them revenue, but rather, use the platforms for promoting their creative work done elsewhere. They are not the ones doing a career our of creating for social media, (with one exception) but they still engage in creative labor for these platforms (Duffy & Meisner, 2022) in order to promote the creative work they are doing outside (or inside) of them.

Indeed, our definition of creators in this article is not as business entrepreneurs who receive some form of renumeration the major social media platforms (Cunnungham and Craig, 2019) but rather artists whose workload consists of marketing their creations on social media platforms, among other things that they do for living.

In this article, we will examine the attitude of creators to the copyright landscape that they have to navigate on social media platforms. Our research question here is "How the automatic copyright moderation on social media platforms influences the creative work?"

The present qualitative study aims to look at the views of content creators and how the Article 17 regulation have affected their production in the last two years.

The study draws upon semi-structured interviews with 14 artists from various EU countries. The sample was drawn from those artists who participated in the survey on digitization and digital access to cultural content, done by Proost & ____ (2022) in the context of a large EU-wide project The artists interviewed used a vide range of social media platforms: Instagram, Facebook, TikTok, YouTube, Behance, Etsy, LinkedIn, Vimeo, Pinterest and Dailymotion. For many interviewees, anticipation of platform punishments directly influenced the cultural products that they produced. Indeed, most of them used self-censorship, avoidance and concerted efforts to circumvent algorithmic intervention. in their creative work before posting it on social media

platforms. Our research has also confirmed previous findings on user folk theories regarding algorithms (De Vito et al., 2017) and algorithmic gossip, searching for the shared meaning of algorithmic moderation practices (Natale, 2019; Bishop, 2020). We have found evidence of "exploitation, insecurity, and a culture of overwork" (Duffy & Meisner, p. 1) among those artists whose work and income relates to social media platforms but not depends on it fully.

The research draws on a multimodal framework to analyze the copyright governance of creative practices and products, focusing on regulative dimension (in our case – adaptation of the Article 17(4) CDSMD)), normative dimension (prevalent assumptions about legitimate and illegitimate behavior in a specific community) and the influence of technological affordances relevant to creative work. (Katzenbach, 2018).

Literature review

One of the main focuses of the research on platform governance of the later years has been on algorithmically driven decision-making mechanisms and how content creators perceive how algorithmic moderation affects their creative production (e.g. see De Vito et al. (2017) on algorithmically-driven content curation and user resistance; West (2018) on folk theories of content moderation; Savolainen (2021) on algorithmic folklore and shadow banning; Duffy & Meizer (2022) on algorithmic invisibility).

The scholarship on 'critical algorithm studies' (Gillespie and Seaver, 2015) studied the major role that algorithmic decision-making has in shaping society and culture. Critical assessment of the decision-making mechanisms in platform studies has shown that the process is not transparent and easily understood by creators or understood at all (Eslami et al. (2015), Gillespie, 2018; Poell et al., 2021).

Bucher (2017) talks about affective dimensions of algorithm – of how it makes people feel, thus analysing algorithmic imaginaries. In earlier research, she studied algorithmic ranking and

visibility on social media, with algorithm establishing a participatory norms through validation and punishment (Bucher, 2012). Influence of the platforms' governance mechanisms on user behavior was further investigated by (Bucher and Helmond, 2018 and Weltevrede and Borra, 2016).

Other studies on algorithmic visibility (e.g. Bishop, 2019) highlighted algorithmic gossip as a way for the creators community engage with algorithms, and Abidin (2016) introduces a concept of "visibility labor" of Instagram creators. Caplan & Boyd (2018) investigate institutional (media industry's) dependency on algorithmic intermediaries. Willson (2017) and Hallinan and Striphas (2016) explore the place of algorithmic content distribution in contemporary 'everyday'.

