BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 38, ISSUE 3

SYMPOSIUM: FROM THE DMCA TO THE DSA

2023

Pages

865–1088

Production: Produced by members of the *Berkeley Technology Law Journal.* All editing and layout done using Microsoft Word.

Printer: Joe Christensen, Inc., Lincoln, Nebraska.
Printed in the U.S.A.
The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Library Materials, ANSI Z39.48—1984.

Berkeley Technology Law Journal
University of California
School of Law
3 Law Building
Berkeley, California 94720-7200
editor@btlj.org
https://www.btlj.org

BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 38 ISSUE 3 2023

ARTICLES

SUBSCRIBER INFORMATION

The *Berkeley Technology Law Journal* (ISSN1086-3818), a continuation of the *High Technology Law Journal* effective Volume 11, is edited by the students of the University of California, Berkeley, School of Law and is published in print three times each year (March, September, December), with a fourth issue published online only (July), by the Regents of the University of California, Berkeley. Periodicals Postage Rate Paid at Berkeley, CA 94704-9998, and at additional mailing offices. POSTMASTER: Send address changes to Journal Publications, Law Library, LL123 South Addition, Berkeley Law, University of California, Berkeley, Berkeley, CA 94720-7210.

Correspondence. Address all correspondence regarding subscriptions, address changes, claims for non-receipt, single copies, advertising, and permission to reprint to Journal Publications, Law Library, LL123 South Addition, Berkeley Law, University of California, Berkeley, Berkeley, CA 94720-7210; (510) 643-6600; JournalPublications@law.berkeley.edu. *Authors*: see section titled Information for Authors.

Subscriptions. Annual subscriptions are $65.00 for individuals and $85.00 for organizations. Single issues are $30.00. Please allow two months for receipt of the first issue. Payment may be made by check, international money order, or credit card (MasterCard/Visa). Domestic claims for non-receipt of issues should be made within 90 days of the month of publication; overseas claims should be made within 180 days. Thereafter, the regular back issue rate ($30.00) will be charged for replacement. Overseas delivery is not guaranteed.

Form. The text and citations in the *Journal* conform generally to THE CHICAGO MANUAL OF STYLE (16th ed. 2010) and to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 21st ed. 2020). Please cite this issue of the *Berkeley Technology Law Journal* as 38 BERKELEY TECH. L.J. ___ (2023).

BTLJ ONLINE

The full text and abstracts of many previously published *Berkeley Technology Law Journal* articles can be found at https://www.btlj.org. Our site also contains a cumulative index; general information about the *Journal*; the *BTLJ Blog*, a collection of short comments and updates about new developments in law and technology written by BTLJ members; and *BTLJ Commentaries*, an exclusively online publication for pieces that are especially time-sensitive and shorter than typical law review articles.

INFORMATION FOR AUTHORS

The Editorial Board of the *Berkeley Technology Law Journal* invites the submission of unsolicited manuscripts. Submissions may include previously unpublished articles, essays, book reviews, case notes, or comments concerning any aspect of the relationship between technology and the law. If any portion of a manuscript has been previously published, the author should so indicate.

Format. Submissions are accepted in electronic format through Scholastica online submission system. Authors should include a curriculum vitae and resume when submitting articles, including his or her full name, credentials, degrees earned, academic or professional affiliations, and citations to all previously published legal articles. The Scholastica submission website can be found at https://btlj.scholasticahq.com/for-authors.

Citations. All citations should conform to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 21st ed. 2020).

Copyrighted Material. If a manuscript contains any copyrighted table, chart, graph, illustration, photograph, or more than eight lines of text, the author must obtain written permission from the copyright holder for use of the material.

SPONSORS
2023–24

The *Berkeley Technology Law Journal* and the Berkeley Center for Law & Technology acknowledge the following generous sponsors of Berkeley Law's Law and Technology Program:

ALLEN & OVERY LLP

ARENTFOX SCHIFF LLP

AXINN, VELTROP & HARKRIDER LLP

BAKER BOTTS L.L.P.

CHARLES RIVER ASSOCIATES

COOLEY LLP

CORNERSTONE RESEARCH

COVINGTON & BURLING LLP

CROWELL & MORING LLP

DESMARAIS LLP

DLA PIPER

DURIE TANGRI LLP

FENWICK & WEST LLP

FISH & RICHARDSON P.C.

GENENTECH, INC.

GEN LAW FIRM

GIBSON, DUNN & CRUTCHER LLP

GILEAD SCIENCES, INC.

GOODWIN PROCTER LLP

GREENBERG TRAURIG, LLP

GTC LAW GROUP PC

HAYNES AND BOONE, LLP

HOGAN LOVELLS

IRELL & MANELLA LLP

JINGTIAN & GONGCHENG

JONES DAY

KEKER, VAN NEST & PETERS LLP

KILPATRICK TOWNSEND & STOCKTON LLP

KING & SPALDING LLP

KING & WOOD MALLESONS

SPONSORS
2023–24

KIRKLAND & ELLIS LLP

KNOBBE MARTENS

LATHAM & WATKINS LLP

MARKS & CLERK

MCDERMOTT WILL & EMERY

MICROSOFT

MORGAN, LEWIS & BOCKIUS LLP

MORRISON & FOERSTER LLP

MUNGER, TOLLES & OLSEN LLP

OCEAN TOMO

ORRICK HERRINGTON & SUTCLIFFE LLP

PAUL HASTINGS, LLP

QUALCOMM TECHNOLOGIES, INC.

QUINN EMANUEL URQUHART & SULLIVAN, LLP

ROBINS KAPLAN LLP

ROPES & GRAY LLP

SIDLEY AUSTIN LLP

TENSEGRITY LAW GROUP LLP

VAN PELT, YI & JAMES LLP

VIA LICENSING CORPORATION

WANHUIDA INTELLECTUAL PROPERTY

WEIL, GOTSHAL & MANGES LLP

WESTERN DIGITAL CORPORATION

WHITE & CASE LLP

WILMER CUTLER PICKERING HALE AND DORR LLP

WILSON SONSINI GOODRICH & ROSATI

WINSTON & STRAWN LLP

WOMBLE BOND DICKINSON

BOARD OF EDITORS
2023–24

Executive Board

BOARD OF EDITORS
2023–24

Editorial Board

Articles Editors

GULNUR BEKMUKHANBETOVA
XUEJIAO CAO
ALEX CHOI
MICHELLE D'SOUZA
GARIMA KEDIA
JOSHUA KUHN
MARLEY MACAREWICH
JOSH MIMURA
BANI SAPRA
SANDEEP STANLEY
BHAVIA SUKHAVASI
NICOLE ZEINSTRA

Notes & Comments Editor
WILLIAM CLARK

Student Publication Editors
MAYA DARROW
JOHN MOORE

Submissions Editors
WILLY ANDERSON
JENNIFER CHENG
VERNON ESPINOZA VALENZUELA
HYEMI PARK

Technical Editors
ASHLEY FAN
EDLENE MIGUEL
BEN PEARCE
ANDREW STONE
CARESSA TSAI

Alumni Relations Editor
EMMA BURKE

External Relations Editor
HUNTER KOLON

Member Relations Editors
JAEYOUNG CHOI
CYRUS KUSHA

Podcast Editors
ERIC AHERN
JULIETTE DRAPER
MEGHAN O'NEILL

Symposium Editors
MARIT BJORNLUND
NICOLE BOUCHER

Production Editors
BEN CLIFNER
SARAH DAVIDSON
KELSEY EDWARDS
ANGELICA KANG
LAUREL MCGRANE

Web & Technology Editors
EMILY WELSCH
LISA YOUNES

LLM Editor
FERNANDA GONZAGA

MEMBERSHIP
2023–24

Members

AYESHA ASAD	AISHWARYA ATHAVALE	CHELSEA BENEDIKTER
DAVID BERNSTEIN	ROMA BHOJWANI	NICOLE BLOOMFIELD
HANNAH BORROWS	OSMANEE CAILLEMER	ALYSON CHIE
ANGELA CHUNG	PETER COE	TIM DABROWSKI
EVELINA DASH	VIVIANA PAOLA DIAZ BAQUERO	HALA EL SOLH
IMAN ESLAMI	SARAH FAROOQ	MAX FRIEND
NADIA GHAFFARI	MRINALINI GOYAT	DAN GRUSHKEVICH
RUI HAN	ANNIKA HANSEN	MARIA HARRAST
ANGELO HERBOSA	MENGRUO HUANG	DYLAN HUGHES
MARIA LUISA ILHARREBORDE	MONICA JEUNG	CORINNE JOHNSTON
YSAMEEN JOULAEE	AARON KAMATH	NAT KAVALER
SRISHTI KHEMKA	TYLER KOTCHMAN	JOSEPH KYBURZ
GAURAV LALSINGHANI	JOSHUA LEE	IRENE LI
KARISSA LIN	WANYI LIN	SARAH LUNT
LILLY MAXFIELD	ZAC MCPHERSON	MAXWELL MELNIK
MARIA MILEKHINA	KIYAN MOHEBBIZADEH	SEAMUS MORIARTY
LEA MOUSTAKAS	GRACIE MURPHY	SAIAISWARYA NAGENDRA

MEMBERSHIP
2023–24

Members (continued)

NIYATI NARANG	NICHOLAS NAVARRO	TUONG-VI NGUYEN
ANTONY NOVAK	YUNFEI QIANG	DEVANGINI RAI
UDAYVIR RANA	SANIDHYA RAO	EMILY REHMET
DELARAI SADEGHITARI	KARINA SANCHEZ	ABBY SANDERS
JULIAN SANGHVI	MASON SEDEN-HANSEN	DHANYA SETLUR KRISHNAN
JACOB SHOFET	AATMAN SHUKLA	GAYATHRI SINDHU
VANSHIKA SINGH	ANKUR SINGHAL	MATT SIOSON
SARAH SISNEY	COLIN STACKPOOLE	HAILEY STEWART
LORENZ STRUB	LESLEY SUN	AMANDA SUZUKI
TRISTAN THREATT	ERIC TING	ITAI TISMANZKY
AMANDA TODD	LINH TRUONG	AARUSHI BAINSLA VERMA
OM SUDHIR VIDYARTHI	SIMON WAGNER	JESSE WANG
SOPHIA WANG	XINRUI WANG	DANIEL WARNER
TIANQI WEI	ETHAN WISEMAN	PAUL WOOD
LIANG-CHU (LUCAS) WU	YING YAO	TWINKLE YE
DUANE YOO	VINCENT ZHAI	TERRY ZHAO
	FAYE ZOU	

BERKELEY CENTER FOR LAW & TECHNOLOGY
2023–24

WAYNE STACY
Executive Director

Staff

MARK COHEN
Senior Fellow & Director, BCLT Asia IP Project

ALLISON SCHMITT
Fellow & Director, BCLT Life Sciences Project

JANN DUDLEY
Associate Director

RICHARD FISK
Assistant Director, Events & Communications

JUSTIN TRI DO
Media Coordinator

ABRIL DELGADO
Events Specialist

Fellow

YUAN HAO
Senior Fellow

KATHRYN HASHIMOTO
Copyright Law Fellow

RAMYA CHANDRASEKHAR
Biometric Regulatory Fellow

ROBERT BARR
BCLT Executive Director Emeritus

BERKELEY CENTER FOR LAW & TECHNOLOGY
2023–24

Faulty Directors

KENNETH A. BAMBERGER
The Rosalinde and Arthur Gilbert Foundation Professor of Law

CATHERINE CRUMP
Robert Glushko Clinical Professor of Practice in Technology Law & Director, Samuelson Law, Technology and Public Policy Clinic

CATHERINE FISK
Barbara Nachtrieb Armstrong Professor of Law

CHRIS JAY HOOFNAGLE
Professor of Law in Residence

SONIA KATYAL
Roger J. Traynor Distinguished Professor of Law & Associate Dean, Faculty Development and Research

ORIN S. KERR
William G. Simon Professor of Law

PETER S. MENELL
Koret Professor of Law

ROBERT P. MERGES
Wilson Sonsini Goodrich & Rosati Professor of Law

DEIRDRE K. MULLIGAN
Professor in the School of Information

TEJAS N. NARECHANIA
Robert and Nanci Corson Assistant Professor of Law

BRANDIE NONNECKE
Associate Research Professor at the Goldman School of Public Policy

OSAGIE K. OBASOGIE
Haas Distinguished Chair, Professor of Law & Professor of Bioethics

ANDREA ROTH
Professor of Law

PAMELA SAMUELSON
Richard M. Sherman Distinguished Professor of Law

PAUL SCHWARTZ
Jefferson E. Peyser Professor of Law

ERIK STALLMAN
Assistant Clinical Professor & Associate Director, Samuelson Law, Technology & Public Policy Clinic

JENNIFER M. URBAN
Clinical Professor of Law & Director, Samuelson Law, Technology & Public Policy Clinic

MOLLY SHAFFER VAN HOUWELING
Harold C. Hohbach Distinguished Professor of Patent Law and Intellectual Property

REBECCA WEXLER
Assistant Professor of Law

FOREWORD

Jennifer M. Urban[†]

I. INTRODUCTION

After more than two decades of the "notice-and-takedown" approach to online copyright infringement and content moderation, the European Union (EU) has moved away from this familiar regime and toward a broader regulatory approach with the Directive on Copyright and Related Rights in the Digital Single Market (CDSMD) and the Digital Services Act (DSA). The Berkeley Center for Law and Technology and the *Berkeley Technology Law Journal*'s 27th Annual Symposium considers this potentially profound shift in copyright enforcement and content moderation policy. On April 6th and 7th, 2023, scholars, policymakers, and industry participants from both Europe and the United States joined in discussion to consider potential benefits and risks of the EU's new approach and whether a new EU/US consensus—or, perhaps, a "Brussels Effect" on US platform liability debates—is likely.

On the first day of the symposium, European experts presented valuable tutorials explaining the architecture of the DSA and the complexities of its core features. They provided US attendees with a map of the DSA's role in the European context, a blueprint of its structure, a breakdown of its interactions with the CDSMD, a comparison to previous approaches, and an analysis of its potential effects on free speech.[1]

On the second day, US experts joined European experts on a series of panels considering how the DSA affects online service providers' responsibilities, what the intended and unintended consequences of the DSA

DOI: https://doi.org/10.15779/Z38697001X

† Clinical Professor of Law at University of California, Berkeley, School of Law; Director of Policy Initiatives, Samuelson Law, Technology & Public Policy Clinic; Co-Director, Berkeley Center for Law and Technology (BCLT). Opinions are my own and should not be attributed to my institution, the California Privacy Protection Agency, or the California Privacy Protection Agency Board. This conference was a transatlantic group effort. Thank you to Professors Martin Senftleben and João Pedro Quintais of the Institute for Information Law (IViR) at the University of Amsterdam and Professors Pam Samuelson and Erik Stallman at UC Berkeley, to the expert BCLT staff, and to the team at the *Berkeley Technology Law Journal.*

1. *27th Annual BTLJ-BCLT Symposium: From the DMCA to the DSA—A Transatlantic Dialogue on Online Platform Liability and Copyright Law Agenda*, BERKELEY LAW (Apr. 6–7, 2023), https://www.law.berkeley.edu/research/bclt/bcltevents/from-the-dmca-to-the-dsa-a-transatlantic-dialogue-on-online-platform-liability-and-copyright-law/agenda/.

may be on fundamental rights, and whether the DSA will influence firm behaviors beyond the EU via a "Brussels Effect."[2]

Attendees also heard from a panel of industry experts on industry perspectives, and benefited from keynote addresses by officials from both sides of the Atlantic. Irene Roche-Laguna, a European Commission official who was key to developing the DSA, discussed the DSA's origins, and goals.[3] She pointed out that the DSA attempts to address a host of critiques of notice-and-takedown, many originating from the US. She asserted: "This is your baby."[4] Shira Perlmutter, the Register of Copyrights for the US, discussed how emerging technologies are currently affecting copyright policy. Among other examples, she walked the audience through the Copyright Office's recent analysis of copyright issues related to generative artificial intelligence technologies.[5]

The five papers in this symposium edition of the *Berkeley Technology Law Journal* both helped constitute this cross-Atlantic discussion and grew from it. They offer viewpoints from both sides of the Atlantic, highlighting potential benefits and risks in the EU's new approach. As Europe moves away from liability rules premised on notice-and-takedown processes and toward horizontal "due diligence" and "accountability" requirements, these papers offer background, optimism, pessimism, and critique. Brief introductions to their rich analyses follow.

II. HUSOVEC: THE DSA AS A BLUEPRINT

In "Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules," Martin Husovec, of The London School of Economics and Political Science, analyzes the DSA as the "first comprehensive attempt to create a second generation of rules for digital services that rely on user-generated content."[6] Though recognizing that some of the regulation's features may be too Europe-specific to travel, Husovec argues that "the principles behind the DSA could be useful in other jurisdictions—perhaps even in the United States" by serving as "the basis for

2. *Id.*

3. *Id.*

4. *See* author's note (on file with author).

5. *27th Annual BTLJ-BCLT Symposium: How Are Emerging Technologies Affecting Copyright Policy?*, BERKELEY LAW, https://bk.webcredenza.com/watch?id=85216 (last accessed Jan. 10, 2024).

