Impact of content moderation practices and technologies on access and diversity

D.6.3 Final Evaluation and Measuring Report

Schwemer, S.F.; Katzenbach, C.; Dergacheva, D.; Riis, T.; Quintais, J.P.

DOI
10.2139/ssrn.4380345
10.5281/zenodo.7705391

Publication date
2023

Document Version
Final published version

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

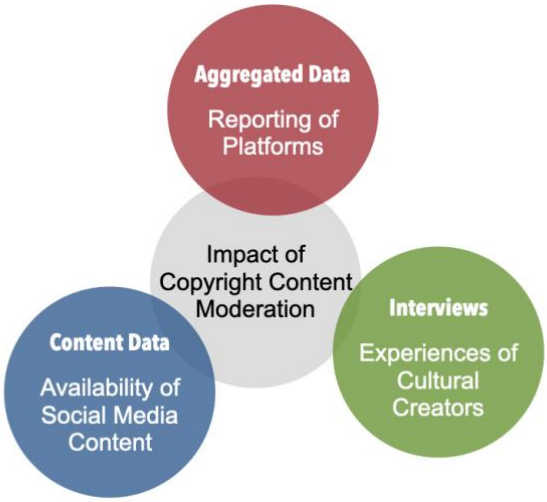
characterization of YouTube over the course of the past 10 years, based on a random sample of channel and video data.

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

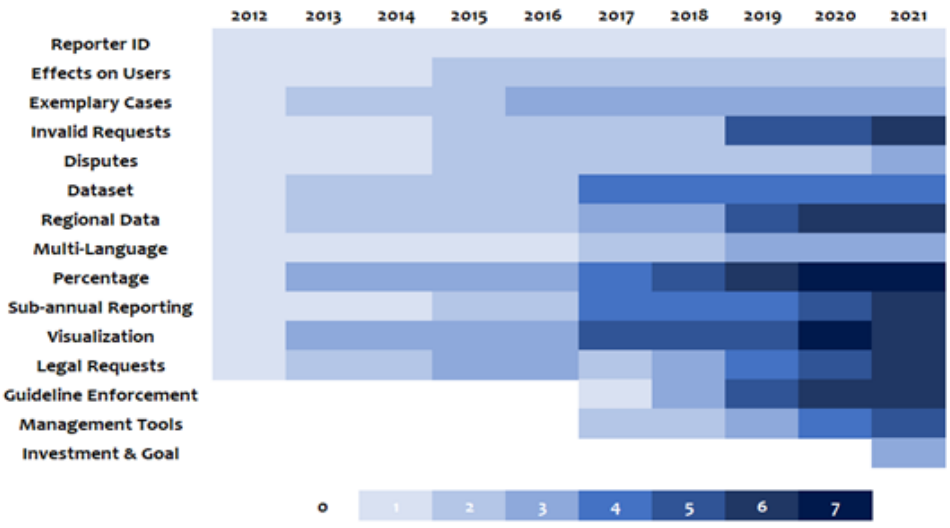
The research design for this sub-study is a longitudinal comparative approach, operationalized with a qualitative content analysis of published transparency reports of major platforms and the information given in their transparency centers. We compare reporting practices and substantial copyright-related 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.

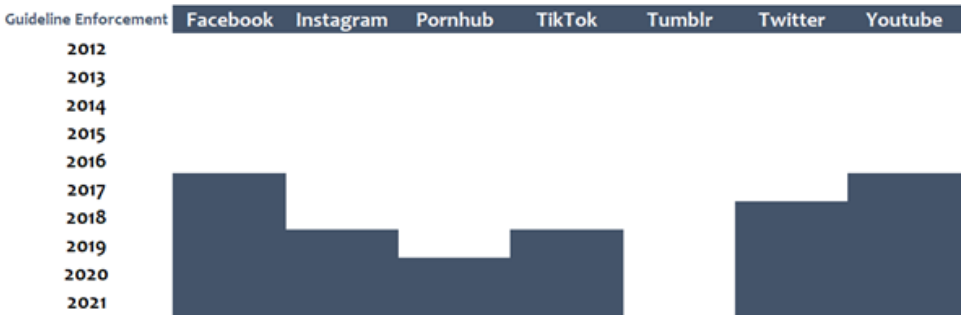


Figure 2: Comparative chronological inclusion of policy-based removal data in platform reporting

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.