More studies have been published recently which study the ways creators and influencers deal with the algorithmic systems of the platforms. Caplan & Gillesebey (2019) examine how YouTube creators develop their own theories about demonetization of their content. Bishop (2020) studies so-called YouTube "experts" who claim to help users "mitigate the risk of algorithmic invisibility"; Cotter (2021) conceptualizes 'practical knowledge' on algorithms through case study of a YouTube community's practices. The main concern of the study by Cunningham and Craig (2019) is that creators are not recognized as stakeholders in contemporary academic or policy debates on platform governance. Duffy et al., 2021 map out methods of studying platformization of the cultural industries. "Algorithmic skills' of enterpreners on platforms are discussed in the study by Klawitter and Hargittai, (2018).

This study contributes to the field of platform governance studies and cultural products creation on social media platforms. Following the previous research, we are providing here an insight into how platforms' treatment of users "can be understood within a framework of institutional power" (Duffy & Meisner, p. 17). In our study we are considering how cultural production is changing in the eyes of artists who are simultaneously social media creators, influenced by governmentality and surveillance, thus adding to the fields of platform studies and the influence of automated algorithmic content moderation on copyright on cultural production

Methods

The design of this survey was such that only those creators who agreed to be interviewed after a large-scale EU survey on artists experiences with digital sphere (Poort & Pervaist, 2022) took part in the research. The present paper focuses on a set of 14 semi-structured interviews conducted between May 2022 and July 2022 withntent creators We interviewed 14 creators from the following countries of the European Union: The Netherlands, Bulgaria, France, Romania, Croatia, Czech Republic, Portugal, Estonia and France. Most of them were visual artists (painters, animators, photographers, illustrators), and one of them had vlogging (on YouTube, educational art topics) as a full-time profession. For others, the use of social media platforms, although associated with their creative work, was more used as a marketing and visibility tool.

This sub-set is derived from the set of participants of a large survey on creative practiced done by Poort & Pervaist, (2022), in the context of the EU in relation to digitization and digital access to cultural content, mainly done through a multilingual survey.

A multimodal framework to analyze the copyright governance of creative practices and products (Katzenbach, 2018) was adopted in the thematic coding of all content. We coded our data set in in an instrumental sense, namely to summarize, identify and organize themes in the corpus of interview transcripts according to the adapted framework.

Interviews followed a semi-structured interview protocol; and lasted between 30 and 90 minutes. Participants received a gift card ($50) in exchange for their time and insight, and interviews were conducted on Zoom. All the interviews were recorded after acquiring consent for this. After the completion of the interviews the audio was transcribed and edited for any discrepancies. From the transcripts, the study's authors developed the coding categories and applied focused codes to the dataset. Throughout the process, we followed a thematic analysis approach (Terry et al, 2017)

Results & Discussion

This study has found the following themes in creators experiences with authomated copyright governance on social media platforms. First of all, there was no understanding of automatic copyright moderation among the artists, thus they used user folk theories regarding algorithms (De Vito et al., 2017) and algorithmic gossip, searching for the shared meaning of algorithmic moderation practices (Natale, 2019; Bishop, 2020). Thus, there were problematic assumptions of content copyright moderation as the first theme. The second theme relates to the cultural content they are producing and how they engage in self-censorship, avoidance and concerted efforts to circumvent algorithmic intervention (Duffy & Meisner, 2022) in their creative work before posting it on social media platformsThe last themes were related to technological dimensions, and they showed that the copyright moderation is also a question of algorithmic gossip (Bishop, 2020), but in reality the creators did face more copyright moderation during the last two years than they did before, and they did not find the appeal process that platforms put in place either helpful or transparent.

Normative dimension: problematic assumptions of content copyright moderation

There is no understanding of the regulatory mechanisms which work on social media platforms among the creative content producers. Assumptions on the 'right' and 'wrong' practices differ, and usually do not correspond to legal realities, thus questioning the 'regulative dimension' of the provision and enforcement of formal rules. A tattoo artist from Croatia, who actively uses Instagram for her creative work, says:"I think every time an artist wants to show their work, if they put it online, it's the moment we put it online, it's we kind of lose a bit of it's not only ours anymore, it's if we want the world to see our work, then we have to accept that someone else might like it enough to steal it. And that's just the way it is. So, I don't think it can be fixed." Another interviewee, a visual artist from Portugal, is unsure when the 'laws' changed. She, like many other interviewees, also confuses the rules and regulations that platforms set for their users with the legal obligations (laws): "Because the law, the laws regarding music, in YouTube videos and posts have changed in the last couple of years, or, like four years. And yes, like I said people were used to having like, no copyrights associate to videos. And then suddenly, like four years ago, three years ago, everything changed".