6. Martin Husovec, *Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules*, 38 BERKELEY TECH. L.J. 882 (2023).

a dialogue between liberal democracies about how to best regulate user-generated content services."[7]

Husovec first traces a history of the DSA's foundations, highlighting the influence of section 512 of the US Digital Millennium Copyright Act (DMCA) on the European E-Commerce Directive and the EU's ensuing "conditional immunity" approach to service provider liability for user-generated content.[8] Husovec praises this approach as a structurally sound method of encouraging the growth of decentralized communication networks, arguing that, via liability exemptions, "everyone commits to constraining themselves in order to facilitate the emergence of an environment from which everyone can benefit."[9]

But, Husovec argues, today this structure "seems insufficient when the clear legislative goal of the liability exceptions was to lay down incomplete and unrestrictive rules that would allow the medium to flourish." We are now in a world of "many societal challenges that require solutions," a task that, in Husovec's view, cannot be completed via liability exemptions alone.[10]

This brings us to the DSA, which Husovec characterizes as resting on "two pillars": due process requirements for content moderation and risk management obligations for service providers.[11]

As to the first, Husovec stresses that the DSA regulates the process by which service providers make content moderation decisions, not the underlying rules for what content is acceptable. Those rules (for lawful content) remain in service providers' hands.[12] Husovec sees the DSA's process requirements as a way of addressing underinvestment by service providers in content moderation decision-making.[13]

The DSA then imposes another layer of regulation—risk mitigation requirements—on online platforms, very large online platforms (VLOPs) and very large online search engines (VLOSEs).[14] The services that fall into these categories must avoid manipulative product design generally, and must consider the effects of their product design on children specifically. The largest services are treated as "public squares" and must make additional risk mitigation efforts; these include engaging in dialogue with regulators about risks to both individual freedoms and democratic institutions.[15] Husovec sees

7. *Id.* at 887.
8. *Id.* at 883–87.
9. *Id.* at 893.
10. *Id.* at 897.
11. *Id.* at 899.
12. *Id.* at 901.
13. *Id.*
14. *Id.* at 900–1.
15. *Id.*

this approach as a recognition of the importance of product design in outcomes and of longstanding asymmetries of information and resources between firms and regulators.[16] At the same time, he recognizes that regulatory attempts to address systemic risk in this way invite suppressing individual expression, especially for "lawful but harmful" content.[17] Husovec considers the key question to be who—regulators or firms "sets the boundaries for the content of communications."[18] In his view, the DSA leaves room for firms to make decisions about legal content, while incentivizing investment in good decisionmaking.

Husovec advocates for other jurisdictions to be guided by five "principles" that he has extracted from the DSA: accountability not liability; horizontality; shared burden; empowerment; and ecosystem solutions.[19]

As to *accountability not liability*, Husovec argues that platforms "as facilitators of user-generated content cannot be expected to bear the liability burden of ordinary publishers."[20] But, he argues, they should be "more accountable" for protecting "individual grievances " by exercising due diligence.[21] He finds the DSA's model superior to liability limitations alone because "[i]n the liability framework, the lack of diligence puts providers at risk of being an accessory to the entire wrongs of others. On the other hand, the accountability framework blames them only for not giving some specific assistance."[22] "Accountability not liability" ties to the principle of *shared burden*, which Husovec summarizes as "everyone is expected to play their part" to limit speech risks. He argues that this principle can be fulfilled by using both liability exemptions and accountability mechanisms to allocate responsibilities.[23] In turn, the principle of shared burden ties to the principle of *user empowerment*, which Husovec uses to argue for users to share risks—but only so far as they are able to counter those risks.[24] The DSA's due diligence obligations, in his view, encourage firms to provide users with the necessary tools.[25]

Husovec is even more complimentary toward the *horizontality* of the DSA, calling it a "digital civil charter that shines through the entire legal system and radiates minimum rights of individuals," regardless of the specific EU

16. *Id.* at 904–5.
17. *Id.* at 906–8.
18. *Id.* at 906.
19. *Id.* at 909.
20. *Id.*
21. *Id.* at 910.
22. *Id.* at 911.
23. *Id.* 913–14.
24. *Id.* at 914–15.
25. *Id.*

jurisdiction.[26] The DSA also sweeps broadly across legal sectors; Husovec argues that this tamps down regulatory arbitrage and forces regulators to consider tradeoffs across the entire landscape of online speech.[27] And relatedly, Husovec compliments the DSA for, in his view, employing the final principle of *ecosystem solutions*. The DSA both sweeps across jurisdictions and sweeps in multiple actors. Husovec argues that previous regimes exhibited a "preoccupation with [online service] providers," giving "little consideration" to others, such as "trusted NGOs . . . fact-checkers, journalists, or researchers."[28] The DSA's allowances for "trusted flaggers," information-sharing, and research will, in Husovec's view, be highly beneficial if they are implemented fully.[29]

Accordingly, in Husovec's analysis, the DSA, if properly implemented, promises to support user-generated content, while "inject[ing] trust" into the system.[30]

III. TUSHNET: RIGHTSIZING REGULATION THROUGH TEST SUITES

In "Three Sizes Fit Some: Why Content Regulation Needs Test Suites," Rebecca Tushnet of Harvard Law School takes a more skeptical view, identifying potential weaknesses in the DSA's novel structure. In Tushnet's assessment, the DSA fails in one of its key features: establishing size-based tiers of online service providers and then differentially imposing obligations by tier. This feature of both the DSA and CDSMD is intended to tailor obligations to relative risk and resources. Yet Tushnet considers them "totalizing," and likely to "damage a thriving online ecosystem," because they fail to capture the true variation within that ecosystem.

Tushnet's skepticism begins at the first gate: establishing the "size" of service providers in order to sort them into regulatory tiers.[31] The DSA requires providers to count monthly active users who have "engaged" with the service for this purpose.[32] Yet, Tushnet points out, there is inherent ambiguity in the required metric. Further, the metric raises potential privacy issues: not all platforms "extensively track users," as not all seek to monetize or prolong

26. *Id.* at 912.
27. *Id.* at 912–13.
28. *Id.* at 917.
29. *Id.* at 918–20.
30. *Id.* at 920.
31. Rebecca Tushnet, *Three Sizes Fit Some: Why Content Regulation Needs Test Suites*, 38 BERKELEY TECH. L.J. 921 (2023).
32. DSA Art. 3(p).

visits.[33] Tushnet points to Wikipedia, the Organization for Transformative Works' Archive of Our Own, and DuckDuckGo as examples of service providers for which the risk of bad behavior seems low, but the potential costs of tracking seem high.[34]

Tushnet also considers the DSA's extensive due process requirements too generalized, and at risk of creating unintended consequences. She points out, for example, that the requirements—which include individualized explanations of platform decisions and a redress process—apply equally to brief comments and longform content, and to acts ranging from demonetization, to removing an "a politician's entire account," and on "to downranking a single post by a private figure."[35] Coupled with protections against bad-faith actors that are, in Tushnet's view, inadequate, particularly in light of demographic differences in who is likely to be willing to use redress systems, this design may lead service providers to reduce their efforts to moderate "lawful but awful" content. Further, the cost of the DSA's requirements could create anticompetitive barriers to smaller and newer market actors.[36]

In Tushnet's analysis, these challenges arise from a regulatory myopia that prompts regulators to focus on "the giant names they know" when crafting regulations. To ameliorate this issue, she argues for regulators to use "test suites" to explore varying types of online service providers, the risks (or relative lack of risk) they present, and the different challenges they face. In her view, "true proportionality" is achievable only with closer attention to the actual diversity of online service providers.[37]

IV. SENFTLEBEN, QUINTAIS, AND & MEIRING: HUMAN RIGHTS IMPLICATIONS OF PLATFORM REGULATION

Martin Senftleben, João Pedro Quintais, and Arlette Meiring, from the University of Amsterdam, complement Tushnet's critique with a detailed analysis of the human rights implications of the CDSMD and DSA, focusing on monetization. In "How the European Union Outsources the Task of Human Rights Protection to Platforms and Users: The Case of User-Generated Content Monetization," the authors take as case studies the content monetization remedies several major providers allow large rightholders to exercise against user-generated content (UGC). These examples illustrate what

33. Tushnet, *supra* note 31, at 924.
34. *Id.* at 923–25.
35. *Id.* at 926–27.
36. *Id.* at 929.
37. *Id.* 930–32.

the authors view as human rights issues created by design deficits in the CDSMD and the DSA.

The authors first offer a detailed analysis of the intricate interaction between the CDSMD and the DSA, highlighting human rights implications. They identify two main human rights effects, which they term *outsourcing* and *concealing*.[38]

Outsourcing stems from the laws' failure to include "concrete solutions for human rights tensions in the law itself."[39] Instead, the law "outsources" safeguards for fundamental rights to private parties—online platforms, in cooperation with the creative industry, and activist users.[40] For example, the DSA requires UCG platforms to "act in a diligent, objective and proportionate manner . . . with due regard to . . . the fundamental rights of [users]"—thus outsourcing the protection of fundamental rights to platforms.[41] The DSA also requires platforms to inform users about how they approach content moderation, including via algorithmic decision-making.[42] And platforms must provide internal systems for handling complaints about content moderation decisions, and information about those systems.[43] The authors take these and similar requirements as evidence of outsourcing not just to platforms, but also to users, who are expected to understand the platforms' policies and use the platforms' systems "to play an active role in the preservation of their freedom of expression and information."[44]

The authors are skeptical about whether legislators can "legitimately 'outsource' the obligation to safeguard fundamental rights" in this way.[45] In part, this is because leaving so much responsibility to private parties may conceal human rights issues from view. For example, both the CDSMD and the DSA rely on user complaints to identify problematic content blocking or removal. But, the authors point out, a "low number of user complaints . . . may be misinterpreted as an indication that content filtering hardly ever encroaches upon freedom of expression and information."[46] Instead, cumbersome complaint procedures and other barriers make it "unrealistic to assume that"

38. Martin Senftleben, João Pedro Quintais, & Arlette Meiring, *How the European Union Outsources the Task of Human Rights Protection to Platforms and Users: The Case of User-Generated Content Monetization*, 38 BERKELEY TECH. L.J. 933, 943–73 (2023).
39. *Id.* at 943.
40. *Id.* at 943–55.
41. *Id.* at 941.
42. *Id.* at 939–40.
43. *Id.* at 941.
44. *Id.*
45. *Id.* at 942.
46. *Id.* at 957.

user complaints "reveal[] the full spectrum and impact of free expression restrictions" at issue.[47] Problems may exist, but may be hidden by practical limitations on users' ability to affirmatively assert their rights.

Overall, the authors see in the CDSMD and the DSA "a worrying tendency of reliance on industry cooperation and user activism to safeguard human rights."[48] Though the Court of Justice for the European Union has guarded free expression by "stating unequivocally" that filtering systems must "be capable of distinguishing lawful from unlawful content,"[49] the Court "did not seize the opportunity to unmask human rights risks . . . inherent in the [CDSMD's] heavy reliance on industry cooperation," nor did it address the "human rights risks that could arise from the ineffectiveness of complaint and redress mechanisms for users."[50] The authors do find promise in the DSA's audit provisions, which could return some responsibility for protecting human rights to the European Commission. Accordingly, the audit provisions "must not be underestimated" as "a promising counterbalance to outsourcing/concealment risks."[51] Still, it remains unclear whether the audit requirements will fulfil this promise. Ultimately, both the intended protections for lawful uses in Article 17(7) of the CDSMD and the audit requirements contained in the DSA are too "underdeveloped" to fully counter the authors' concerns.[52]

The authors then apply their analysis to one method of content moderation: monetization programs. As the authors point out, content removal and blocking/filtering garner much more attention from commentators, but 'monetization'—the opportunity to capture "advertising revenue that accrues from the continued online availability of UGC"—is a very popular choice for rightholders who have access to it.[53] Indeed, the authors report, rightholders eligible for YouTube's ContentID chose monetization as the remedy for over 90% of claims made over a six-month period.[54] Yet the CDSMD "largely ignores the topic" of monetization.[55] The DSA does include "demonetization" in its framework, including it specifically in the set of negative actions users (or others) can appeal through platforms' complaint

47. *Id.* at 959.
48. *Id.* at 973.
49. *Id.* at 964 (citing CJEU, 26 April 2022, case C-401/19, Poland v Parliament and Council).
50. *Id.* (citing CJEU, 26 April 2022, case C-401/19, Poland v Parliament and Council).
51. *Id.* at 972.
52. *Id.* at 973.
53. *Id.*
54. *Id.* at 986 (internal citations omitted).
55. *Id.* at 974 (internal citations omitted).

systems.[56] Still, the authors view the DSA as addressing monetization "at a superficial level, mostly by outsourcing its regulation to private parties."[57] Due to this outsourcing, the authors point out, the "workings of [monetization systems] are mostly concealed behind complex terms and conditions and opaque algorithmic systems" employed by platforms in cooperation with rightholders.[58]

After undertaking a thorough review of (the admittedly limited) publicly available information about several large companies' [59] approaches to monetization, the authors conclude that outsourcing monetization remedies to private actors leads to, and conceals, at least three important human rights issues. First, major rightholders can appropriate and exploit transformative UGC, invading and "usurp[ing] this freedom of expression space."[60] Second, relatedly, misappropriating user creativity in this manner encroaches on the user's fundamental right to property by treading on the user's intellectual property rights.[61] And third, favoring large-scale rightholders over user-creators "gives rise to the question of whether it violates the principle of equal treatment" in the Charter of Fundamental Rights of the European Union.[62]

Accordingly, though the DSA contains some promising features, Senftleben, Quintais, and Meiring consider it insufficient to the task of protecting human rights. They call for collective licensing with "non-waivable remuneration" for UGC creators, and for a general redesign of monetization systems to benefit user-creators as well as large rightholders.[63]

V. GRIMMELMANN & ZHANG: AN ECONOMIC MODEL OF INTERMEDIARY LIABILITY

In "An Economic Model of Online Intermediary Liability," James Grimmelmann and Pengfei Zhang take a different tack. Rather than focusing on the DSA from the outset, these authors take a step back in order to "clarify the terms of the debate" over how best to structure intermediary liability by developing a generalized economic model.[64] They argue that standardizing

56. *Id.* at 982–83 (internal citations omitted).
57. *Id.* at 974.
58. *Id.*
59. The authors review YouTube, Meta, TikTok, and third-party offerings from Audible Magic and Pex. *Id.* at 984–98.
60. *Id.* at 1000.
61. *Id.* at 1004.
62. *Id.* at 1006.
63. *Id.* at 1010.
64. James Grimmelmann & Pengfei Zhang, *An Economic Model of Online Intermediary Liability*, 38 BERKELEY TECH. L.J. 1011, 1013 (2023).

arguments into a formal economic model promotes communication, intuition, visualization, rigor, proof, and empiricism.[65] By standardizing the terms of the debate and making its assumptions explicit, the authors believe, they can order and improve the intermediary liability debate. They then use their model to compare the relative benefits and drawbacks of different approaches to platform regulation, including section 230 of the US Communications Decency Act, and section 512 of the DMCA, and the DSA.[66]

Reviewing the available literature on platform liability, the authors find that there is very little formal economic analysis; varied views on the best approach (ranging from no liability, to conditional liability, to strict liability (or even criminal liability) for certain harms); and some descriptive empirics on platform behavior.[67] But there is an "immense" literature exploring economic theories of liability.[68]

Drawing on this literature, Grimmelmann and Zhang seek to determine which is economically optimal: "online intermediary liability" or "online intermediary immunity."[69] They take as initial assumptions two observations: platforms have *imperfect information* about the harmfulness of content they host; and content can have *positive externalities* that go beyond the benefits the platform can internalize. Taken together, these features of the online content ecosystem, they argue, could plausibly cause platforms to overmoderate.[70]

Relying on these assumptions, the authors illustrate the uncertainty platforms face with a simple probability model. Any given piece of content carries a probability of being harmless or harmful. Platforms do not know whether a given piece of content actually is harmful, but they can know something about the probability that it is.[71] The authors then include the probabilities of various consequences flowing from hosted content: that the platform receives some benefit; that society receives some benefit; and that harmful content causes someone harm. To sharpen the model, they assume that there exists some set of "good" content that benefits the platform, benefits society, and is always harmless. Likewise, they assume that there exists some set of "bad" content that is bad for society and always harmful. This allows them to visualize a "moderation threshold" at which a rational moderator will shift from removing content to leaving it up, along with

65. *Id.* at 1013–14.
66. *Id.* at 1060–64.
67. *Id.* at 1014–18.
68. *Id.* at 1014.
69. *Id.*
70. *Id.* at 1019.
71. Later, the authors add options for costless and costly investigations of content by platforms. *Id.* at 1032–39.

changes in platform profit, social benefit, and social harm as the threshold shifts.[72]

Armed with this model, the authors test various models of liability. Giving platforms *blanket immunity*, perhaps surprisingly, can result in both undermoderation (where platforms leave up too much harmful content) and overmoderation (where platforms remove too much socially beneficial content). This is because platforms don't fully internalize the benefits of hosted content (and so might remove content that benefits society), and also don't internalize harms suffered by third parties (and so might leave up harmful content).[73] On the other hand, imposing *strict liability* on platforms always causes overmoderation, a conclusion the authors can nicely demonstrate with their model.[74] The authors complicate the picture by testing the effects of platforms engaging in *costless investigations* (which are always to the good) or *costly investigations* (which will cause some overremoval).[75]

Clarifying assumptions and formalizing policy components in this way allows the authors to compare different policy approaches to content moderation. Regulators wishing to address undermoderation have a few traditional tools to choose from. They could impose liability based on *actual knowledge* by the platform of harmful content; the authors consider this option to be an improvement over strict liability if "actual knowledge" is not distorted into a lower threshold (at which point platforms begin to overmoderate).[76] Regulators could impose *liability on notice* from victims, which leaves some uncompensated harm (due to victims' investigation costs), but at first appears to enhance social welfare.[77] However, if victims can shirk proper investigation and send notices for content that is not harmful, then liability on notice "might collapse into strict liability" because the bad notices "are of no use to the platform in distinguishing harmful from harmless content," but still trigger strict liability for the platform.[78] This is an observed problem with section 512 notice-and-takedown that likely causes overmoderation.

Regulators could also impose standards-based models of liability. They could turn to *negligence* and impose a standard of care that requires some amount of investment by platforms in preventing harm.[79] This can run into

72. *Id.* at 1019–25.
73. *Id.* at 1025–29.
74. *Id.* at 1029–32.
75. Note: here I have radically simplified seven pages of close and careful reasoning. *Id.* at 1032–39.
76. *Id.* at 1045–46.
77. *Id.* at 1046–47.
78. *Id.* at 1047.
79. *Id.* at 1049–53.

difficulty because it's difficult to choose the optimal standard.[80] Or regulators could create *conditional immunity* by setting a threshold of harm and providing immunity to platforms that don't cross it.[81] These methods sound very similar, but are distinct because negligent platforms are liable for specific pieces of content for which they didn't exercise sufficient care, while platforms that lose conditional immunity lose it for all content by blowing their harm "budget."[82]

After briefly considering approaches to overmoderation (subsidies and must-carry requirements),[83] the authors use their findings to evaluate existing and proposed approaches.[84] In their model, Section 230 functions as blanket immunity for the content it covers, and reform proposals vary.[85] The Citron-Wittes proposal, which turns on overall moderation efforts, is a conditional immunity approach. Efforts to impose common-law distributor liability function as liability on notice. And the Platform Accountability and Consumer Transparency Act would impose liability on notice, but where relevant "notice" requires a court order. The model allows some important trade-offs inherent in these approaches—for e.g., the cost of investigations, or the loss or accrual of social benefits—to be made explicit and compared.