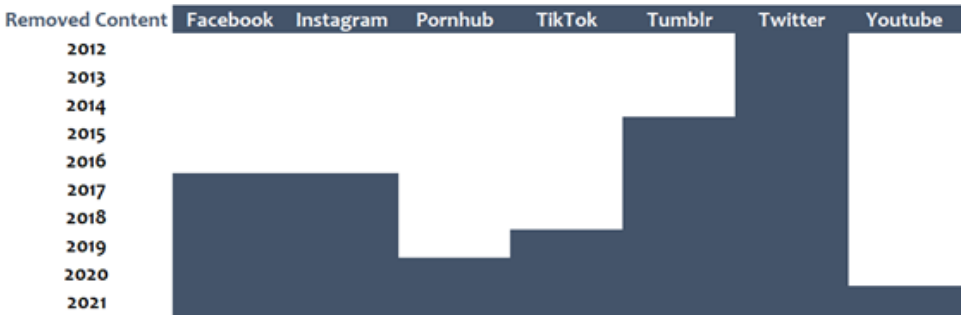


Figure 3: Comparative chronological inclusion of copyright-based removal data in platform reporting

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.

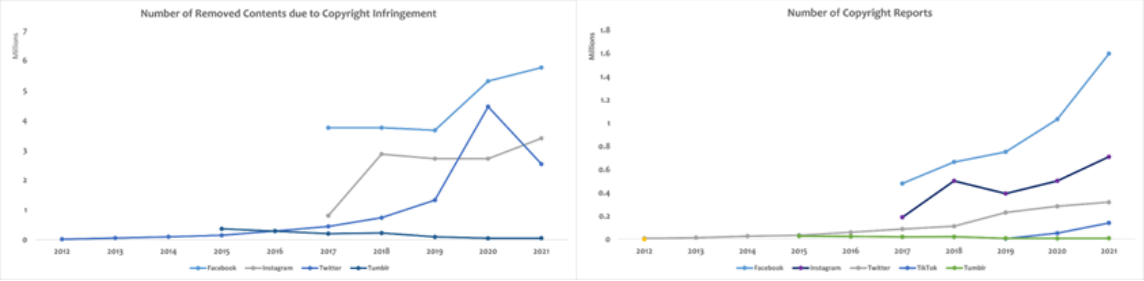


Figure 4: Amount of content removed due to copyright claims per platform & Figure 5: Number of takedown notices received per platform

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.

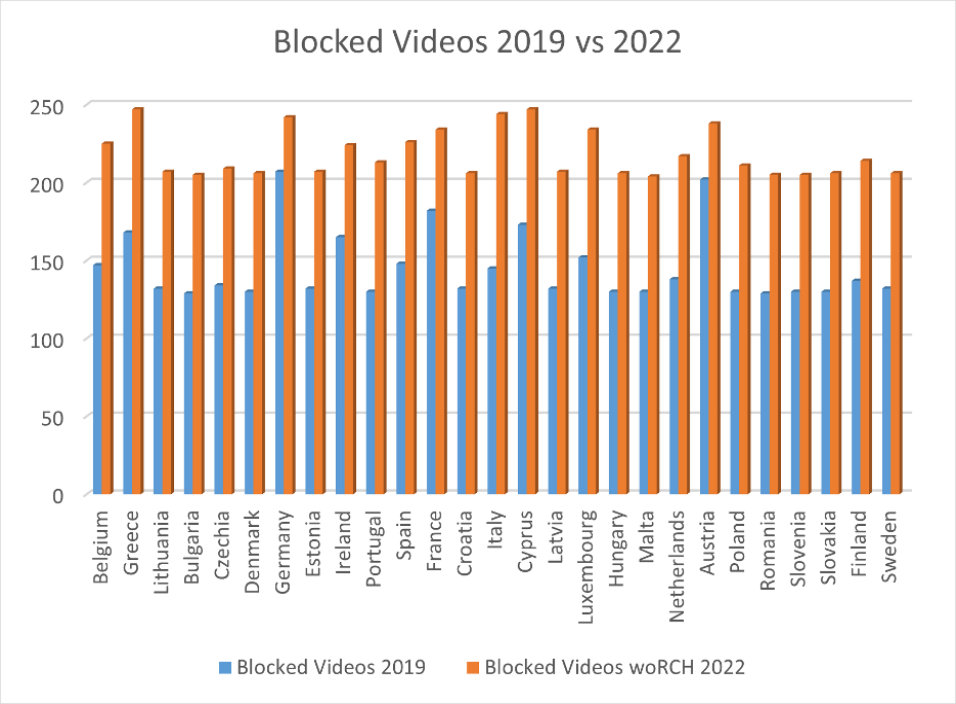
Against this background, this empirical study investigates the changes and influences in access and cultural diversity on social media and streaming platforms, specifically YouTube, in the timeframe 2019 to 2022, exactly focusing on the period between the closure of CDSM Directive negotiations and today, 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)