Consistent with previous studies on understanding the algorithms, creatives on social media platforms hey often gain knowledge on issues of copyright through 'algorithmic gossip' (Bishop, 2019). "I don't know much actually. Like what I know. It's from like stories, like close stories, or even personal stories never like an actual study about it. And like an actual thing explaining is what and how we can do stuff." Learning by mistakes is also the common practice: "I think it's, it's not that explained. For us. We just learn by mistakes for example, putting a song and then understanding that we can or we learn by what fellow colleagues tell us never like I never came across like an actual workshop or seminar, anything like that", - adds another interviewee. Creative from Spain explains that he does not know anything but two aspects of automated platform moderation: "I don't think I'm aware of it. The only thing that I know that it's that they don't allow like very explicit images like sexual very sexual images. And yeah, that's all know. And also copy. Coping someone's work. I'm not sure if because I see a lot of the time someone copying someone's work without permission before anyone notices actually and reports it. "Bulgarian visual artist agrees with him: "Like, for example, not use other people's content in terms of music, and I don't know everything pictures without crediting them. And that's pretty much it."They often feel that the lack of knowledge of automated algorithmic copyright moderation holds them back from doing other forms of creative work: "That's something I want to make more clear for myself. Because I want to have music for background for my stories, or that people use a lot of different music. And I always think that, how, like, is this allowed? Or how does it work?", - says a visual artist from Estonia. A visual artist from Greece added: "I don't have reels, for example because of the copyright laws, because the reels need music to be able to play. And sometimes I watch a video on mute, it's a very uncomfortable viewing of the feed. But it means a lot of work for me to figure out which one will go with which platform and which algorithm will suppress it, because it's copyright violated. So even if you have read the rules, it feels a bit like a guessing game. Yes, at the moment, with the laws in Greece."This observation leads us to another theme that has been evident in the interviews. Does regulatory dimension of platforms copyright rules and legal obligations together with technological dimension of copyright algorithmic moderation hold creators back from creating? We have found out that it does.

Technical dimension influencing normative: self-censorship, cultural content adjustment due to algorithmic copyright moderation

We investigated the attitude to algorithmic content moderation in this study. The findings were that creators, even if they themselves did not experience the moderation, are anticipating it and thus adjust the content beforehand. Bulgarian visual artist said: “the music I've had (to adjust) too, especially when we're like shooting from shows and we have music playing at the shows. I have to… either delete or distort the whole sound of the videos that I'm posting. Because I've heard people (having problems because of that). Well, it's not what I would wish to do since it is influencing the work. I don't know how to explain it well, it changes the content when you change the music. And I think I've done it once or twice with paintings, because they have been like, replicas are very close to somebody else's work. So I would change them a bit for social media.”

Several other interviewees have talked about issues with the music that they have to “solve” before posting anything to social media“I wasn't used to it at first and then I had to adjust a few things around like for example, asking my friends that have bands or, or my boyfriend who is a composer to do like, his own thing, to send me his own music. So that I don't have any copyright issues”, - said a creator from Romania. She added that she did not know how others were coping with this strain, as not many people have “boyfriends who write music”. A creator from France added that: “when I need to have like cover music, I post my own or like I post one without copyright claims. So I'm very careful about not having any (issues) with copyrights”.Some interviewees confessed to quitting posting certain products to platforms all together, since they were not sure about copyright moderation:“I stopped uploading videos on Facebook because I don't know how to make them interesting without all the filters and music and everything so I when I post video I always post it only on Instagram”. Instagram, they add, have a special “list” of music that one could use without the copyright claim, thus making it on the one hand easier for creators but on the other hand governing the creative process by assigning pre-approved music. “I think we had a Lip Sync Battle with gallery workers. And we sang a very popular song. And we uploaded it because it was like, a promo with gallery workers. And it's got taken down in matters of an hour, I think. Yes, it was Instagram”, - said a participant from Bulgaria.After this, she only used the “approved” music from the list. “I don't upload my own music. Yeah, I just use