The authors' model is especially helpful in bringing analytical order to the hodge-podge that is section 512 of the DMCA. According to their analysis, section 512 combines multiple approaches, starting with blanket immunity, but then adding five exceptions, each a different "flavor" of liability.[86] First, the platform loses immunity with actual knowledge.[87] Second, it loses immunity if it fails to remove infringing material when it has a sufficient level of awareness (negligence).[88] Third, it loses immunity if it has the ability to control and is strongly under-investing in investigations[89] Fourth, the platform loses immunity if it receives a notice of claimed infringement and fails to remove it (liability on notice).[90] Finally, it loses immunity if it fails to ban "repeat infringers" according to some threshold. This exception, the authors point out, has functioned as a conditional immunity standard, with some platforms staying on the "safe" side of the harm threshold while others (most famously,

80. *Id.* at 1052.
81. *Id.* at 1053–55.
82. *Id.* at 1054.
83. *Id.* at 1055–61.
84. *Id.* at 1061–65.
85. *Id.* at 1061–62.
86. *Id.* at 1062–64.
87. *Id.* at 1062.
88. *Id.* at 1062–63.
89. *Id.* at 1062.
90. *Id.*

Cox Communications) ending up on the wrong side of the line and thus, without immunity for their users' infringement.[91]

Informed by their model, the authors find several things to like in the DSA's approach.[92] First, it more sharply distinguishes between "mere conduits" and "hosting providers" than the DMCA does. Under the DSA, and like the DMCA, conduits have no content moderation requirements. However, the DSA does not, in the authors' view, condition platforms' immunity on terminating repeat infringers. Nor does it have vicarious-liability-like provisions. This approach more cleanly focuses content moderation responsibilities on hosting providers, which are subject to notice-and-takedown requirements.[93] The authors also compliment the DSA's "trusted flagger" system, which sets an investigation standard for trusted flaggers to meet. They characterize this a "clever response to the signaling problem" evident in the DMCA (which lacks sufficient disincentives to sending under-investigated notices).[94] Finally, the DSA, like section 230 of the CDA, neither requires platforms to actively monitor hosted content nor punishes them for investigating and moderating. Together, this "prevent[s] the *Stratton Oakmont* trap," in which platforms could face strict liability for all harmful content if they remove any at all.[95]

VI. CHANDER: GLOBAL EFFECTS OF THE DSA

In his essay, "When the Digital Services Act Goes Global," Georgetown University's Anupam Chander argues that the DSA is likely to influence jurisdictions beyond Europe via a "Brussels Effect" and considers the ensuing risk to civil society and freedom of expression.[96]

Chander considers it likely that the DSA will "likely carry a Brussels Effect, both de facto through changes in the practices of multinational corporations, and de jure through changes in foreign law."[97] He does not delve deeply into the details, but follows Dawn Nunziato in pointing out the DSA's extraordinary financial enforcement mechanisms—fines of up to six percent of a targeted platform's worldwide turnover—as a source of pressure on firms

91. *Id.* at 1055 (internal citations omitted).
92. *Id.* at 1064–65.
93. *Id.*
94. *Id.* at 1064.
95. *Id.*
96. Anupam Chander, *When the Digital Services Act Goes Global*, 38 BERKELEY TECH. L.J. 1067, 1067–68 (2023).
97. *Id.* at 1071.

to err on the side of European norms when developing content policies.[98] He also points out firms might find it convenient to standardize content policies in response to the DSA's transparency requirements,[99] and that European regulators have stated that they hope to effect "global standards" through the DSA and DMA.[100]

More important to Chander, however, is his view that governments "might find much to envy in the Digital Services Act" leading to a so-called "de jure" Brussels Effect as governments adapt their laws to reflect the DSA.[101] Whereas European leaders hope to encourage "democracy, fundamental values, and the rule of law,"[102] Chander worries that some of the DSA's mechanisms may have very different effects in the hands of "governments with authoritarian tendencies."[103]

To analyze these possible effects, Chander sets a "Putin Test" for various aspects of the DSA.[104] In essence, he asks, "What would Putin do?" with each mechanism. First up are the DSA's Digital Services Coordinators—national regulators who are to be established in each European Member State. Chander points out that the Digital Services Coordinator is entrusted with substantial powers that touch on speech, including choosing "trusted flaggers," investigating user complaints, requesting information from VLOPs and VLOSEs, choosing "vetted researchers," ordering content removal, and issuing those extraordinary six-percent fines.[105] Though the DSA imposes constraints on each of these activities to ensure the protection of fundamental rights, Chander points out that an interested Digital Services Coordinator could act in accordance with narrow political, personal, or ideological preferences to harass platforms or otherwise use its power to achieve anti-democratic goals.[106] Next, Chander worries about the DSA's establishment of emergency powers and its requirement that all EU-serving intermediaries designate local EU representatives. Emergency powers create the potential for abusive government coercion, as do requirements to place a representative within physical reach.[107]

98. *Id.* (internal citations omitted).
99. *Id.* at 1071–72.
100. *Id.* at 1074.
101. *Id.* at 1073.
102. *Id.* at 1075.
103. *Id.* at 1077.
104. *Id.* at 1075–80.
105. *Id.* at 1077–79.
106. *Id.* at 1079.
107. *Id.* at 1079–80.

Ultimately, Chander calls for a recognition that both corporate actors and governments can threaten speech, and for vigilant attention to the ways in which the DSA could be misused in non-EU jurisdictions.[108]

VII. CONCLUSION

The DSA is an exceptionally complicated law, with far-reaching effects and much for scholars to unpack. But the five papers in this symposium issue—complimentary and critical, underpinned by various methods, and from both EU and US perspectives—make an excellent start. With gratitude for the careful analysis and trenchant observations of the symposium presenters and these five authors, and for the able stewardship of the *BTLJ* symposium editors, I commend this collection to you.

108. *Id.* at 1081–83.

This Page Intentionally Left Blank.

This Page Intentionally Left Blank.

RISING ABOVE LIABILITY: THE DIGITAL SERVICES ACT AS A BLUEPRINT FOR THE SECOND GENERATION OF GLOBAL INTERNET RULES

Martin Husovec[†]

ABSTRACT

Twenty-five years ago, in 1998, the United States Congress developed a blueprint for the global regulation of the internet. Section 512 of the Digital Millennium Copyright Act (DMCA) recognized that user-generated content will be crucial to most digital services and offered up-front assurances from liability to some providers subject to conditions. What started as a sectorial conditional immunity system in copyright law was immediately scaled up into an all-encompassing horizontal rulebook in the European Union through the E-Commerce Directive (ECD) in 2000—recently updated into the Digital Services Act (DSA). The last two decades have largely validated the DMCA's conditional immunity as a feasible baseline approach to the regulation of internet communications that power global exchanges of ideas, goods, and services. However, the conditional immunity model has its limits. It was not designed to offer a complex solution for new challenges. The DSA is the first comprehensive attempt to create a second generation of rules for digital services that rely on user-generated content. Unlike previous sectorial initiatives, its approach is sweepingly horizontal. The DSA requires some level participation from both state and non-state institutions for its system of checks and balances to work, and some of its solutions can be "too European." However, the principles behind the DSA could be useful in other jurisdictions—perhaps even in the United States. The United Kingdom, which is currently developing its own set of post-Brexit rules, continues to build on some of the same principles as the DSA.

TABLE OF CONTENTS

DOI: https://doi.org/10.15779/Z38M902431

† Associate Professor of Law at London School of Economics and Political Science (LSE). I am grateful for feedback from editors and reviewers which enriched this Article. The mistakes are solely mine.

I. INTRODUCTION

Twenty-five years ago, in 1998, the United States Congress developed a blueprint for the global regulation of the internet. Section 512 of the Digital Millennium Copyright Act[1] (DMCA) recognized that user-generated content will be crucial to most digital services and offered up-front assurances from liability to some providers subject to conditions. What started as a sectorial, conditional immunity system in copyright law was immediately scaled up into an all-encompassing horizontal rulebook in the European Union through the E-Commerce Directive (ECD) in 2000[2]—recently updated into the Digital Services Act (DSA).[3] The two jurisdictions inspired many other countries to start granting conditional immunity—liability exemptions that require at least providers' knowledge of others' actions to expose them to liability for those actions.[4]

1. 17 U.S.C. § 512.

2. Directive 2000/31 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, O.J. (L 178), 1–16 (commonly and hereinafter referred to as the E-Commerce Directive).

3. 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), O.J. (L 277) 1 EU.

4. *See, e.g.*, The Information Technology Act, 2000, § 79 (Indian law covering conduit and hosting services); Information Technology Framework Act, R.R.Q. 2001, c C-1.1 (Canadian law covering hosting and search engine services); Lei No. 12.965, de 23 de Abril de 2014, Diario Oficial da Uniao [D.O.U] de 24.04.2014 (Brazilian law covering conduit and hosting services). Attempts to introduce exemptions sometimes took different turns; for example, South Korean liability exemptions were turned into liability norms. Act on Promotion of Information and Communications Network Utilization and Information Protection, art. 44-2, *translated in* Korea Legislation Research Institute's online database, https://elaw.klri.re.kr/eng_service/main.do (search required).

Unlike its older sister, § 230[5] of the Communication Decency Act (CDA) adopted in 1996,[6] § 512 of the Digital Millennium Copyright Act is not widely credited as having created the internet.[7] Yet, upon closer look, while § 230 of the CDA might continue to guarantee the internet as we know it in the legal system of the United States, it is the DMCA's model that continues to run the internet globally. For many countries for which § 230 offers a constitutionally unacceptable immunity model for application-layer services,[8] the DMCA offers a more acceptable version. The DMCA-style conditional immunity is therefore also increasingly present in bilateral trade agreements.[9] If we ever witness international harmonization on the issue, this type of conditional immunity model is probably more likely to prevail.[10]

In Europe, conditional immunity was powerfully used in the infancy of new digital markets to unite countries under one set of rules. The ingenuity of

5. 47 U.S.C. § 230.

6. Today's broad reading of § 230 CDA is a result of the judicial reading in Zeran v. Am. Online Inc., 129 F.3d 327 (4th Cir. 1997) that rejected a narrower understanding that would allow distributors to be held liable based on their knowledge of illegal content, and Batzel v. Smith, 333 F.3d 1018, 1033 (9th Cir. 2003) that allowed providers to participate in the selection process to a limited degree.

7. Kosseff makes this point most forcefully in his book. *See generally* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (Cornell Univ. Press 2019).

8. In the European legal system, denial of remedy in cases like *Batzel*, 129 F.3d 327 or *Zeran*, 333 F.3d 1018 would constitute violation of Article 8 of the European Convention on Human Rights (ECHR), which is evident in cases like K.U. v. Finland, App. No. 2872/02 (Dec. 2, 2008), https://hudoc.echr.coe.int/fre?i=001-89964; Delfi AS v. Estonia, App. No. 64669/09, ¶ 110 (Jun. 16, 2015), https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-155105&filename=001-155105.pdf; and most recently Sanchez v. France, App. No. 45581/15, ¶ 162 (Sept. 2, 2021), https://hudoc.echr.coe.int/fre?i=001-211599 ("While the Court acknowledges that important benefits can be derived from the internet in the exercise of freedom of expression, it has also found that the possibility of imposing liability for defamatory or other types of unlawful speech must, in principle, be retained, constituting an effective remedy for violations of personality rights").

9. Daniel Seng, *The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices*, 18 VA. J. L. & TECH. 369, 374 (2014) ("[T]he DMCA safe harbors have indeed gone global. And the world has embraced the DMCA."). Seng lists some of the FTAs at pages 373–75.

10. *See* WTO, WTO Electronic Commerce Negotiations Updated Consolidated Negotiating Text, WTO INF/ECON/62/Rev.2 (Sept. 2021) (limiting liability through Article B.1(2)). However, even Article 19.17.2 of the Canada-US-Mexico Trade Agreement, which contains a provision inspired by Section 230 of the CDA, was interpreted by Canadian courts as permitting a Canadian DMCA-inspired notice-based liability exemption in Article 22 of the IT Framework Act. Superior Court of Québec, A.B. v. Google LLC, 2023 QCCS 1167, https://www.canlii.org/en/qc/qccs/doc/2023/2023qccs1167/2023qccs1167.pdf. As noted by judges: "Article 19.17.2 CUSMA does not require Canada to have an immunity provision that is identical to the expansiveness of the American provision, section 230(c)(1) CDA." *Id.* ¶ 182.

the European solution rests in focusing on a one-size-fits-all compromise to rule the legal system of each of its Member States instead of searching for compromises in areas of unharmonized domestic law. Thus, conditional immunity was held as a single standard to which liability in all areas of law in the Union must converge. Section 4 of the E-Commerce Directive greatly simplified the immunity part of § 512 of the DMCA by stripping it of its tricky parts.[11] This allowed technology companies to retain the benefits of the European Union's E-Commerce Directive regime by simply complying with more demanding U.S. copyright law. In practice, the much more detailed DMCA rules about notice-and-takedown choreography became the de facto standard across the world.[12]

The last two decades have largely validated the DMCA's conditional immunity as a feasible baseline approach to the regulation of internet communications that power global exchanges of ideas, goods, and services. However, the conditional immunity model has its limits. It was not designed to offer a complex solution for new challenges. Firstly, many of them were not known or debated at the time. Second, only a tiny fraction of humanity used the internet, and if people did use it, it was not a large part of their lives. At the time of the E-Commerce Directive's adoption in 2000, less than seven percent of the world's population used the internet.[13]

By 2016, a new mainstream sentiment concerning digital services started spreading in Europe and the United States. The Court of Justice of the European Union's (CJEU) newly invented "right to be forgotten" was rapidly taking off and putting pressure on the responsibility of search engines to individuals.[14] Facebook's neglect of content moderation in Myanmar exposed the grave risks of providers' chronic under-investment in less lucrative

11. E-Commerce Directive, O.J. (L 178), 1–16. The E-Commerce Directive did not incorporate general requirements, such as the implementation of a reasonable repeat-infringer policy (§ 512(i)(1)(A)), standard technical measures (§ 512(i)(1)(B)), or special requirements, such as lack of "a financial benefit directly attributable to the infringing activity" (§ 512(c)(1)(B)). On the other hand, in contrast to the DMCA, the ECD opens the doors much more extensively to injunctions.

12. Seng, *supra* note 9; Jennifer M. Urban, Joe Karaganis & Brianna Schofield, *Notice and Takedown in Everyday Practice*, UC BERKELEY PUB. L. RSCH. PAPER NO. 2755628 (2017), https://ssrn.com/abstract=2755628 ("Beyond its influence as a model, the DMCA also operates as de facto international law because the vast majority of notices are sent to US-based companies, which operate under it.").

13. *Individuals Using the Internet*, WORLD BANK, https://data.worldbank.org/indicator/IT.NET.USER.ZS (last visited Sept. 8, 2023).

14. Case C-131/12, Google Spain SL v. Agencia Espanola de Proteccion de Datos, ECLI:EU:C:2013:424 (Jun. 25, 2013).

markets.[15] The run-up to the 2016 U.S. elections inevitably politicized the topic of content moderation on social media. Social media in Europe was caught in the middle of the European migration crisis, which surfaced incredible amounts of organized support—but also toxic hate speech—among the general population.[16] It's likely that at this point, European governments began to question if self-regulation was the right approach. It became evident that the space that the conditional immunity model left to providers must soon be filled by regulation.

The DSA is the first comprehensive attempt to create a second generation of rules for digital services that rely on user-generated content. Unlike previous sectorial initiatives,[17] its approach is sweepingly horizontal. The DSA requires some level participation from both state and non-state institutions for its system of checks and balances to work, and some of its solutions can be "too European." However, the *principles* behind the DSA could be useful in other jurisdictions—perhaps even in the United States. The United Kingdom, which is currently developing its own set of post-Brexit rules, continues to build on some of the same principles as the DSA.

My hope is that these high-level principles might form the basis for a dialogue between liberal democracies about how to best regulate user-generated content services.[18] After all, if Europeans in the late 1990s could simplify and scale up the U.S. rules to fit their goals, maybe today other countries can do the same with the new E.U. rules. Having interoperable policies continues to be important for the flourishing of a truly global network of communications that generates unprecedented benefits for humanity.

---

15. Steve Stecklow, *Hatebook*, REUTERS (Aug. 15, 2018), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

16. EUR. COMM'N, RACISM AND DISCRIMINATION IN THE CONTEXT OF MIGRATION IN EUROPE (Mar. 31, 2017), https://ec.europa.eu/migrant-integration/library-document/racism-and-discrimination-context-migration-europe_en; EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS, CURRENT MIGRATION SITUATION IN THE EU: HATE CRIME, (Nov. 2016), https://fra.europa.eu/sites/default/files/fra_uploads/fra-2016-november-monthly-focus-hate-crime_en.pdf.

17. The German and French parliaments previously adopted anti-hate speech rules that mostly imposed tight reaction periods for providers, see Netzwerkdurchsetzungsgesetz [NetzDG] [The Network Enforcement Act of 2017], Jan. 9, 2017, Bundesgesetzblatt, Teil I [BGBL I] at 3352 (Ger.); Loi 2020-766 du 24 juin 2020 visant a lutter contre les contenus haineux sur internet [Law 2020-766 of 24 June 2020 to Combat Hate Content on the Internet] [Loi Avia], Journal Officiel de la Republique Francaise [J.O.] [Official Gazette of France] (Jun. 25, 2008), p. 156.

18. For a broader debate, see MARTIN HUSOVEC, PRINCIPLES OF THE DIGITAL SERVICES ACT (Oxford Univ. Press forthcoming 2024).