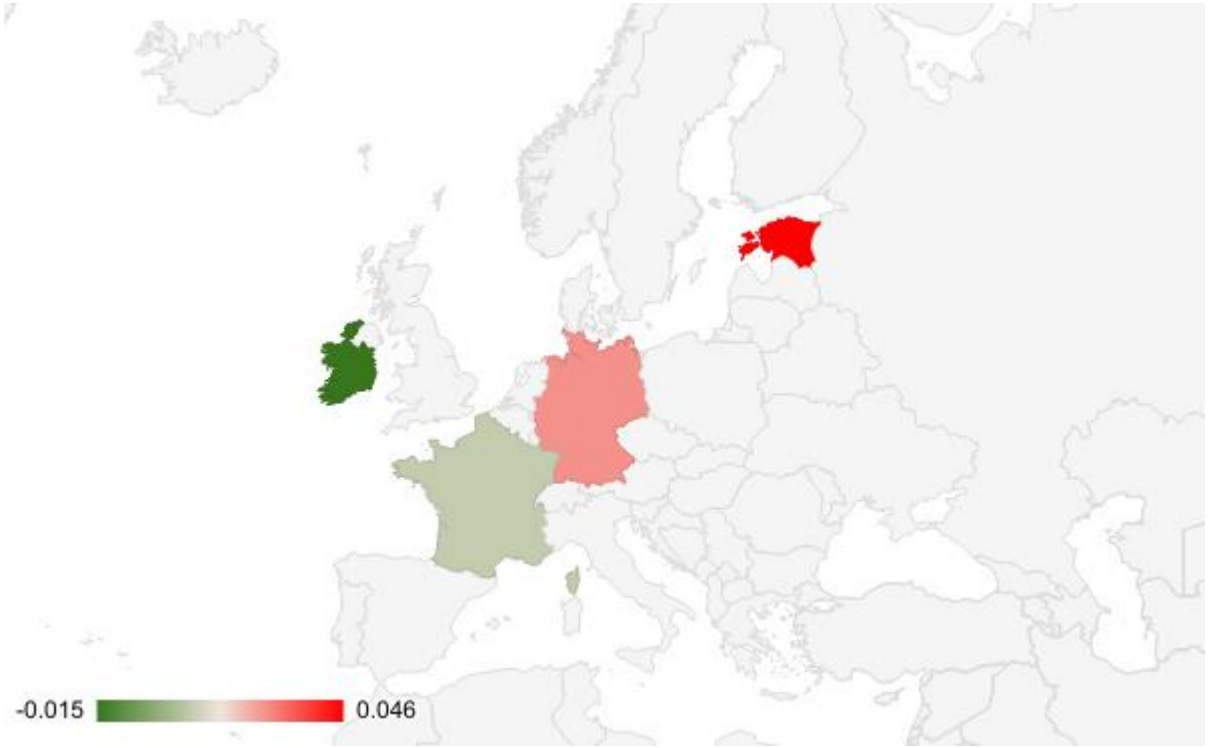


In order to better understand the specifics of blocked or deleted videos vis-à-vis available content, we have run statistical analyses (dominance and regression analysis) to determine key factors. As a result, we found that the main predictor of videos being blocked or deleted was their belonging to 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