the music that is there, because it would take too much time. And I think if I upload my own music, they will recognize, and you will get an email that it's the copyright and a scam, or you know, they will delete it. Or unless you compose the music yourself, obviously, then you're good, but I'm not a musician.", - said another interviewee.An artist from the Netherlands, who also uses the pre-approved music for 'reels' on Instagram, said that this influences his creative process in a direct way: "I think it's a collaboration, more than stealing someone's copy rights or something. Like for example, my art is many many times inspired by vintage or classical stuff. So, I use like music from the 1920s or 1930s. For example, last time, I drew a Marilyn Monroe, obviously, many artists do that. But you will put in that reel… you will put the song of Marilyn Monroe, right? To enchant it, to highlight it. And that's it."In the next section, first of all, we discussed which cases of copyright moderation did the creators encounter. Secondly, we were interested whether the instances of such moderation have increased in the last year. And finally, it was important for us to find out what were the practices of platforms in terms of appealing content moderation based on copyright issues.

Process: how platforms block and what types of creations they block

Nine out of fourteen interview participants either had their creations taken down for copyright infringement or knew someone who had. One of the creators from the Netherlands remembered: "I remember once or twice using, pieces from movies to post on Instagram, like once, something from a generally good movie from the 60s. And strangely, it was blocked. And it's because within that movie there was a soundtrack is by The Rolling Stones, and I guess the algorithm works really well. Although it really not featuring anything related to the Rolling Stones. I tried several times to upload it. And it was immediately removed." An illustrator from Portugal described similar situation with the colleague: "I had a colleague of mine who does posters for rock bands. And he uploaded a video from the drop of his most current poster, and on the video, music from the band was playing in the background. And not in the background of the music, like it wasn't edited in. It was from the video itself, like it was playing live where we shot it. And he got taken down, even though he works with the band."

Timing: did copyright moderation increase during the last year in the views of creators

When asked directly, most of the participants did not think that moderation due to copyright issues increased during the last year. However, when they remembered instances of such moderation, often their cases were from not so distant past, such as “Last Christmas” (Lip Sync video, creator from Bulgaria), “last year” (reel, creator from Croatia), “during the past three or four years” (video, creator from Romania) . Another detail is that some interviewees had their old videos or posts taken down not long ago: “I had a show reel from 2016. That then I had to change the music when that law started to be applied in YouTube. They sent me something, but they blocked it.”Process: how complains and appeals work on social media platforms

Our participants in general have found the appeal, report and complain processes on platforms not very helpful. Sometimes they have to use network of friends and followers in order to solve the issue, so not directly complaining to a platform but using other mechanisms:“It was very difficult of a process to solve. I had a artists friend who was catfished. So someone created an account, sharing her work as if it was by them on Instagram. And then, she asked everyone to report it. And then we all (her friends and followers) reported and the account was taken out.”Others do not appeal at all:“I didn't appeal because I thought that it was in there, right. Like the song wasn't ours. And we were only lip syncing. So we weren't singing over it. Or we were only singing over, like some parts. So I thought that it's, you know, stupid, but fair, I guess”, - said a Bulgarian artist.

Another artist from Croatia remembered how she had no idea of why her video was taken down so she had to use Google to find out what was happening since she never got any answers. She did not appeal afterwards, too. “I was struggling with it for a few days before I realized what happened. It just saw my video was just taking down. And there was a note saying, like a little yellow sign, that something was wrong. And then it said that the video cannot be played right now. No other explanation. I tried to write and I tried to upload it again. But nothing happened. Then I Googled why”Another one gave up on appealing after reading an article (not on the platform) which explained that it would not “make sense” for platforms to hire someone to respond to all the questions. So she “just stopped expecting the answer every time I have a problem.”

Regulative dimension: do the laws work?