II. FROM DMCA TO DSA

The European regulation of user-generated content services is clearly inspired by U.S. law. In this section, I first briefly explain how this has happened and then why, despite today's controversies, conditional immunity is an approach that has been arguably validated over the last two decades.

A. A BRIEF HISTORY OF LIABILITY EXEMPTIONS

Unlike the first liability exemption of its kind, § 230 of the CDA, which did not attract much stakeholder attention at the time,[19] § 512 of the DMCA is a product of hard negotiations between content industries and technology companies.[20]

The debate about the copyright liability of providers was power-charged by the 1995 White Paper issued by the Clinton administration's Information Infrastructure Task Force, which supported its view with two earlier rulings from U.S. courts regarding bulletin boards.[21] The White Paper presented strict direct copyright liability of providers, including internet access providers, as a given and argued that it would be "premature to reduce the liability of any type of service provider[.]"[22] The report implicitly encouraged plaintiffs to test the waters against all providers, not just bulletin boards. In 1995, the Church of Scientology sued another bulletin board operator, along with an internet access provider, Netcom, in a U.S. district court.[23] While the court quickly ruled that companies are not directly and strictly liable, it established that contributory knowledge-based liability remains an option.[24] The *Netcom* case undoubtedly put telecommunications companies, an established industry, on alert about

19. JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET 67 (Cornell Univ. Press 2019) ("Despite its monumental statements about a new, hands-off approach to the internet, the bill was virtually unopposed on Capitol Hill. Lobbyists focused primarily on the telecommunications bill's impacts on phone and cable television service.").

20. UNITED STATES COPYRIGHT OFFICE, SECTION 512 OF TITLE 17: A REPORT OF THE REGISTER OF COPYRIGHTS 18 (2020), https://www.copyright.gov/policy/section512/section-512-full-report.pdf.

21. Playboy Enters., Inc. v. Frena, 839 F. Supp. 1552 (M.D. Fla. 1993); Sega Enters. Ltd. v. MAPHIA, 857 F. Supp. 679 (N.D. Cal. 1994).

22. INFO. INFRASTRUCTURE TASK FORCE, INTELLECTUAL PROPERTY AND THE NATIONAL INFORMATION INFRASTRUCTURE: THE REPORT OF THE WORKING GROUP ON INTELLECTUAL PROPERTY RIGHTS 128 (1995), https://www.eff.org/files/filenode/DMCA/ntia_dmca_white_paper.pdf [hereinafter *White Paper*].

23. Religious Tech. Ctr. v. Netcom On-Line Comm. Servs., Inc., 907 F. Supp. 1361 (N.D. Cal. 1995).

24. Providers were not acting volitionally with respect to copyright-relevant acts, and thus cannot be held strictly liable. However, given that Netcom was served with notice, this triggered a duty to investigate the matter to avoid contributory copyright liability.

potential liability risks even though the outcome was favorable to them.[25] Those companies eventually lobbied to codify *Netcom* in the DMCA.[26]

After the White Paper's proposals failed in the 104th United States Congress,[27] the next Congressional session starting in January 1997 hoped to find a quick solution between opposing interests to successfully implement the World Intellectual Property Organization (WIPO) Internet Treaties.[28] In the legislative process, liability exemptions became a precondition to the passage of the entire piece of legislation.[29] As noted by the Senate Judiciary Committee Report, although the issue "[was] not expressly addressed in the actual provisions of the WIPO treaties, the Committee is sympathetic to the desire of . . . service providers to see the law clarified in this area."[30] It was understood that "without clarification of their liability, service providers may hesitate to make the necessary investment in the expansion of the speed and capacity of the internet."[31]

The Judiciary Committee report initially only included a liability exemption for mere conduits.[32] The final compromise with four liability exemptions—conduits, caching, hosting, and information location tools—only materialized after three months of direct negotiations between providers and content

25. JESSICA D. LITMAN, DIGITAL COPYRIGHT 128 (Prometheus Books 2d ed. 2006).

26. *See* H.R. Rep. No. 105-551, pt. 1 at 11 (1998), http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=105_cong_reports&docid=f:hr551p1.105.pdf ("As to direct infringement, liability is ruled out for passive, automatic acts engaged in through a technological process initiated by another. Thus, the bill essentially codifies the result in the leading and most thoughtful judicial decision to date: Religious Tech. Ctr. v. Netcom On-Line Commc'n Servs., Inc., 907 F. Supp. 1361 (N.D. Cal. 1995). In doing so, it overrules those aspects of Playboy Enters., Inc. v. Frena, 839 F. Supp. 1552 (M.D. Fla. 1993), insofar as that case suggests that such acts by service providers could constitute direct infringement, and provides certainty that *Netcom* and its progeny, so far only a few district court cases, will be the law of the land").

27. JESSICA D. LITMAN, DIGITAL COPYRIGHT 122 (Prometheus Books 2d ed. 2006).

28. *Id.* at 126, 130 ("After the bruising copyright fight in the last Congress, it wanted to satisfy the Hollywood and Silicon Valley communities but did not want to have to expend significant pollical capital to do so.").

29. *Id.* at 134–35.

30. S. Rep. No. 105-190, at 19 (1998). WCT only indirectly mentions the position of providers that can be found in an agreed statement to Article 8 which was the result of lobbying by providers and telecommunications companies who failed to include liability exemptions into the WIPO Internet Treaties themselves. *See* MIHALY FICSOR, THE LAW OF COPYRIGHT AND THE INTERNET: THE 1996 WIPO TREATIES, THEIR INTERPRETATION AND IMPLEMENTATION 509 (Oxford Univ. Press 2002).

31. S. Rep. No. 105-190 (1998).

32. H.R. Rep. No. 105-551 (1998).

owners.[33] The compromise text was already captured in the Commerce Committee in June 1998,[34] which argued that:[35]

> Title II preserves strong incentives for service providers and copyright owners to cooperate to detect and deal with copyright infringements that take place in the digital networked environment. At the same time, it provides greater certainty to service providers concerning their legal exposure for infringements that may occur in the course of their activities.

The DMCA was signed into law in October 1998. Congress was "keenly aware that other countries will use U.S. legislation as a model."[36]

In Europe, the European Commission published a communication to the European Parliament and Council in October 1996 explaining that providers will need legal assurances to be able to properly operate in the online market. The communication stated that:[37]

> Internet access providers and host service providers play a key role in giving users access to Internet content. It should not however be forgotten that the prime responsibility for content lies with authors and content providers. It is therefore essential to identify accurately the chain of responsibilities in order to place the liability for illegal content on those who create it . . . . The law may need to be changed or clarified to assist access providers and host service providers, whose primary business is to provide a service to customers, to steer a path between accusations of censorship and exposure to liability.

Two years later, the European Commission introduced the proposed E-Commerce Directive. Its Section 4 included three liability exemptions—conduits, caching, and hosting. At the time, only two European countries had liability exemptions. Germany adopted its two horizontal liability exemptions in July 1997[38] (termed the IuKDG) and Sweden adopted a law on bulletin

33. S. Rep. No. 105-190 at 7 ("These negotiations continued under the supervision of the Chairman for three months, from January to April, 1998."). *See* JESSICA D. LITMAN, DIGITAL COPYRIGHT 135 (Prometheus Books 2d ed. 2006).

34. H.R. Rep. No. 105-551 (1998).

35. *Id.*

36. S. Rep. No. 105-190 (1998).

37. Communication from the Commission on Illegal and Harmful Content on the Internet, COM (1996) 487 final, at 12–13 (Oct. 16, 1996).

38. Informations- und Kommunikationsdienste-Gesetz [IuKDG] [Act on Information and Communication Services of 1997] (Jun. 13, 1997), BGBL I at 52. Section 5 of IuKDG was elegantly condensed in the following four parts establishing the following: (1) liability is for own content remains to be governed by generally applicable law; (2) liability for other people's content on services that can be "used by others" ("die sie zur Nutzung bereithalten")

boards in May 1998.[39] Both the new German laws and DMCA made the basic distinction between services giving "access" and "space" to other people's information. Thus, unlike § 230 of the CDA, both the IuKDG and the DMCA differentiated liability exemptions based on the proximity of providers to users' actions. Conduits as distant facilitators were given the broadest immunity, while nearer hosts were granted more cautious exemptions based on their knowledge. In terms of scope, the German laws seemed more far-reaching, as they extended to conduits and all services which were being "made available for use[.]"[40] In contrast, § 512 focused on specific technical functions—conduits, caching, storage, and information location tools.

The main inspirations for Section 4 of the E-Commerce Directive were § 512 of the DMCA and Section 5 of the IuKDG. The Commission borrowed three liability exemptions from the DMCA, and a horizontal approach from the IuKDG. Unlike the U.S. copyright statute, the E.U. proposal was not driven by the need to implement the WIPO Internet Treaties but rather the European Union's desire to create an internal market without frontiers in the early stage of the internet's development. The newly found U.S. copyright compromise concerning the internet was thus extended to all areas of law.

The European Commission's proposal was adopted in June 2000. The Commission's approach followed the American definitions of categories of services and thus arguably narrowed down the scope of services which could rely on conditional immunity. For instance, the German provision could have easily covered information location tools, which were not given any explicit immunity.[41] In 2002, the E-Commerce Directive became law for fifteen E.U. Member States and, two years later, for another ten newly joined member states. As of now, both the E-Commerce Directive and the Digital Services Act apply across 27 member states. The Digital Services Act, as an E.U. regulation, is applicable directly without a need for local implementation. Post-Brexit, the United Kingdom so far has not repealed its implementation of the ECD liability exemptions, and E.U. case law until the end of 2020 continues

is possible only once they acquire knowledge; (3) liability for giving access to other's people content is barred; (4) blocking remains possible in accordance with generally applicable law.

39. Lag om ansvar for elektroniska anslagstavlor (Svensk forfattningssamling [SFS] 1998:112) (Swed.).

40. Section 5(2) of the IuKDG ("die sie zur Nutzung bereithalten").

41. Their qualification under hosting is complicated in the European Union due to questions about whether the information is "provided by" the indexed websites in all cases.

to be binding in British courts.[42] The United Kingdom is currently developing its own set of online safety rules that will supplement the existing exemptions.[43]

While the differences between the statutory language of the E.U. and U.S. laws were not insignificant, they were mostly reconcilable.[44] Generally, one can say the European Union simplified the DMCA—but also omitted some of its key components. In particular, the European Union omitted a liability exemption for information location tools and the DMCA's elaborate conditions for injunctive relief; the latter omission became a major point of divergence. Under E.U. law, injunctions were, in principle, left unconstrained if they conformed to the notion of "specific" monitoring.[45] The DMCA, in contrast, limited injunctions with a myriad of conditions.[46] As a result, while under § 512 of the DMCA all preventive injunctions—such as those imposing filters or website blocking—remained practically impossible, under Section 4 of the ECD they soon became the primary driver of European litigation efforts.[47] Eventually, the CJEU allowed plaintiffs who successfully litigated their grievances to seek injunctions that saddled providers with more responsibility to identify infringing content.[48]

The introduction of liability exemptions in the United States and European Union was clearly driven by the same rationale: to encourage investment by giving more legal certainty. As a result, the legal system can "steer a path

42. *See* The Electronic Commerce (EC Directive) Regulations 2002, SI 2001/2555 (Eng.), https://www.legislation.gov.uk/uksi/2002/2013, along with the European Union (Withdrawal) Act 2018, c.16, § 6 (UK), https://www.legislation.gov.uk/ukpga/2018/16/section/6/enacted.

43. *See* Online Safety Bill 2022-3, HL Bill [362] (UK), https://bills.parliament.uk/bills/3137.

44. Miquel Peguera, *The DMCA Safe Harbors and Their European Counterparts: A Comparative Analysis of Some Common Problems*, 32 COLUM. J. L. & ARTS 481, 481–82 (2009).

45. *See* E-Commerce Directive, art. 15(1), O.J. (L 178), 13; *id.*, r. 47, 6 ("Member States are prevented from imposing a monitoring obligation on service providers only with respect to obligations of a general nature; this does not concern monitoring obligations in a specific case and, in particular, does not affect orders by national authorities in accordance with national legislation.").

46. 17 U.S.C. § 512(j) (significantly limiting forms of injunctions) and 17 U.S.C. § 512(m) ("Nothing in this section shall be construed to condition the applicability of [liability exemptions on] a service provider monitoring its service or affirmatively seeking facts indicating infringing activity, except to the extent consistent with a standard technical measure ").

47. *See generally* MARTIN HUSOVEC, INJUNCTIONS AGAINST INTERMEDIARIES IN THE EUROPEAN UNION: ACCOUNTABLE BUT NOT LIABLE? (2017).

48. The biggest shift was brought by the CJEU in Case C-18/18, Glawischnig-Piesczek v. Facebook Ir. Ltd., ECLI:EU:C:2019:458 (June 4, 2019).

between accusations of censorship and exposure to liability."[49] The problem was acute to different degrees in different areas of law; however at the time, when CEOs of some technology companies were sentenced in criminal proceedings for distributing pornography, the concerns certainly were not trivial or overblown.[50] The growing national case law in the E.U. was seen as both too unpredictable and too unwieldy to provide clarity on how to reliably build a legal framework for the new environment that showed so much promise. Since user-generated content is so central to the digital *communications* network, the liability question was *the* question of internet regulation.

B. LIABILITY EXEMPTIONS AND SPECIFICITY OF THE INTERNET

The law has a key role in guaranteeing the shape and form of the internet. The decentralized nature of the internet as a network is inseparable from the underlying liability regime for those who facilitate its functioning. Without the sympathy of the law, there is no internet as we know it. In a hypothetical world where technology facilitates decentralization but the law provides incentives against it, no rational actors would have created spaces or tools without editorial control. A liability regime for the actions of others is a key incentive factor. Unless legislatures want to reinstate editors, some form of conditional immunity is necessary.

The European plan for most of the user-generated content services that host content is to ask victims to use nonjudicial notice-and-takedown systems and rely on the help of authorities, including courts, where possible. This mix of routes, while more generous to victims than the immunity-based framework of § 230 of the CDA, constrains victims' and the state's abilities to solve any social problem. But it does so for a good reason: to maintain the benefits of a decentralized communication network. By observing liability exemptions, everyone commits to constraining themselves in order to facilitate the emergence of an environment from which everyone can benefit. This is the essence of the digital social contract.

Strict liability, in contrast, demands total control, and such legal rules would become very expensive for society. By way of analogy, printers who are strictly liable for everything they print for others would inevitably need to first read and vet everything they print. Printing would become very slow and

49. Communication from the Commission on Illegal and Harmful Content on the Internet, COM (1996) 487 final, at 13 (Oct. 16, 1996).

50. In Germany, the law was also a reaction to the controversial CompuServe case. *See* Stefan Engel-Flechsig, Frithiof Maennel & Alexander Tettenborn, *Das neue Informations- und Kommunikationsdienste-Gesetz,* NJW 1997 2981, 2984 (1997).

expensive as a result, and people would be increasingly unable to use it to share ideas.

The link between such liability and freedom of speech has been recognized by the United States Supreme Court, the European Court of Human Rights (ECtHR), and the Court of Justice of the European Union in their human rights jurisprudence.[51] These highest courts set the limits for how user-generated services can be regulated by legislatures responsible for a little over 1 billion people.[52] At the moment, the strict liability of providers for user-generated content is treated on both sides of the Atlantic as unthinkable and fundamentally unconstitutional. U.S. and E.U. courts in unison continue to advocate for "medium-specific"[53] or "graduated and differentiated"[54] regulation that differs from regulation of editorial media. The European Court of Human Rights, for instance, despite its complex case law,[55] makes it clear

51. *See* Reno v. Am. C.L. Union, 521 U.S. 844 (1997); Case C-401/19, Poland v. Council & Eur. Parliament, ECLI:EU:C:2021:613 (July 15, 2021); MTE and Index.hu v. Hungary, App. No. 22947/13 (Feb. 2, 2016), https://hudoc.echr.coe.int/fre?i=001-160314.

52. To be precise: 690 million in the Council of Europe, of which 447 million are in the European Union, and then 331 million in the United States. *COE—Council of Europe 2023*, COUNTRY ECON., https://countryeconomy.com/countries/groups/council-europe (last visited Sept. 9, 2023) (noting that Russia is not a member anymore).

53. The "medium-specific" approach is relied upon by Judge Dalzell in Am. C.L. Union v. Reno, 929 F. Supp. 824, 873 (E.D. Penn. 1996) ("My examination of the special characteristics of internet communication, and review of the Supreme Court's medium-specific First Amendment jurisprudence, lead me to conclude that the internet deserves the broadest possible protection from government-imposed, content-based regulation.").

54. *See* Council of Eur., Recommendation on a New Notion of Media, CM/Rec (2011)7 ¶ 7 (2013), https://edoc.coe.int/en/media/8019-recommendation-cmrec20117-on-a-new-notion-of-media.html ("A differentiated and graduated approach requires that each actor whose services are identified as media or as an intermediary or auxiliary activity benefit from both the appropriate form (differentiated) and the appropriate level (graduated) of protection and that responsibility also be delimited in conformity with Article 10 of the European Convention on Human Rights and other relevant standards developed by the Council of Europe."), cited by the ECtHR in Delfi AS v. Estonia, App. No. 64569/09, ¶ 113 (June 16, 2015), https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-155105&filename=001-155105.pdf.