The European Copyright in the Digital Single Market (CDSM) Directive (2019/790) has been adapted and came into force in June 2019. Countries had the time period of two years to implement it within their national law but almost one year after the deadline, on May 19th, 2022, the EU Commission sent out a press-release saying that Belgium, Bulgaria, Cyprus, Denmark, Greece, France, Latvia, Poland, Portugal, Slovenia, Slovakia, Finland and Sweden have not yet notified the Commission on changes to national law. However, some creators do not think that the legislation works. Even those from the countries which have adapted the Directive, are either not aware or have not seen the laws put into action when it comes to protection of their own work.An artist from Romania said: "My opinion on platform copyright regulation is that it is hectic and illogical for things that I would like to be copyrights regulated, and should be like, obviously regulated, since it's like, art theft or something like that. Then it's not regulated at all, because the content was changed, so the platform doesn't care. I find it's just confusing and illogical, and I don't find it that it works. Like when it should work, it doesn't and when it shouldn't work, it does."A Greek painter thinks that his creations are not protected via copyright on social media platforms:"Let's say, I drew a picture, and somebody copied my image and posted it on another social media. I can't track that, you know. That's the issue, I think. And I don't want to track every social media to see what's going on. I'm gonna waste a lot of time, you know, doing that."During our research we have found out other issues that worry creators in relation to their work on social media platforms. Themes related to visual labor (or "algorithmic treadmill", as one of the participants described it); benefits and losses that the creators experience in relation to social media; issues of algorithmic moderation which does not relate to the copyright have all come up in our research. However, we are not including them in this paper in order to keep the main focus of the current study on copyright moderation.

Conclusion

For a number of years researchers have expressed concerns about algorithms exercising too much power in influencing social realities (Beer, 2009; Gillespie, 2014; Kitchin and Dodge, 2011). In addition, algorithmic content moderation on private platforms has been compared to a "black box" (Perel & Elkin-Koren, 2017). "There are multiple sources of opacity – institutional, legal and

technological – that make it difficult to evaluate automated private regulatory systems." (Gray & Suzor: 2020).

As a result of our study, we have found that users of social media platforms which do creative work on those, are influenced by algorithmic content moderation. Perhaps our most important finding which extends understanding on how algorithmic content moderation influence creative work on platforms, is that creators engage in self-censorship, avoiding posting certain content or adjusting it in advance.For many artists, anticipation of platform punishments directly influenced the cultural products that they produced. In addition, because the regulative dimension of algorithmic copyright moderation is opaque for creators, they engage in algorithmic gossip (Bishop, 2019) and use user folk theories (De Vito et. Al, 2019) trying to guess which practices are accepted and which are not. We have also found examples of "exploitation, insecurity, and a culture of overwork" (Duffy & Meisner, p. 1) among those interviewed.

Thus, the normative dimension of copyright governance on social media platforms experiences changes due to the uncertainty of regulative dimension, which may lead to decrease in cultural production on social media platforms.

We have also found that the technological dimension of appeal and complaint processes on social media platforms is currently not very helpful for the creators. In terms of timing, it is difficult to say whether social media platforms have started to implement more algorithmic copyright moderation after the nation states of the EU approved European Copyright in the Digital Single Market (CDSM) Directive (2019/790). some interviewees had their old and new videos or posts taken down during the last year due to the copyright moderation.

For artists engaging in creative work on social media platforms, there are multi[le uncertainties about automated content moderation which leads to changes and self-censorship of their creative products. There are important policy implications from this research, such as more transparency in platform governance, both from policy makers and the tech giants themselves. As a conclusion, we will use the words of Croatian artist that we interviwed: "I would have more actual people look at copyright moderation, because when it's just a computer or an algorithm,

scanning the internet for repeating sounds and repeating images, it doesn't really work and content gets taken down, when it has the right to be used."