55. The European Court of Human Rights signaled that the Member of the Council of Europe might be exceptionally allowed to legislate that discussion forum providers should do more than only operate notice-and-takedown to avoid civil liability for hate speech. *See Delfi AS*, App. No. 64569/09. The decision is often mischaracterized as imposing a particular liability framework on the states. The case law only gives discretion to states to do this. Even more controversially, in a case concerning Facebook page administrators, the ECtHR also allowed the criminal financial liability of politicians for comments posted by others if they have some—albeit not specific—knowledge about those comments. Sanchez v. France, App. No. 45581/15, ¶ 162 (Sept. 2, 2021), https://hudoc.echr.coe.int/fre?i=001-211599. However, neither of the two rulings allows unconditional strict liability.

that "the notice-and-take-down-system could function in many cases as an appropriate tool for balancing the rights and interests of all those involved."[56]

The different treatment of the internet as a medium is not an act arising from a rose-tinted, naïve love for new technology.[57] It comes down to what the American Judge Dalzell in 1996 called "the special attributes of internet communication" that make it "the most participatory form of mass speech yet developed[.]"[58] The Court of Justice of the European Union referred to the internet as "one of the principal means by which individuals exercise their right to freedom of expression and information[,]"[59] and supported the view of the European Court of Human Rights that "user-generated expressive activity" is "an unprecedented platform for the exercise of freedom of expression."[60]

For Judge Dalzell and his colleagues in the late 90s, these "special attributes" were very low barriers to entry for speakers and readers leading to "astoundingly diverse content" and "significant access to all who wish to speak in the medium[.]"[61] For top European judges looking at it in the early 2010s, the special attributes of the internet are: its "accessibility"; its "capacity to store and communicate vast amounts of information"; its ability to support "user-generated expressive activity"; and its role in "facilitating the dissemination of information in general[.]"[62]

---

56. MTE and Index.hu v. Hungary, App. No. 22947/13, ¶ 91 (Feb. 2, 2016), https://hudoc.echr.coe.int/fre?i=001-160314 (presented as an application of the Grand Chamber decision in Delfi AS v. Estonia).

57. Discussing "internet exceptionalism" is beyond the space limitations of this Article, but the two essays worth reading on this are Mark Tushnet, *Internet Exceptionalism: An Overview from General Constitutional Law*, 56 WM. & MARY L. REV. 1637 (2015), and Tim Wu, *Is Internet Exceptionalism Dead?*, *in* THE NEXT DIGITAL DECADE: ESSAYS ON THE FUTURE OF THE INTERNET (Berin Szoka et al. eds., 2011), https://scholarship.law.columbia.edu/faculty_scholarship/1676.

58. Am. C.L. Union v. Reno, 929 F. Supp. 824, 867, 883 (E.D. Pa. 1996).

59. Case C-401/19, Poland v. Council and European Parliament, ECLI:EU:C:2021:613, ¶ 46 (July 15, 2021).

60. Delfi AS v. Estonia, App. No. 64569/09, ¶ 110 (June 16, 2015), https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-155105&filename=001-155105.pdf.

61. *Am. C.L. Union*, 929 F. Supp. at 877.

62. Case C-401/19, Poland v. Council and European Parliament, ECLI:EU:C:2021:613 (July 15, 2021), at ¶ 46 ("In the light of their accessibility and their capacity to store and communicate vast amounts of information, internet sites, and in particular online content-sharing platforms, play an important role in enhancing the public's access to news and facilitating the dissemination of information in general, with user-generated expressive activity on the internet providing an unprecedented platform for the exercise of freedom of expression"). The Grand Chamber is citing the ECtHR decisions in Cengiz and Others v. Turkey, Apps. No. 48226/10 and 14027/11, ¶ 52 (Dec. 1, 2015), https://hudoc.echr.coe.int/app/conversion/docx/pdf?library=ECHR&id=001-159188&filename=CASE%20OF%20

Any exemptions are naturally suspect to legislative favoritism towards the industry. And liability exemptions can be naturally fashioned in different ways. However, it has long been recognized that the DMCA-modelled liability exemptions are not necessarily major liability carve-outs when compared to ordinary applications of liability.[63] As noted by Advocate General Jääskinen, "these provisions are better qualified as restatements or clarifications of existing law than exceptions thereto."[64] This is also clear when looking at the text of § 512 of the DMCA, which incorporates many requirements of American copyright secondary liability.[65] Thus, while conditional immunities like those laid out in the ECD and DMCA might bring about some changes, they are usually not major liability carve-outs. The case for internet exceptionalism is somewhat stronger with the prohibition of general monitoring. However, its strongest legitimacy is in the protection against indiscriminate surveillance of people and their content, not as a rule to protect providers against increased costs.[66]

One could object that liability exemptions are therefore not *necessary* because courts would have gradually arrived at the right solution after years of litigation by simply applying general laws. While it is impossible to prove this with a counterfactual, the early history of liability in many countries[67] and even numerous recent examples of inconsistent case law show that legislative clarity has a unique value. For instance, while the recent ECtHR case law on liability does not in principle allow the states to depart far from knowledge-based immunity for hosts, the Court is clearly incapable of fashioning a predictable

---

CENG%C4%B0Z%20AND%20OTHERS%20v.%20TURKEY.pdf&logEvent=False, and Kharitonov v. Russia, App. No. 10795/14, ¶ 33 (Jun. 23, 2020), https://hudoc.echr.coe.int/fre?i=002-12866.

63. This is obviously different in the case of § 230 of the CDA, which lifts the constitutionally compelled immunity required by the First Amendment. *See generally* Eric Goldman, *Why Section 230 Is Better Than the First Amendment*, 95 NOTRE DAME L. REV. 33 (2019).

64. Case C-324/09, L'Oreál v. eBay, ECLI:EU:C:2010:757, ¶ 136 (Dec. 9, 2010).

65. *Compare* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984), *with* Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

66. As rightly pointed out by one of the reviewers, especially when low-cost means cannot be imposed due to the prohibition, the argument about existence of material carve-outs from the general framework might be valid. However, in such cases, the different treatment is not a result of favouring companies but favouring the privacy and expression rights of their users.

67. The early controversial U.S. cases concerned defamation law. *See, e.g.*, Stratton Oakmont, Inc. v. Prodigy Servs. Co., 23 Media L. Rep. 1794 (N.Y. Sup. Ct. 1995). The early German cases, on the other hand, concerned child abuse images and protection of minors *See, e.g.*, Entscheidungen des Amtsgericht München in Strafsachen [Munich Local Court] Az. 8340 Ds 465 Js 173158/95 (May 28, 1998).

test and is constantly creating endless pockets of new, sub-case law.[68] It seems that judges trained to engage in granular balancing are less interested in devising bright-line rules. Had the E.U. statutory law not been as clear as it was, human rights law would have hardly offered predictability.

C. THE NEED FOR A SECOND GENERATION OF RULES

The last two decades have drawn contours indicating many societal challenges that require solutions, ranging from: the protection of children; problems with hate speech or terrorism; to subversive activities that attack the basis of our democratic systems. All these problems are exacerbated by the "special features" of the internet as a medium: its lack of editorial approval, low barriers of entry (including omnipresent zero cost of services), incredible speed and scale of distribution, its broad social and geographical inclusiveness, and resilience of communications. Regulators across the globe are thus rightly considering how to address these challenges.

Simply pointing to the existing digital social contract seems insufficient when the clear legislative goal of the liability exceptions was to lay down incomplete and unrestrictive rules that would allow the medium to flourish. The tendency of some stakeholders to see liability exemptions as a magical limit on any future regulation mischaracterizes their key contribution. The key contribution is not in stopping any new rules from being adopted but in keeping one set of sufficiently enabling rules on the books. In any federal system, federal liability exemptions help to coordinate national or state laws by preempting national- or state-level experimentation. This is the added benefit of such rules both in the United States and European Union. However, this does not mean that such rules must be carved in stone. In fact, the E.U. and U.S. experiences both show that the inability of federal legislatures to update federal rules can lead states to test their limits.[69]

---

68. The two leading Grand Chamber cases, Sanchez v. France and Delfi AS v. Estonia, are basically painted as exceptions in other cases like MTE and Index.hu v. Hungary. App. No. 45581/15 (Sept. 2, 2021), https://hudoc.echr.coe.int/fre?i=001-211599; Delfi AS v. Estonia, App. No. 64569/09 (June 16, 2015), https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-155105&filename=001-155105.pdf; Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, App. No. 22947/13 (Feb. 2, 2016), https://hudoc.echr.coe.int/fre?i=001-160314.

69. In Europe, the lack of early Union legislation led Germany and France to adopt their own hate speech laws for social media. *See* Case C-131/12, Google Spain SL v. Agencia Espanola de Proteccion de Datos, ECLI:EU:C:2013:424 (June 25, 2013). In the US, the lack of any federal regulation led to state laws in Florida and Texas. *See* S.B. 7072, 2021 Leg. (Fla.), https://www.flsenate.gov/Session/Bill/2021/7072/; H.B. 20, 2021 Leg., 87th Sess. (Tex.), https://capitol.texas.gov/BillLookup/History.aspx?LegSess=872&Bill=HB20.

Additionally, the harms and victims of various societal challenges come in different forms. Some harms are amplified by the design of services; others are caused by other people and only facilitated by lack of intervention. Some victims of such harms lack means, while some are well-resourced; some can use technology to uncover violations of their rights, while others cannot. Before the DSA, Section 4 of the ECD left all these concerns to self-regulation or national experimentation. However, to effectively regulate a global network, the regulatory action must be big enough for global companies to start paying attention to it. For instance, despite three decades of European data protection law, it took the GDPR—which was adopted in 2016—to fully bring the laws to everyone's attention.

Horizontal liability exemptions, such as the one found in Chapter 2 of the Digital Services Act (formerly Section 4 of the E-Commerce Directive) are about creating breathing space for speech and markets while allowing enforceability of the rights of victims, but they do not address specific challenges. The rules of the first generation—§ 230 of the CDA, § 512 of the DMCA, and Section 4 of the ECD—all suffer the same insufficiency. They excel at coordinating expectations to encourage investment but fail at offering tools to solve a wide range of societal problems that emerged along with the use of these services.

The European Union's Digital Services Act is one example of how to update the digital social contract without undermining the decentralized nature of the internet. The DSA re-affirms democratic legitimacy for the rules of conditional immunity, and even extends them on margin. Providers' liability for user-generated content thus mostly does not change.[70] What changes are regulatory expectations when companies make their decisions about other people's content or behavior, and, for some providers, what they need to think about when designing digital services. These companies are accountable to the public through regulation. However, such regulation is specifically designed for them. Instead of fitting user-generated services into ill-suited preexisting categories, they are given a regulatory category of their own based on their size and technical functions.

70. The DSA introduces a few changes to the text of the liability exemptions, which arguably expands them; especially the mere conduit liability exemption (Article 4) now applies to a broader set of infrastructure services; hosting exemption receives some minor additions (e.g., Article 6(3)), which arguably already follow from the pre-existing case law; the newly inserted Article 7 about own investigation arguably will have limited effect, and again builds upon the case law.

III. THE TWO PILLARS OF THE DSA

The Digital Services Act has two main pillars: (1) due process requirements for content moderation, and (2) risk management obligations for services. Content moderation is defined and regulated as the process of decision-making that emerges from providers' reliance on the liability exemptions, such as hosting. Risk management focuses on the system and product design of services and invites providers to consider the broader effects of their advertising infrastructure, recommendation algorithms, and other systems. The table below provides an overview of all the main DSA obligations.

Table 1

Two pillars of rules	Technical activity	Company or service size	Main types of due diligence obligations imposed by the DSA
Content moderation	• Conduit • Caching • Hosting	Companies of all sizes	• Contact points or legal agents • Clarity of terms and conditions
		Medium-size[71] and bigger companies	• Content moderation reports
	• Hosting	Companies of all sizes	• Notice submission rules • Justification of decisions • Crimes notification to authorities
	• Online platforms (a subset of hosting services)	Medium-size and bigger companies	• Prioritization of trusted flaggers • Measures against abuse • Internal appeal systems • External appeal systems • Transparency of advertising and recommender systems
Risk management	• Online platforms (a subset of hosting services)	Medium-size and bigger companies	• Protection of minors • Dark patterns • Know-Your-Client obligations of marketplaces
	• Very large online platforms (VLOPs) • Very large online search engines (VLOSEs)	45 million average active monthly users, regardless of the company's size or turnover	Upon designation by the European Commission: • Risk assessment and auditing of all design features of the product • Enhanced data access for researchers to study risks • Crisis response mechanism • Special advertising transparency • Choice on recommender systems • Internal compliance officers

A. CONTENT MODERATION

The DSA recognizes that content moderation decisions by private companies can have a large impact on people's livelihoods and their freedoms

71. Small enterprises are defined by a Council Recommendation as "an enterprise which employs fewer than 50 persons and whose annual turnover and/or annual balance sheet total does not exceed EUR 10 million." Commission Recommendation of 6 May 2003 Concerning the Definition of Micro, Small and Medium-sized Enterprises, 2003 O.J. (L 124) 36.

to share and receive information from others. The last decade has shown that companies are not always willing to invest sufficient resources into such decision-making, especially in smaller countries or markets. The E.U. legislature's solution is to regulate the *process* through which such content moderation decisions are made.

Content moderation rules must be clear and predictable (Article 14(1)), and decisions must be based on existing policies (Article 14(4)). A wide range of content moderation decisions is subject to an obligation of individual explanation (Article 17) and annual transparency reporting (Article 15). Each decision must be subject to free internal appeals (Article 20) and potentially external dispute resolution (Article 21). In addition, many expectations are purposefully vague. Notification systems must be "user-friendly" (Article 16(1)), decisions made "in a timely, diligent, nonarbitrary and objective manner" (Article 16(6)), and enforcement practices must pay "due regard to the rights and legitimate interests of all parties involved" (Article 14(4)).

The above provisions set the rules for the *process side* of content moderation. Underlying contractual rules about acceptable content or behavior remain to be set by companies. However, providers' rule-making space is indirectly constrained by the limits placed on the procedure. Due to the obligation to disclose rules upfront (Article 14(1)), companies cannot retroactively change their policies, or invent sanctions ex post facto that have no basis in their existing rules (Article 14(4)). Providers can continue to contractually constrain speech beyond illegality according to their preferences; however, they must apply the rules in a nonarbitrary and non-discriminatory manner. Any contractual policies will be interpreted by out-of-court dispute settlement bodies which cannot consider "secret rulebooks" of any kind.

This clearly shows that the DSA does not take away all the content moderation discretion from platforms. It generally does not limit what legal content can be prohibited by providers under their community guidelines—that is a power that providers retain. Thus, if providers do not like how out-of-court bodies read their rules, they can change them and make them clearer. But once they put the rules in black and white, they cannot claim a contrary meaning without actually changing them. The DSA limits only some grossly unfair policies (Article 14(4)) that would likely already struggle with other areas of explicit legal prohibitions, such as consumer law.

The goal of these procedural guarantees is to script the process of content moderation into a tighter choreography that better reflects the impact of content moderation decisions on individuals. The mix of very specific procedural rules and vague aspirational regulatory expectations is meant to provide the basis for standard-setting but also a north star for content

moderation decision-making. For individuals, the rules give them more credible due process rights which go well beyond the standard delivered by markets alone.

Scholars like Douek criticize regulatory due process expectations as an unnecessary "process theatre"[72] which does not solve the overall problems because it resembles "using a teaspoon to remove water from a sinking ship."[73] But is that the right framing? First, for the affected individuals, even a teaspoon of hope that their grievances can lead to proper resolution are good enough reasons to institute them. This rationale is hardly diminished by the fact that such personal disputes do not resolve the larger problems. Second, the DSA tries to use the personal dimension of disputes as a source of broader learning, something favored by Douek, and as a pressure to improve the overall quality of the processes.[74] Finally, for very large online platforms, content moderation is only one part of their overall risk management.

B. RISK MANAGEMENT

The DSA's second pillar concerns risk management, which comprises a set of rules that address how companies design their products and other behind-the-scenes processes. Unlike the United Kingdom's upcoming Online Safety Bill,[75] the DSA legislatively and explicitly doses responsibility by the size or impact of the services. Risk must be mitigated only by digital services known as online platforms; that is, services that distribute user-generated content to the public as their main feature.[76] Platforms operated by micro and small companies—those employing less than fifty employees or earning less than ten million euros annually[77]—have no risk management obligations. The intuition

72. Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526, 577 (2022).

73. *Id.* at 606.

74. *See* Digital Services Act art. 21, 2022 O.J. (L 277) against the background of a lab experiment concerning ADR system as a solution to rational bias against over-blocking. Lenka Fiala & Martin Husovec, *Using Experimental Evidence to Improve Delegated Enforcement*, 71 INT'L REV. OF L. & ECON. (2022), https://www.sciencedirect.com/science/article/pii/S0144818822000357.

75. The UK's Online Safety Bill, *supra* note 43, is still in the legislative process. According to the recent impact assessment, out of 25 thousand forecasted regulated organizations, roughly 20 thousand are likely micro. Online Safety Bill 2022-3, Impact Assessment ¶ 109, (UK), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061265/Online_Safety_Bill_impact_assessment.pdf.

76. Digital Services Act art. 3(i), 2022 O.J. (L 277).

77. *See* Commission Recommendation of 6 May 2003 Concerning the Definition of Micro, Small and Medium-sized Enterprises, 2003 O.J. (L 124) 36–41, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32003H0361.

behind this is that with more power comes more responsibility. For risk management, platforms are divided into two groups:

- In the *lower* tier, mid-sized or bigger companies are subject to limited and prescriptive rules covering design practices.
- In the *upper* tier, online platforms or search engines which serve more than 45 million monthly active users in the European Union are subject to a more expansive and vaguer set of rules: general risk management.

The companies in the lower tier must mostly think about how their product design protects *minors*, and against *manipulative* and *aggressive* practices—also known as dark patterns. The companies in the upper tier must do the same, plus much more. Specific businesses are designated as quasi-public squares where many Europeans meet and exchange. By state designation, they are placed under special regulatory dialogue with the European Commission and national authorities (regulators) about *any relevant risks to democratic institutions and individuals*, including risks to people's freedoms and well-being. Given that these are interests that are hard to delineate, the scope is very broad.

In the first round in Spring 2023,[78] the following digital services were designated:

- *Social media*: Instagram, LinkedIn, Pinterest, Snapchat, TikTok, Twitter, Facebook, YouTube
- *Marketplaces*: Alibaba, AliExpress, Amazon Store, Booking.com, Google Shopping, Zalando
- *App stores*: Apple AppStore, Google Play
- *Other*: Google Maps, Wikipedia
- *Search engines*: Bing, Google Search.

The newly imposed risk management obligations are clearly meant to legislatively complement liability assurances with some societal responsibilities as to trust, safety, and fairness in these services.

Risk management is a result of two realizations. First, the importance of design to the health of any ecosystem. This point has been reinforced by

78. *See* the designations published in European Commission Press Release IP/23/2413, Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines (Apr. 25, 2023), https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413. For an explanation of the DSA's scope, see Martin Husovec, *The DSA's Scope Briefly Explained* (2023), https://ssrn.com/abstract=4365029. At the moment two platforms, Zalando and Amazon, are seeking invalidation of their designations before the General Court. Case T-367/23, Amazon v. European Commission (July 5, 2023); Case T-348/23, Zalando v. European Commission (June 27, 2023).

Francis Haugen's Facebook revelations[79] that put the spotlight on how amplification encourages certain types of behavior. Second, constant information and resource asymmetry between authorities (regulators) and providers realistically dictate that providers have the primary responsibility to find new solutions. The DSA's obligations relate more to the process or systems put in place. However, as shown below, this is more easily stated than practiced. In recent years, some type of "systemic regulatory approach" has been advocated by many scholars;[80] however, the details of such proposals differ significantly.

A particularly influential concept was Lorna Wood's and William Perrin's proposal which inspired the United Kingdom's Online Safety Bill (OSB). The proposal argued for the safety "by design" approach described as follows:[81]

> The regulator should be given substantial freedom in its approach to remain relevant and flexible over time. *We suggest the regulator employ a harm reduction method similar to that used for reducing pollution: agree tests for harm, run the tests, the company responsible for harm invests to reduce the tested level, test again to see if investment has worked and repeat if necessary* . . . . The regulator would then work with the largest companies to ensure that they had measured harm effectively and published harm reduction strategies addressing the risks of harm identified and mitigating risks that have materialised.

The framing of their model, including its placement under the statutory "duty of care" umbrella,[82] requires redistribution of responsibility for individual harms. This in turn evokes supervision of recommendation systems and product design features that change user behavior, including what individual content is being posted by them. While Wood and Perrin insist that content regulation is not the result of their approach,[83] they also envisage regulators'

79. Statement of Frances Haugen: Hearing before the S. Sub-Comm. on Consumer Protection, 177th Cong. (2021) (statement of Frances Haugen, former Facebook employee and whistleblower), https://www.commerce.senate.gov/services/files/FC8A558E-824E-4914-BEDB-3A7B1190BD49.

80. For an overview of (mostly) U.S. scholarship, see Kate Klonick's response to Douek, *supra* note 72. Kate Klonick, *Of Systems Thinking and Straw Men*, 136 HARV. L. REV. 339, 347 (2023).

81. LORNA WOODS & WILLIAM PERRIN, ONLINE HARM REDUCTION—A STATUTORY DUTY OF CARE AND REGULATOR 7, 13 (2019) (emphasis added), https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf.

82. *Id.* at 29–30.

83. *Id.* at 12 (discussing types of content).

ability to limit "harmful behavior"[84] prophylactically.[85] It thus hardly avoids addressing the substance of the environment—the underlying rules of engagement for users.

In contrast, Evelyn Douek's proposal equally centers around systems but in a very different way:[86]

> Instead of focusing on the downstream outcomes in individual cases, it focuses on the upstream choices about design and prioritization in content moderation that set the boundaries within which downstream paradigm cases can occur .... And in *focusing on procedural accountability rather than the pursuit of some substantive conception of an ideal speech environment*, it is more politically feasible and less constitutionally vulnerable.

Thus, her "substance-agnostic approach"[87] is much more limited because it allows companies to experiment with any (legal) content policies. However, it seems to be focused on regulation of amplification, which, as explained by Keller, is not always easily substance-agnostic either.[88]

In any risk management system, the relationship between substance and process is the most difficult one. First, any proposal that tries to tackle "risks" or overall "harmful behavior" cannot ignore that on user-generated content services, what users say or do remains a key risk factor. While user behavior can be encouraged by the design of services, in some form it will continue to exist irrespective of this encouragement; usually, the risks on such services result less frequently from purely non-human external factors.

Managing the risks of crowds *often* requires telling individuals how they must behave. If authorities subject the occurrence of selected *illegal* user expressions to some metrics, the legitimacy of such policy is straightforward. The legislatures already agreed that such behavior is illegal, and the authority is only trying to enforce compliance. However, if authorities subject the occurrence of some *legal* expressions to the same metrics, they can easily end up policing the bounds of what people can say—the content of their communications. Putting direct quotas on user expression, when taken to its logical conclusion, means telling some people what they cannot say.

---

84. *Id.* at 48. *See* Graham Smith, *Speech Is Not a Tripping Hazard—Response to the Online Harms White Paper*, CYBERLEAGLE (June 28, 2019), https://www.cyberleagle.com/2019/06/speech-is-not-tripping-hazard-response.html (discussing consequences).

85. *See generally* Woods & Perrin, *supra* note 81.

86. Douek, *supra* note 72, at 585 (emphasis added).

87. *Id.* at 606.

88. *See* Daphne Keller, *Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard*, 1, J. FREE SPEECH L. 227 (2021).

Whether addressing "harm" or "risk," the key litmus test is *who* sets the boundaries for the content of communications. One approach gives such power to decide to authorities; others leave it to individuals, platforms and legislatures.

- The *full* risk management approach gives the broadest power to authorities to ask companies about how their service design influences what happens on the platforms. Authorities observe, compare, analyze, and ask for changes, including by imposing tailored standards or quotas of "problematic" user behavior, regardless of the behavior's legality. The mandate of authorities thus extends to lawful but awful content and permits them to become surrogate legislatures policing the boundaries of free expression.
- The *limited* risk management approach shares the concerns about system design that might encourage various risks but stops before giving the authorities (agencies) the power to rewrite what lawful individual behavior should be banned or suppressed by quotas. This approach recognizes that authorities do not have the legitimacy of parliaments. Parliaments should remain responsible for setting the goalposts of illegal content of communications. If a specific risk or harm is particularly damaging, parliaments can move the goalposts further.[89] As a result, the authorities limit their demands regarding legal content to solutions that preserve people's agency by giving them freedom of choice; such solutions mostly empower or re-design the users' choice architecture.

Arguably, the Digital Services Act adopts the limited risk management approach. The DSA does *not* explicitly address the problem of whether the European Commission can require providers to change their contractual standards of "lawful but harmful content" as part of the risk management strategies.[90] However, in the absence of any explicit legal mandate, any attempts by the Commission to suppress specific legal expressions would arguably violate the rule of law.[91] The United Kingdom's Online Safety Bill is

89. In the UK, the self-harm debate led to the empowerment obligation and a proposal to create a new offence of "encouraging or assisting serious self-harm." Online Safety Bill 2022-3 Amendments, HL Bill [87] (later 362), p. 1 (2022), https://bills.parliament.uk/publications/51205/documents/3437.

90. Digital Services Act art. 35(1)(b), 2022 O.J. (L 277) speaks of "adapting their terms and conditions and their enforcement." In my view, this does not necessarily mean prohibiting lawful behavior or content. It speaks to the clarity and predictability of rules.

91. The argument is that Digital Services Act art. 34, 2022 O.J. (L 277) on its own is not sufficient to fulfil the human rights requirements under the E.U. Charter to legitimize prohibitions of speech to be "prescribed by the law." *See* Charter of Fundamental Rights of

currently moving in the DSA's direction too,[92] although the original proposal could have led to a full risk management approach.[93] The Australian Online Safety Act of 2021, however, seems to go the farthest by allowing authorities to ask for the removal of lawful but awful content.[94]

The limited risk management approach can be best explained in an analogy with managing risks during public protests. Imagine a public assembly protesting immigration policies that gathers in the streets of a city. The role of providers can be analogized to the position of protest organizers.[95]

The DSA designates the largest services in the European Union as controlled public spaces; it tasks their designers—the providers—to analyze risks created by bringing crowds together and to intervene if needed. What is the role of the state and providers in such cases?

The state can impose safety measures on organizers and protesters to protect them from others and others from them; and to avoid hurting bodies, property, or businesses. Physical safety measures benefit freedom of expression because they make everyone more comfortable in expressing their views. To achieve this, the authorities can ask organizers to take various safety

---

the European Union art. 52, 2012 O.J. (C 326) 391. For an excellent article on the rule of law requirement in this context, see Graham Smith. *Online Harms and the Legality Principle*, CYBERLEAGLE (2020), https://www.cyberleagle.com/2020/06/online-harms-and-legality-principle.html ("[T]he regulator's views about harm would sit alongside, and effectively supplant, the existing, carefully crafted, set of laws governing the speech of individuals.").

92. Douek, *supra* note 72.

93. The UK government construed Wood and Perrin's proposal in its initial proposal of the Online Safety Bill by creating a controversial clause about safety duties for "harmful but lawful" content for adults. The relevant clause was dropped and the bill left with an empowerment obligation under the system known as "triple lock" in the later versions of the Bill. *See* Online Safety Bill 2022-3, HL Bill [362], § 12 (UK), https://bills.parliament.uk/bills/3137 (creating a duty for some services to "include in a service . . . features which adult users may use or apply if they wish to increase their control over content" that "reduce[s] the likelihood of the user encountering" or "alert" users to some types of content, such as hate speech, self-harm or eating disorders).

94. *See Online Safety Act 2021* (Cth), pt 4, 9 (Austl.), https://www.legislation.gov.au/Details/C2021A00076. According to Professor Nicolas Suzor, "The classification scheme has long been criticized because it captures a whole bunch of material that is perfectly legal to create, access and distribute." *See* Ariel Bogle, *Australia's Changing How it Regulates the Internet—and No-one's Paying Attention*, ABC NEWS (Sept. 20, 2022),https://www.abc.net.au/news/science/2022-09-21/internet-online-safety-act-industry-codes/101456902.

95. One of the reviewers made an excellent point, with which I nevertheless do not fully agree. The reviewer argues that a better analogy would be with the owner or operator of the property. In my view, this would evoke a very passive role of the platforms that do not influence the created risks by the design of their services. While the metaphor of organisers might better fit social media with active recommender systems than Wikipedia (also an online platform), my illustration is meant to show how self-imposed rules should not be adopted by authorities as the reason for intervention for otherwise legal protests.

measures particularly to prevent illegal behavior by protesters or counter-protesters, including proscribing the use of excessive disruption or noise. However, beyond illegal modes of expression, the *authorities* cannot control who speaks or protests, what posters or chants they use, or where they present them. That said, *organizers* can go beyond illegality, whatever their motivation. They can self-impose stricter rules on crowds.

Imagine now that this public assembly has two teams of rule enforcers dressed in red and blue jackets. Red enforcers represent the state, and they can only intervene when protesters violate a set of red rules—the behavior that the legislature has determined to be illegal. Blue enforcers are paid by organizers. They are the analogue of content moderators. Because organizers want a legitimate assembly where families can gather, they ask all the participants to respect some of their own basic rules. These blue rules differ from red rules. Among other things, they allow organizers much earlier intervention. For instance, they can say that posters with profanities are not permitted because they are likely to lead to illegal behavior.

For efficiency reasons, the state will expect blue enforcers to also enforce red rules. This is the analogue of delegated enforcement in which providers engage daily when they remove illegal content. However, red enforcers *cannot* enforce blue rules. Blue rules are the analogue of contractual self-restraint that platforms adopt to make their services appealing to users and advertisers. Red enforcers cannot turn a self-imposed ban on profanities against the organizers to end the protest or arrest protesters. To justify such intervention, authorities must stick to the red rules. Logically, they cannot tell organizers what blue rules to adopt either because that is the prerogative of legislatures. The state can, however, require that protesters inching closer to escalation must take extra measures to keep bystanders safe from violence.

The DSA's very large online platforms (VLOPs) and very large online search engines (VLOSEs) manage huge crowds constantly. As a result, they must periodically assess the risks, submit their reports to auditors, and follow up in case the auditors are not satisfied. The entire dossier of documents is then submitted to the European Commission for the ultimate assessment and release for the public to see and criticize.

IV. PRINCIPLES FOR A NEW GENERATION OF RULES

Now that the reader is familiar with the rules in the Digital Services Act, I would like to extract some of the main principles that define the regulatory approach. As pointed out by Daphne Keller, "differences between American and European approaches shouldn't prevent us from finding common ground

on other functional aspects of platform regulation."[96] The DSA has a lot to offer, but one needs to look beyond the exact wording and "under the hood" to understand the thinking. In my view, the following set of principles can be derived from the DSA and could serve as "common ground" to guide the legislative design of a new generation of rules:[97]

1. Accountability, not liability
2. Horizontality of regulations
3. Shared burden: everyone is responsible
4. Empowerment of users
5. Ecosystem solutions

A. ACCOUNTABILITY, NOT LIABILITY

Platforms as facilitators of user-generated content cannot be expected to bear the liability burden of conventional publishers, such as newspapers. As much as their content moderation might resemble quasi-editorial functions, the special features of the internet demand different legal regimes. The existence of some sensible legal immunities for liability generated by the actions of others is the basic precondition of the viability of the user-generated services which harness the internet's special benefits. These include no requirements for editorial approval, low barriers of entry, incredible speed and scale of distribution, broad social and geographical inclusiveness, and resilience of communications. Instead of devising restrictions which may negate these advantages, the focus should be on how to align providers' business operations with socially optimal practices that maximize freedoms of individuals—thus making the businesses more accountable to public interest.

Prior to the DSA, most of the relevant laws tried to influence providers' behavior by threatening them with accessory liability for what their users do.[98] Save for some areas of law, most notably intellectual property law in the European Union,[99] the courts often faced a binary decision: impose liability, with all its consequences, or deny it entirely and confirm a liability exemption. The DSA ends this binary. Self-standing regulatory expectations created by the

96. Daphne Keller, *For Platform Regulation Congress Should Use a European Cheat Sheet*, HILL (Jan. 15, 2021), https://thehill.com/opinion/technology/534411-for-platform-regulation-congress-should-use-a-european-cheat-sheet/.

97. *See* MARTIN HUSOVEC, PRINCIPLES OF THE DIGITAL SERVICES ACT (Oxford Univ. Press forthcoming 2024).

98. *See* Martin Husovec, *Remedies First, Liability Second: Or Why We Fail to Agree on Optimal Design of Intermediary Liability?*, *in* THE OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY (Oxford Univ. Press 2020) (criticizing a one-size fits all approach).

99. *See* MARTIN HUSOVEC, INJUNCTIONS AGAINST INTERMEDIARIES IN THE EUROPEAN UNION: ACCOUNTABLE BUT NOT LIABLE? (Cambridge Univ. Press 2017).

legislature give courts and authorities a third option. A failure to satisfy such expectations is enforced separately. Thus, similar to banks that are usually not liable for the illegal financial transactions of their clients, they can still be held accountable and fined for not adopting the right anti-money laundering processes.

The DSA leaves existing liability exemptions almost intact. Section 4 of the ECD is incorporated into Chapter 2 of the DSA. Its novelty is in the creation of new regulatory expectations named "due diligence obligations" that are foreseen in Chapter 3. They are unrelated to legal immunities for third-party content. As noted by Recital 41 of the DSA, "[t]he due diligence obligations are independent from the question of liability of providers of intermediary services which need therefore to be assessed separately." If due diligence obligations are violated, they trigger a separate enforcement system envisaged by the DSA; they do not expose providers to a flood of claims for individual grievances. Due diligence obligations aim to improve the operations of systems and procedures that companies are using to moderate users' content or manage other overall risks.

To illustrate this, consider the following example. In American copyright law, under § 512(i) of the DMCA, a failure to terminate accounts of repeat infringers leads to the loss of a liability exemption and thus the potential joint liability of providers for the actions of users who infringe copyright when using their services. In European copyright law, such failure has no impact on liability exemptions. However, post-DSA, a failure to terminate accounts of repeat infringers can lead to a violation of Article 23(1) of the DSA, which can be enforced privately or publicly even though the liability exemption continues to apply. Thus, in both cases, the consequences are substantially different.

The DSA's accountability-but-not-liability design was not an automatic policy choice. In the legislative process, the European Parliament strongly pushed to make the liability exemptions dependent on compliance with due diligence obligations. Thus, any violation of Chapter III of the DSA would make liability exemptions unavailable. The opposite approach, where due diligence duties act as preconditions, exists under Section 79 of the Indian Information Technology Act (2000),[100] and is being proposed by Professor

100. Section 79 of the Indian Information Technology Act states that "the intermediary observ[ing] due diligence while discharging his duties under this Act and also observ[ing] such other guidelines as the Central Government may prescribe in this behalf." The Information Technology Act, 2000, § 79. Part II(4)(4) of the Indian Ministry Guidance, https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf (imposing filters on "significant social media intermediaries").

Danielle Citron as a solution for the revision of § 230 of the CDA in the United States.[101] Under such a system, the liability exemptions would become a truly hard-earned "prize" or a "privilege" given only to those who respected the DSA in its entirety. The more due diligence obligations are added to the list, the more impossible walking of the tightrope becomes. The E.U. legislature consciously decided against this approach—for good reasons, as it would basically nullify the existence of liability exemptions.

In the liability framework, the lack of diligence puts providers at risk of being an accessory to the entire wrongs of others. On the other hand, the accountability framework blames them only for not giving some specific assistance.[102] The legal culpability implied in the two settings is very different and it translates into the seriousness of the consequences for the platforms. While liable platforms face injunctions and joint liability for damages and are called to account by many victims who were wronged by the actions of others, accountable platforms only face the pain of enforcement efforts to bring them into compliance. Thus, while liable platforms restore a lawful state by making the victims whole, accountable platforms restore it by simply adjusting their behavior in ways that comply with regulatory expectations.[103]

If accountability is further narrowed down to *systemic* legal obligations in the design and operation of systems and processes, the difference is even more significant.[104] Under such systems, if a provider violates a systemic due diligence obligation, only one obligation to correct the outcome is owed to individuals or regulators. In contrast, if such obligation is embedded into a liability exemption, one failure to operate a specific policy leads to separate

101. Danielle Keats Citron, *How To Fix Section 230*, VA. PUB. L. & LEGAL THEORY RSCH. PAPER NO. 2022-18 (2022), https://ssrn.com/abstract=4054906 (arguing that § 230 should be narrowed in scope, and made subject to duties of care that can be further fleshed out by administrative agencies).