Bibliography

Abidin, C. (2016). Visibility labour: Engaging with Influencers' fashion brands and #OOTD advertorial campaigns on Instagram. *Media International Australia*, *161*(1), 86–100. https://doi.org/10.1177/1329878X16665177

Andersen, J. (2020). Understanding and interpreting algorithms: Toward a hermeneutics of algorithms. *Media, Culture & Society*, *42*(7–8), 1479–1494. https://doi.org/10.1177/0163443720919373

Article 17 survives, but freedom of expression safeguards are key: C-401/19 - Poland v Parliament and Council. (2022, April 26). Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2022/04/26/article-17-survives-but-freedom-of-expression-safeguards-are-key-c-401-19-poland-v-parliament-and-council/

Bärtl, M. (2018). YouTube channels, uploads and views: A statistical analysis of the past 10 years. *Convergence*, *24*(1), 16–32. https://doi.org/10.1177/1354856517736979

Bechky, B. A. (2003). Sharing Meaning across Occupational Communities: The Transformation of Understanding on a Production Floor. *Organization Science*, *14*(3), 312–330.

Bishop, S. (2019). Managing visibility on YouTube through algorithmic gossip. *New Media & Society*, *21*(11–12), 2589–2606. https://doi.org/10.1177/1461444819854731

Bishop, S. (2020). Algorithmic Experts: Selling Algorithmic Lore on YouTube. *Social Media + Society*, *6*(1), 2056305119897323. https://doi.org/10.1177/2056305119897323

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, *14*(7), 1164–1180. https://doi.org/10.1177/1461444812440159

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Caplan, R., & boyd, danah. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 2053951718757253. https://doi.org/10.1177/2053951718757253

Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society*, *6*(2), 2056305120936636. https://doi.org/10.1177/2056305120936636

Copyright law could put end to net memes. (2018, June 8). *BBC News*. https://www.bbc.com/news/technology-44412025

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, *21*(4), 895–913. https://doi.org/10.1177/1461444818815684

Cotter, K. (2022). Practical knowledge of algorithms: The case of BreadTube. *New Media & Society*, 14614448221081802. https://doi.org/10.1177/14614448221081802

Cunningham, S., & Craig, D. (2019). Creator Governance in Social Media Entertainment. *Social Media + Society*, *5*(4), 2056305119883428. https://doi.org/10.1177/2056305119883428

Custodians of the Internet. (n.d.). *Yale University Press*. Retrieved July 29, 2022, from https://yalebooks.yale.edu/9780300261431/custodians-of-the-internet

DENARDIS, L. (2014). *The Global War for Internet Governance*. Yale University Press. https://www.jstor.org/stable/j.ctt5vkz4n

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017). "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3163–3174. https://doi.org/10.1145/3025453.3025659

Devito, M., Birnholtz, J., & Hancock, J. (2017). Platforms, People, and Perception: Using Affordances to Understand Self-Presentation on Social Media. *Proceedings of the Annual ACM Conference on Computer Supported Cooperative Work and Social Computing*. https://doi.org/10.1145/2998181.2998192

Doctorow, C. (2019, February 13). *The Final Version of the EU's Copyright Directive Is the Worst One Yet*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2019/02/final-version-eus-copyright-directive-worst-one-yet

Duffy, B. E., & Meisner, C. (2022). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, 01634437221111923. https://doi.org/10.1177/01634437221111923

European Commission back-tracks on user rights in Article 17 Guidance. (2021, June 4). Kluwer Copyright Blog. http://copyrightblog.kluweriplaw.com/2021/06/04/european-commission-back-tracks-on-user-rights-in-article-17-guidance/

Feiner, L. (n.d.). YouTube and its users face an existential threat from the EU's new copyright directive. CNBC. Retrieved October 31, 2022, from https://www.cnbc.com/2019/05/10/youtube-faces-existential-threat-from-the-eus-new-copyright-directive.html

Foundation, L. and I. (n.d.). *Article 17 of the Directive on Copyright in the Digital Single Market and why it "broke" the Internet*. Law and Internet Foundation. Retrieved July 29, 2022, from https://www.netlaw.bg/en/a/article-17-of-the-digital-single-market-directive-and-why-it-broke-the-internet

Geiger, C., & Jütte, B. J. (2021). *Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match* [SSRN Scholarly Paper]. https://papers.ssrn.com/abstract=3776267