102. This should hold true for both public and private enforcement. Even for damages claims for violations of due diligence (Digital Services Act art. 54, 2022 O.J. (L 277)), the damage must be causally connected with the violation of the diligence obligation (Digital Services Act recital 122, 2022 O.J. (L 277) and only compensate the corresponding part of the damage. Thus, damages caused by third parties who uploaded the content are distinct.

103. Arguably, there are situations where liability exemptions will be lost, due diligence obligations violated, and the damage caused by a third party is closely related to that caused by violation of a due diligence obligation. In such cases, the DSA can indicate to national law that a component of the duty of care for domestic liability rules was violated. However, in many cases, the two harms are unlikely to be related (e.g., transparency rules or non-arbitrariness standards hardly relate to damage caused by third-party content).

104. The primary example of such obligations is Digital Services Act art. 21(2), 2022 O.J. (L 277) ("[E]ngage, in good faith, with the selected certified out-of-court dispute settlement body with a view to resolving the dispute"). Other examples are Article 22(2)(c), Article 23, and arguably many open-ended standards of Article 20(4).

debts to many who were wronged. Accountability for systemic obligations means owing one type of assistance to all affected people, while liability for others means owing all wronged people full liability for the actions of many other people. The difference is stark. Moreover, the DSA prohibits super-compensatory damages for due diligence violations (Article 54).

B. HORIZONTALITY OF REGULATIONS

The second principle implicit in the DSA's and ECD's design is its horizontal character. The horizontal approach cuts through the entire legal system and thus creates baseline expectations. Sectorial rules remain possible; however, they are forced to interact with the horizontal approach. In the European Union, the DSA thus becomes a *digital civil charter* that shines through the entire legal system and radiates minimum rights of individuals. Unless the European legislature suspends it in various areas, it creates a baseline that holds across the entire ecosystem of user-generated content services. In the DSA, the horizontality of liability exceptions is complemented by the horizontality of due diligence obligations. This supports my earlier argument about the updated digital social contract for user-generated content services. For instance, in the European system, the DSA's rules substantially improved the situation under several sectorial rules dealing with copyright issues.[105]

The horizontality of rules is not only useful for complying companies and individuals, but it also prevents gaming the system. The typical problem with sectorial rules imposing different standards is that they invite regulatory arbitrage. For instance, in the US, where Section 230 of the CDA provides even post-notification immunity to hosting services, there is a strong incentive to formulate any claims as copyright issues because § 512 of the DMCA is much more accommodating.[106] This turns defamation claims into copyright claims and distorts copyright policy in the long run. In a situation where the identity of claims is fluid and the plaintiffs can shop around for the strongest

105. The DSA updated the safeguards applicable under Article 17 of the Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *See* Martin Husovec, *Mandatory Filtering Does Not Always Violate Freedom of Expression: Important Lessons from Poland V. Council and European Parliament*, 60 COMMON MKT. L. REV. 173 (2023); João Pedro Quintais & Sebastian Felix Schwemer, *The Interplay between the Digital Services Act and Sector Regulation: How Special is Copyright?*, 13 EU. J. OF RISK REGUL. 191 (2022).

106. The most famous of the abuses of this kind are U.S. doctors asking for copyright assignment to text of future reviews to be able to require their takedown. *See* Mike Masnick, *Why Doctors Shouldn't Abuse Copyright Law to Stop Patient Reviews*, TECH DIRT (Apr. 14, 2011), https://www.techdirt.com/2011/04/14/why-doctors-shouldnt-abuse-copyright-law-to-stop-patient-reviews/.

cause of action, diverging standards for different legal areas are bound to cause regulatory arbitrage. The only way to avoid this is to adopt one set of uniform rules for all areas of law.

Horizontality also allows for better balancing of different trade-offs. For example, protecting minors might come at the expense of the freedoms of adults. Enforcement of hate speech policies can have unintended effects on legitimate discourse. Having a holistic policy allows the regulators to better balance one against the other, as their mandates extend to both. Thus, the European Commission, when looking at risks and technological solutions, must equally consider the under-detection of hate speech and over-blocking of legitimate speech. Given that content moderation and risk management stretch into all areas of human interactions, having the broadest possible focus is key to any balanced policy.

Politically, the horizontal approach also moderates the excessive strength of some interest groups because it broadens the conversation and dilutes their voice with the equally valid concerns of others. The E-Commerce Directive and the Digital Services Act could hardly have been adopted as sectorial measures. In fact, both the American and European examples show that copyright rules, an area that powerful lobbies of interest groups exercise influence over, constantly diverge from the baseline in favor of copyright holders. Section 512 of the DMCA is stricter than § 230 of the CDA. Similarly, Article 17 of the Copyright DSM Directive is stricter than Article 6 of the DSA.[107]

C. SHARED BURDEN: EVERYONE IS RESPONSIBLE

The DSA renews democratic support for the shared burden model for societal risks on digital services. Under the principle of shared burden, everyone is expected to play their part—to do something to protect oneself. It also means resisting the temptation to blame one actor for all ills.

In liability systems around the world,[108] it is an established principle that if victims contribute to their own damage by failing to exercise due care, the

107. Article 17 of the CDSM Directive introduces a system of strict liability for unlicensed content unless case providers can meet very strict cumulative conditions: inability to obtain a license, the stay-down obligation, and notice-and-takedown system. *See* Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC Council Directive 2019/790, art. 17, 2019 O.J. (L 130) 92.

108. *See Principles of the European Tort Law (PETL), Art. 8.1.1*, http://www.egtl.org/PETLEnglish.html ("Liability can be excluded or reduced to such extent as is considered just having regard to the victim's contributory fault and to any other matters which would be relevant to establish or reduce liability of the victim if he were the tortfeasor."); Martin Turck,

person who is otherwise liable will face decreased or no liability. This principle of comparative negligence was famously formulated by Lord Ellenborough in 1809 who said that: "One person being in fault will not dispense with another's using ordinary care for himself."[109] Arguably, the ECD builds upon this principle in the design of its liability exemptions, and the DSA designs its due diligence obligations the same way.

Under liability exemptions, victims or their representatives must notify providers about infringing content or seek redress before authorities, and providers must act upon notifications or state-issued orders. Providers are usually not expected to prevent all individual grievances; instead, hosting providers must investigate them mostly once they are brought to their attention. Even the ex ante risk management due diligence obligations do not change that. Reporting illegal content, disputing providers' decisions, organizing with others, and learning and teaching others how to avoid risks remain the key ingredients of the DSA's content moderation system.

One of the expressions of the shared burden principle is also the prohibition of general monitoring in the ECD and DSA. Article 15 of the ECD, now Article 8 of the DSA, prohibits the following:

> No general obligation to monitor the information which providers of intermediary services transmit or store, nor actively to seek facts or circumstances indicating illegal activity shall be imposed on those providers.

The provision thus also embodies the idea[110] that the law generally structures the allocation of responsibilities to various actors. Providers are not subject to general obligations to intervene in other people's affairs. Such implicit allocation is not exhausted by the liability exemptions. This is also why any other rules imposed on providers, such as injunctions or any permitted national regulatory expectations, remain curtailed. As much as the burden under the liability exceptions system is shared, so must the burden under the accountability for risk management system be similarly split.

The sharing of the burden under liability exemptions allowed the user-generated content universe to flourish because it spreads responsibility and

Contribution Between Tortfeasors in American and German Law--A Comparative Study, 41 TUL. L. REV. 1 (1966–1967); Giuseppe Dari-Mattiacci & Eva S. Hendriks, *Relative Fault and Efficient Negligence: Comparative Negligence Explained*, 9 REV. OF LAW & ECON. 1 (2013).

109. Butterfield v. Forrester, Eng. Rep. 926, 927 (1809).

110. Advocate General Øe in his Opinion in C-401/19, ¶ 106 ("I am inclined to regard the prohibition laid down in Article 15 of Directive 2000/31 as a general principle of law governing the internet, in that it gives practical effect, in the digital environment, to the fundamental freedom of communication.").

thus expectations. Burden sharing under the accountability-for-risk-management framework will be equally crucial to avoid moral hazard.[111] While the DSA clearly puts accountability for risks on VLOPs/VLOSEs, it does not require the eradication of risks. Not all risks can be controlled by providers in the same way. While inherent risks cannot be mitigated at all, other risks can be increased by the behavior of providers, their users, or third parties.

For instance, the risk of fraud via digital scams depends not only on platforms' protective systems but also on their users' behavior, skills, and awareness. Providers can do a lot to prevent such scams; however, they can only partly influence users' behavior, skills, and awareness. The risk thus needs to be distributed, and users must share their part of the burden. This is how we deal with risks in most areas because protecting people against their own irresponsibility sometimes only breeds more irresponsible behavior.[112] The same starting point should be used to approach the regulation of issues such as the manipulation of votes by disinformation campaigns. The VLOPs' and VLOSEs' accountability for these harms is significant, but not absolute and not exclusive.

This brings me to my next principle.

D. USER EMPOWERMENT

The users can only be asked to learn how to share part of the risks if they are able, and thus empowered, to mitigate them. The principle of user empowerment means that ultimately, users can share only parts of those risks that they are given a chance to control. Typically, this means the provision of tools that grant people agency in deciding what they wish to see and from whom. If platforms leave little agency to users, they should assume more risks. The more agency users gain, the more they can control their own digital experience. Thus, undeniably, more user empowerment means less central responsibility of providers, which might not appeal to everyone. But it does not mean that such tools will allow providers to shrug off any accountability for risks; if coupled with reasonable expectations on the users' side, control given to users can at best reduce it.

The DSA tries to give users new levers of control over their user experience, such as the ability to challenge decisions, receive compensation for

111. *See generally* John M. Marshall, *Moral Hazard*, 66 AM. ECON. REV. 880 (1976).

112. For instance, in the EU, liability for unauthorized payments, such as those caused by phishing attacks, is primarily with banks. However, if clients behave grossly negligently, the banks do not have to compensate the clients. *See* Article 73 of the Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on Payment Services in the Internal Market, Amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 2007/64/EC.

moderation mistakes, rely on representation before platforms, benefit from new parental tools and choice on recommender systems. As explained by Recital 40 of the DSA, the due diligence obligations:[113]

> should aim in particular to guarantee different public policy objectives such as the safety and trust of the recipients of the service, . . . the protection of relevant fundamental rights enshrined in the Charter, the meaningful accountability of those providers and the *empowerment* of recipients and other affected parties, whilst facilitating the necessary oversight by competent authorities.

Thus, empowerment of individuals is encoded in the DSA and invites providers to harness its power. The trade-off for VLOPs/VLOSEs is clear. Relinquish part of control in exchange for lesser accountability for risks or keep full control and assume more responsibility for what transpires on the platform. Risk-sharing is thus an incentive to delegate to users and enhance their agency as individuals with free will and preferences.

When I am talking about empowerment tools, I do not mean the obvious tools. Realistically, all platforms give users some agency in their digital experience. We all want to follow people based on our preferences and block people who cross our personal red lines.[114] However, platforms still assume too much central control over many decisions where the personal preferences of their users can legitimately diverge. By definition, this is most important for the category of legal content that can be controversial to host. While few users will diverge on their preferences for commercial spam, many might have different sensitivities for shocking, sensational, nude, or vulgar content.

In the literature, Fukuyama and others have argued for empowerment through a system of middle-ware tools that could help users to personalize their content moderation experience.[115] The idea of polycentric content moderation that puts users in charge of more decisions arguably already exists, however, before the DSA could not have been legally compelled. Consider a new start-up, TrollWall,[116] that offers social media page administrators a machine learning-based content moderation tool that is meant to address the slow removal of illegal content by Facebook, but also offers a scalable solution

113. Digital Services Act recital 40, 2022 O.J. (L 277) (emphasis added).

114. Naturally, any preference for illegal content is simply illegal and thus irrelevant.

115. FRANCIS FUKUYAMA, BARAK RICHMAN, ASHISH GOEL, ROBERTA R. KATZ, A. DOUGLAS MELAMED & MARIETJE SCHAAKE, REPORT OF THE WORKING GROUP ON PLATFORM SCALE, https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/platform_scale_whitepaper_-cpc-pacs.pdf.

116. *See AI Autopilot for Comment Moderation*, TROLL WALL, https://www.trollwall.ai/ (last accessed Sept. 26, 2023).

to preserve the civility of online discussions. This tool gives administrators the ability to adjust content categories, sensitivity, and what should happen with the detected content. Although the tool is offered by a third party to page administrators,[117] Facebook has a key role in creating APIs that facilitate it and approves such apps for distribution in its platform. While far from being error-free, the tool gives administrators more agency to deal with problems with a scale that is prohibitively big for full human oversight. The DSA can pave the way to more of such tools that puts users and other individuals in charge.

E. ECOSYSTEM SOLUTIONS

If the responsibility for societal challenges is shared, everyone needs to be part of the solution. While providers and the state navigate their respective roles, civil society holds both to account.

Countering extremism or disinformation can be successful only if providers are assisted by an ecosystem of actors, such as trusted NGOs who notify the content, fact-checkers, journalists, or researchers. One of the shortcomings of the first generation of rules like the DMCA, CDA, and ECD is their preoccupation with providers and the little consideration they pay to those other players in the ecosystem.[118] Under the E-Commerce Directive, only platforms were relieved of liability. The other parties involved in solving the societal challenges in play were not given any specific tools to do their work. The self-regulatory approach was meant to solve this in the European Union. However, this often led to disparate arrangements across different services that can be taken away from civil society at the whim of new owners or leadership of providers.[119] For civil society, disparities mean difficulties in scaling the response.

The Digital Services Act puts the ecosystem front and center. It recognizes that content moderation is a product of decision-making by providers, but its quality is equally dependent on inputs (the quality of notifications) and feedback (the ability of users to correct the mistakes).

---

117. Naturally, Facebook can offer its own tools to page administrators, but these have so far very limited usefulness, especially in smaller markets. Thus, one can see how user empowerment can play out in small.

118. Jessica Litman has argued that the DMCA "sells the public short." And yet, § 512 DMCA at least includes some safeguards—even if ineffective in practice—such as details for notices (§ 512(c)(3)) and rules on counter-notice (§ 512(j)) or misrepresentation (§ 512(f)). *See* JESSICA D. LITMAN, DIGITAL COPYRIGHT 145 (Prometheus Books, 2d ed. 2006).

119. Brian Fung, *Academic Researchers Blast Twitter's Data Paywall as 'Outrageously Expensive,'* CNN (Apr. 5, 2023), https://edition.cnn.com/2023/04/05/tech/academic-researchers-blast-twitter-paywall/index.html.

On the side of inputs, the DSA tries to incentivize the quality of notifications. Providers are tasked with designing their submission interfaces in user-friendly ways to help other actors with their work.[120] It gives preferential treatment to trusted flaggers who have a track record of quality.[121] Trusted flaggers that abuse their position might be suspended or have their certification removed by regulators.[122] Providers are asked to suspend or terminate the accounts of those who repeatedly submit abusive notifications or manifestly illegal content.[123] The DSA encourages standardization[124] of how notices are exchanged which should lead to the emergence of more automated cross-platform solutions.

On the side of feedback, the DSA tries to decrease the information asymmetry between providers and their content creators. Providers must properly disclose their rules up front and describe what automated tools they use to enforce them.[125] They must issue individualized explanations for a wide range of content moderation decisions and allow appeals free of charge.[126] If content creators or notifiers are dissatisfied, they can file external appeals to out-of-court dispute resolution bodies.[127] The providers must pay for the complainant's costs of initiating external appeals whenever they lose cases, which should motivate them to improve the quality of their decisions internally.[128] Specialized organizations can be included in the dispute resolution process, thus allowing content creators to improve the quality of their representation.[129] Consumer groups are given a collective redress in the form of injunctions which can be sought to cure noncompliance.[130]

In the risk management pillar, the DSA asks researchers, civil society, and auditors to formulate relevant risks and invent new ways to mitigate them. For the largest digital services, regulators conduct a regulatory dialogue about societal challenges in public to intensify scrutiny.

120. Digital Services Act art. 16(1), 2022 O.J. (L 277),
121. *Id.* art. 22.
122. *Id.*
123. *Id.* art. 23.
124. *Id.* art. 44(1).
125. *Id.* arts. 14–15.
126. *Id.* arts. 17, 20.
127. *Id.* art. 21.
128. For an empirical test of this proposition, see *supra* note 74. Given that the system offers a more credible remedy, one can also expect that the use of it will increase, thus the impact will be higher than under the current system, where no independent third party is involved, and the only available remedy—courts—are not as de-risked for the complainants.
129. Digital Services Act art. 86, 2022 O.J. (L 277).
130. *Id.* art. 90.

In other words, the DSA gives other actors in the digital ecosystem tools that they can rely on when protecting private or public interests. By doing this, the DSA heavily relies on societal structures that the law can naturally only foresee and incentivize but cannot build. These structures—such as local organizations analyzing threats, consumer groups helping content creators, and communities of researchers—are the ones that give life to the DSA's tools. They need to be built from the bottom up by people, perhaps even locally in each member state. If their creation fails, the regulatory promises might turn out to be nothing more than glorious aspirations.

V. CONCLUSIONS

In 2023, content moderation continues to be a politically divisive topic in the United States. The Republican Party wants companies to moderate less content that is not prohibited by the legislature.[131] The Democratic Party wants them to moderate more of such content. The political currents have not yet swept Europe in a similar way, although the political situation is evolving.[132] While the two sides cannot agree on how to exercise content moderation discretion, they should be able to agree that legislative acts reinstating ex ante editors are in no one's interest.

The internet is a special medium that should not be regulated as broadcasting or newspapers. Content moderation discretion can only exist if providers have very limited liability for the distribution of the content of others. If liability is strict or close to strict, their discretion must morph into editorial discretion because no one can offer digital spaces or tools for expression without vetting information in advance.

Running our digital services—ranging from social media and marketplaces to search engines—on the infrastructure of editorial control is impossible. Thus, what policymakers should aim for is to increase providers' accountability while keeping their liability limited. Platforms need more *accountability, not liability*. Their design practices should be subject to regulation without immediately expanding the underlying content laws.

131. *See* S.B. 7072, 2021 Leg. (Fla.); H.B. 20, 2021 Leg., 87th Sess. (Tex.).