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

Hallinan, B., & Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media & Society*, *18*(1), 117–137. https://doi.org/10.1177/1461444814538646

How Europe's new copyright laws will change the creator economy. (2021, August 3). Music Business Worldwide. https://www.musicbusinessworldwide.com/how-europes-new-copyright-laws-will-change-the-creator-economy/

Inside Facebook's fight against European regulation. (2019, January 23). POLITICO. https://www.politico.eu/article/inside-story-facebook-fight-against-european-regulation/

Katzenbach, C. (2018). There Is Always More than Law! From Low IP Regimes to a Governance Perspective in Copyright Research. *Journal of Technology Law & Policy*: 22(2). Available at: https://scholarship.law.ufl.edu/jtlp/vol22/iss2/2

Katzenbach, C., Herweg, S., & Roessel, L. van. (2016). Copies, Clones, and Genre Building: Discourses on Imitation and Innovation in Digital Games. *International Journal of Communication*, *10*(0), 22.

Killeen, M. (2022, May 19). *Commission chides 13 states for failure to transpose Copyright Directive*. Www.Euractiv.Com. https://www.euractiv.com/section/digital/news/commission-chides-13-states-for-failure-to-transpose-copyright-directive/

Klonick, K. (n.d.-a). *The New Governors: The People, Rules, and Processes Governing Online Speech*. Retrieved July 29, 2022, from https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Klonick, K. (n.d.-b). THE NEW GOVERNORS: THE PEOPLE, RULES, AND PROCESSES GOVERNING ONLINE SPEECH. *HARVARD LAW REVIEW*, *131*, 73.

Lomborg, S., & Kapsch, P. H. (2020). Decoding algorithms. *Media, Culture & Society*, *42*(5), 745–761. https://doi.org/10.1177/0163443719855301

Malcolm, D. O. and J. (2018, June 12). *70+ Internet Luminaries Ring the Alarm on EU Copyright Filtering Proposal*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2018/06/internet-luminaries-ring-alarm-eu-copyright-filtering-proposal

Malcolm, J. (2017, October 16). *Digital Rights Groups Demand Deletion of Unlawful Filtering Mandate From Proposed EU Copyright Law*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2017/10/digital-rights-groups-demand-deletion-unlawful-filtering-mandate-proposed-eu

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Quintais, J. P., Frosio, G., van Gompel, S., Hugenholtz, P. B., Husovec, M., Jütte, B. J., & Senftleben, M. (2020). Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive: Recommendations from European Academics. *JIPITEC*, *10*(3). https://www.jipitec.eu/issues/jipitec-10-3-2019/5042

Release, P. (2022, July 5). *EFF Statement on EU Parliament's Adoption of Digital Services Act and Digital Markets Act*. Electronic Frontier Foundation. https://www.eff.org/press/releases/eff-statement-eu-parliaments-formal-approval-digital-services-act-and-digital-markets

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, 01634437221077174. https://doi.org/10.1177/01634437221077174

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. The SAGE handbook of qualitative research in psychology, 2, 17-37.

Van Dijck, J., de Winkel, T., & Schäfer, M. T. (2021). Deplatformization and the governance of the platform ecosystem. *New Media & Society*, 14614448211045662. https://doi.org/10.1177/14614448211045662

Villasenor, A. B. and J. (2021, February 2). The European Copyright Directive: Potential impacts on free expression and privacy. *Brookings*. https://www.brookings.edu/blog/techtank/2021/02/02/the-european-copyright-directive-potential-impacts-on-free-expression-and-privacy/

What is Article 13? The EU's copyright directive explained. (2019, February 14). *BBC News*. https://www.bbc.com/news/technology-47239600

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150. https://doi.org/10.1080/1369118X.2016.1200645

YouTube: We've invested $100 million in Content ID and paid over $3 billion to rightsholders. (2018, November 7). *VentureBeat*. https://venturebeat.com/2018/11/07/youtube-weve-invested-100-million-in-content-id-and-paid-over-3-billion-to-rightsholders/