132. Among the E.U. countries, only Polish conservatives introduced a bill similar to the United States' Florida and Texas proposals. The Polish bill was meant to protect against "censorship" by prohibiting moderation of legal content. *Law To Protect Poles From Social Media "Censorship" Added To Government Agenda*, NOTES FROM POLAND (Oct. 5, 2021), https://notesfrompoland.com/2021/10/05/law-to-protect-poles-from-social-media-censorship-added-to-government-agenda/. However, the bill was never adopted. In the UK Online Safety Bill, the controversy around "lawful but harmful content" for adults led a new prime minister, Rishi Sunak, to drop the clause and only rely on empowerment obligation and extension of some offences.

Because non-editorial content lacks editors, some think it will also always lack *trust*. This leads policymakers to push for tighter content standards or even editorial discretion. However, there are ways to inject trust into the ecosystem without abandoning its decentralized character. The solution of the Trusted Content Creators,[133] for instance, draws entirely on the principles of *shared burden* and *ecosystem solutions*. Instead of banning or suppressing that what is not trusted, TCC rewards trusted content by asking providers to give extra benefits to those content creators who self-organize and commit to abide by their own shared norms. Decentralization is not the antithesis of trust.

Similarly, there are many ways to overcome different views on *how* to exercise content moderation discretion over legal content. The *user-empowerment* principle shows the way for a middle ground between two positions on how to exercise content moderation. It invites policymakers to think about solutions that delegate the choice of what legal content to display from advertisers or providers to individuals. The legislature can also facilitate user choice by making the underlying markets more competitive[134] or open up the content moderation experiences within dominant services to more alternatives.[135]

People voting with their feet show that they are interested in non-editorial content much more than they are in editorial content. Among the top fifty visited websites on the internet globally, the great majority rely on users—other people—to generate the content.[136] It seems that humans are primarily interested in what other humans have to say. No one can beat the educating and entertaining power of crowds. While we often fret about issues of the legality and trustworthiness of such content, only a few think the solution is to go back to the age of editorial media.

The proposed five principles offer common ground for liberal democracies to think about the challenges of our day without sacrificing what we have gained: an inclusive, decentralized and open global communication network.

133. Martin Husovec, *Trusted Content Creators*, LSE LAW POLICY BRIEFING PAPER NO. 52, (2022), https://ssrn.com/abstract=4290917.

134. This is the approach taken by the Digital Markets Act. *See* Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828, 2022 O.J. (L 265).

135. *See* Fukuyama et al., *supra* note 115 (proposing middle-ware).

136. *See List of Most-visited Websites*, WIKIPEDIA, https://en.wikipedia.org/wiki/List_of_most_visited_websites (last accessed Sept. 26, 2023).

THREE SIZES FIT SOME: WHY CONTENT REGULATION NEEDS TEST SUITES

Rebecca Tushnet†

ABSTRACT

The European Union's Digital Services Act (DSA) offers a new model for regulating online services that allow users to post things. It uses size-based tiers to delineate the different levels of obligation imposed on various services. Despite the tiers of regulation in the DSA, and very much in its copyright-specific companion Article 17, it's evident that the broad contours of the new rules were written with insufficient attention to variation. Instead, regulators assumed that "the internet" largely behaved like YouTube and Facebook. Using three examples of how that model is likely to be bad for a thriving online ecosystem—counting users, providing due process, and implementing copyright-specific rules—this Article concludes that, to improve policymaking, regulators should use test suites of differently situated services to ensure that they are at least considering existing diversity and properly identifying their targets.

TABLE OF CONTENTS

I. INTRODUCTION

The European Union's Digital Services Act (DSA) offers a new model for regulating online services that allows users to post content online, as does its copyright-specific companion Article 17. Both sets of rules attempt to use tiers to distinguish among types of services. In general, smaller or otherwise less-commercial endeavors have fewer obligations. Despite these gestures towards customization, the broad contours of the new rules were written with

DOI: https://doi.org/10.15779/Z386688K73

† Frank Stanton Professor of First Amendment Law, Harvard Law School.

insufficient attention to variation, in part because the regulators were thinking about YouTube and Facebook as shorthand for "the internet" in full. This brief Article will discuss three examples of how that totalizing model is likely to damage a thriving online ecosystem. Problems of service variation—even among platforms that host substantial amounts of user-generated content—arise in counting users, providing due process, and implementing copyright-specific rules. The crude tiers in the system risk creating the situation they presume: an internet with substantially less variation. And this is unlikely to be good for creators, consumers, or anyone else.

This Article concludes by recommending the use of test suites in which regulators ask whether a variety of differently situated services have the features about which the regulations are concerned. This will increase the chances that regulators at least consider the existing diversity of internet services and increase the chances that they properly identify their targets.

II. COUNTING USERS

The first issue is the smallest but reveals the underlying complexity of the problems of regulation at the very outset of the regulatory process. As Martin Husovec wrote,[1] placement in the tiers depends on monthly active users of the service, which explicitly extends beyond registered users to recipients who have "engaged" with an online platform "by either requesting the online platform to host information or being exposed to information hosted by the online platform and disseminated through its online interface."[2] While a recital clarifies that multi-device use by the same person should not count as multiple users,[3] that leaves many other measurement questions unsettled, and Husovec concludes that "[t]he use of proxies (e.g., the average number of devices per person) to calculate the final number of unique users is thus unavoidable. Whatever the final number, it always remains to be only a better or worse approximation of the real user base."[4] And yet, as he writes, "Article 24(2) demands a number."[5] This obligation applies to every service because it determines which tier, including the small and micro enterprise tier, a service falls into.

1. Martin Husovec, *The DSA's Scope Briefly Explained* (July 4, 2023) https://ssrn.com/sol3/papers.cfm?abstract_id=4365029, at 1–2.

2. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), O.J. (L 277) 1 EU, art. 3(p).

3. *Id.* Recital 77.

4. Husovec, *supra* note 1, at 4.

5. *Id.*

This demand is based on assumptions that are simply not uniformly true about how online services monitor their users, especially in the nonprofit or public interest sector. It seems evident—though not specified by the law—that a polity that passed the European Union's General Data Protection Regulation (GDPR) would not want services to engage in tracking just to comply with the requirement to generate a number. As the search engine provider DuckDuckGo pointed out, by design, its search engine doesn't track users, create unique cookies, or have the ability to create a search or browsing history for any individual.[6] So, to approximate compliance, it used survey data to generate the average number of searches conducted by users—despite basic underlying uncertainties about whether surveys could ever be representative of a service of this type—and applied it to an estimate of the total number of searches conducted from the European Union.[7] This doesn't seem like a bad guess, but it's a pretty significant amount of guessing.

Likewise, Wikipedia assumed that the average E.U. visitor used more than one device, but estimated devices per person based on global values for 2018, rather than for 2023 or for Europe specifically.[8] Perhaps one reason Wikipedia overestimated was because it was obviously going to be regulated no matter what, so the benefits of reporting big numbers outweighed the costs of doing so, as well as the stated reason that there was "uncertainty regarding the impact of Internet-connected devices that cannot be used with our projects (e.g., some IoT devices), or device sharing (e.g., within households or libraries)."[9] But it reserved the right to use different, less conservative assumptions in the future. In addition, Wikipedia noted uncertainty about what qualified as a "service" or "platform" with respect to what Wikipedia's specific, and somewhat unusual, organization[10]—is English Wikipedia a different service or platform for DSA purposes than Spanish Wikipedia? That question obviously has profound implications for some services. And Wikipedia likewise reserved the right to argue that the services should be treated separately,[11] though it's still not clear whether that would make a difference if none of Wikipedia's projects qualify as micro or small enterprises.

6. *Digital Services Act (EU 2022/2065) Regulatory Reporting*, DUCKDUCKGO, https://help.duckduckgo.com/duckduckgo-help-pages/r-legal/regulatory-reporting/ (last visited May 15, 2023).

7. *Id.*

8. *EU DSA Userbase Statistics*, WIKIMEDIA FOUND., https://foundation.wikimedia.org/wiki/Legal:EU_DSA_Userbase_Statistics (last visited May 15, 2023).

9. *Id.*

10. *Id.*

11. *Id.*

The nonprofit I work with, the Organization for Transformative Works (OTW), was established in 2007 to protect and defend fans and fanworks from commercial exploitation and legal challenge. OTW members make and share works commenting on and transforming existing works, adding new meaning and insights—from reworking a film from the perspective of the "villain," to using storytelling to explore racial dynamics in media, or to retelling the story as if a woman, instead of a man, were the hero. The OTW's nonprofit, volunteer-operated website hosting transformative, noncommercial works, the Archive of Our Own (AO3), as of early 2023 had over 4.7 million registered users, hosted over 11 million unique works,[12] and received approximately two billion page views per month—on a budget of under $300,000 a year.[13] Like DuckDuckGo, the OTW doesn't collect anything like the kind of information that the DSA assumes online services have at hand, even for registered users (which, again, are not the appropriate group for counting users for the DSA's purposes).

The DSA is written with the assumption that platforms extensively track its users. If that isn't true, because a service isn't trying to monetize them or incentivize them to stay on the site, it's not clear what regulatory purpose is served by imposing many DSA obligations on that site. The dynamics that led to the bad behavior targeted by the DSA can generally be traced to the profit motive and to choices about how to monetize engagement.[14] Although DuckDuckGo does try to make money, it doesn't do so in the kinds of ways that make platforms seem different from ordinary publishers (monetizing

12. *April 2023 Newsletter, Volume 177*, ARCHIVE OF OUR OWN (May 9, 2023), https://archiveofourown.org/admin_posts/25846.

13. *OTW Finance: 2023 Budget*, ARCHIVE OF OUR OWN (Apr. 22, 2023), https://archiveofourown.org/admin_posts/25468.

14. Katherine J. Wu provides a good summary of a common thesis: "Originally designed to drive revenue on social media platforms, recommendation algorithms are now making it easier to promote extreme content." Katherine J. Wu, *Radical Ideas Spread Through Social Media. Are the Algorithms to Blame?*, PBS (Mar. 28, 2019), https://www.pbs.org/wgbh/nova/article/radical-ideas-social-media-algorithms. Thus the regulators' focus on algorithms deployed by the large, for-profit services. *See, e.g.*, Maximilian Gahntz & Claire Perhsan, *Action Recommended: How the Digital Services Act Addresses Platform Recommender Systems* (Feb. 27, 2023), https://verfassungsblog.de/action-recommended/; Paddy Leerssen, *Algorithm Centrism in the DSA's Regulation of Recommender Systems* (Mar. 29, 2022), https://verfassungsblog.de/roa-algorithm-centrism-in-the-dsa; *The EU's Attempt To Regulate Big Tech: What it Brings and What is Missing*, EDRI (Dec. 18, 2020), https://edri.org/our-work/eu-attempt-to-regulate-big-tech/ (identifying different regulatory needs for the dominant providers); *27th Annual BTLJ-BCLT Symposium: From the DMCA to the DSA: Keynote and Copyright Interactions* (Apr. 7, 2023), https://tushnet.blogspot.com/2023/04/27th-annual-btlj-bclt-symposium-from_7.html (statement of Matthias Leistner that the DSA starts from the premise that platforms will use algorithms to moderate content).

information about users and trying to keep them scrolling). Likewise, as a nonprofit's website, AO3 doesn't try to make itself sticky for users or advertisers even though it has registered accounts.

The AO3's tracking can tell its maintainers how many page views or requests it gets per minute and how many page views come from which browsers, since those things can affect site performance. The AO3 can also get information on which sorts of pages or areas of the code see the most use, which coders can use to figure out where to put their energy when optimizing the code and fixing bugs. But the AO3 can't match that up to internal information about user behavior. The AO3 doesn't even track when a logged-in account is using the site, only the date of every initial login, and one login can cover many, many visits across months.

AO3 users regularly say they use the site multiple times a day (one game on social media is to report how many tabs users have open on the site). One can divide the number of visits from the European Union by some number to gesture at a number of monthly average users, but that number is only an estimate of the proper order of magnitude. AO3's struggles are perhaps extreme, but they are clearly not unique in platform metrics, even though counting average users must have sounded simple to policymakers. Perhaps the drafters didn't worry too much because they wanted to impose heavy obligations on almost everyone, but it seems odd to have important regulatory tiers without a reliable way to tell who's in which one.

These challenges in even initially sorting platforms into the DSA's tiers illustrates why regulation often generates more regulation—Husovec suggests that, "[g]oing forward, the companies should publish actual numbers, not just statements of being above or below the 45 million user threshold, and also their actual methodology."[15] But even that, as Wikipedia and DuckDuckGo's experiences show, would not necessarily be very illuminating. And the key question would remain: why is this important? What are we afraid of DuckDuckGo doing and is it even capable of doing those things if it doesn't collect this information? Imaginary metrics lead to imaginary results—Husovec objects to porn sites saying they have low monthly average users,[16] but if you choose a metric that doesn't have an actual definition it's unsurprising that the results are manipulable.

15. Husovec, *supra* note 1, at 4.
16. *Id.* at 4–5.

III. PROVIDING DUE PROCESS

A second example of the DSA's "one size fits some" design draws on the work of Philip Schreurs in his paper, *Differentiating Due Process In Content Moderation*.[17] Along with requiring hosting services to accompany each content moderation action affecting individual recipients of the service with statements of reasons, the DSA also obligates platforms—that aren't micro or small enterprises—to put specific due process protections in place. These obligations apply not just to account suspension or removal, but to acts that demonetize or downgrade any specific piece of content.[18]

The DSA also requires online platform service providers to provide recipients of their services with access to an effective internal complaint-handling system.[19] Although there's no notification requirement before *acting* against high-volume commercial spam, even when action is taken against high-volume commercial spam, platforms must provide redress systems. Platforms' decisions on complaints can't be based solely on automated means.

Further, when platforms are large enough, the DSA allows users affected by a platform decision to select any certified out-of-court dispute settlement body to resolve disputes relating to those decisions.[20] Such platforms must bear all the fees charged by the out-of-court dispute settlement body if the latter decides the dispute in favor of the user, while the user does not have to reimburse any of the platforms' fees or expenses if they lose unless the user manifestly acted in bad faith. Nor are there other constraints on bad-faith offenders, since Article 23 prescribes a specific method to address the problem of repeat offenders who submit manifestly unfounded notices: an initial warning explaining what was wrong with the notices, and then only a temporary suspension if the behavior continues. The platform must provide the notifier, who need not be a user, with the possibilities for redress identified in the DSA. Although platforms may "establish stricter measures in case of manifestly illegal content related to serious crimes,"[21] they still must provide these procedural rights.

This means that due process requirements are the same for removing a one-word comment as for removing a one-hour video: for removing a politician's entire account and for downranking a single post by a private figure

17. Philip Schreurs, *Differentiating Due Process in Content Moderation* (unpublished manuscript) (on file with author).

18. Digital Services Act art. 17, 2022 O.J. (L 277).

19. *Id.* art. 20.

20. *Id.* art. 21.

21. *Id.* art. 64.

that uses a slur. Schreurs suggests that the process due should instead be more flexible, depending on the user, violation, remedy, and type of platform.[22]

The existing inflexibility is a problem because every anti-abuse measure is also a mechanism of abuse. There may well be significant demographic differences in who is likely to appeal a moderation decision: Such differences are common in other areas in which the law provides the ability to make claims of right.[23] Meta's Oversight Board, for example, reported that more than two-thirds of all appeals of content moderation decisions came from the United States, Canada, and Europe in 2022, which was unrepresentative of actual user activity.[24] Differences in willingness to appeal can increase the impact of content moderation policies that already disfavor specific groups,[25] just as

22. Schreurs, *supra* note 17.

23. In general, the evidence suggests that willingness to make rights claims, and contest such claims, varies across predictable demographic lines. *See, e.g.*, Anna-Maria Marshall, *Injustice Frames, Legality, and the Everyday Construction of Sexual Harassment*, 28 LAW & SOCIAL INQUIRY 659 (2003) (finding gendered differences in willingness to make legal claims); Hugh M. McDonald & Julie People, *Legal Capability and Inaction for Legal Problems: Knowledge, Stress and Cost*, 41 UPDATING JUSTICE 1 (2014) (finding that willingness to make a legal complaint varies by socioeconomic status and education level); Roger Michalski, *The Pro Se Gender Gap*, 88 BROOKLYN L. REV. 563 (2023) (finding a gender gap in self-representation); Calvin Morrill, Karolyn Tyson, Lauren B. Edelman & Richard Arum, *Legal Mobilization in Schools: The Paradox of Rights and Race Among Youth*, 44 L. & SOC'Y REV. 651 (2010) (finding racial differences in willingness to make legal claims).

24. Oversight Board, Annual Report 2022, at 32; *see also id.* at 33 ("We recognize that these figures do not reflect the spread of Facebook and Instagram users worldwide, or the actual distribution of content moderation issues around the world.").

25. *See, e.g.*, Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, & Andrea Wegner, *Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas*, 5 PROCEEDINGS OF THE ACM ON HUMAN-COMPUT. INTERACTION 1, 27, https://dl.acm.org/doi/abs/10.1145/3479610 (noting that disproportionate content removals occurred for political conservatives, transgender people, and Black people; the first group of removals "often involved harmful content removed according to site guidelines to create safe spaces with accurate information, while transgender and Black participants' removals often involved content related to expressing their marginalized identities that was removed despite following site policies or fell into content moderation gray areas"); Brittan Heller, *Coca-Cola Curses: Hate Speech in a Post-Colonial Context*, 29 MICH. TECH. L. REV. 259, 263 (2023) ("It is doubtful that calling someone a Coca-Cola bottle [a racial slur in some African contexts] would violate the terms of service of a social media company utilizing a predominantly American perspective, unless the reference was seen as an infringement of intellectual property. These layers of social meaning likely would have evaded automated content moderation filters."); Oversight Board, *supra* note 24, at 12 ("Meta's policies on adult nudity result in greater barriers to expression for women, trans, and non-binary people on Facebook and Instagram."); Adi Robertson, *Tumblr is Settling With NYC's Human Rights Agency Over Alleged Porn Ban Bias*, VERGE (Feb. 25, 2022), https://www.theverge.com/2022/2/25/22949293/tumblr-nycchr-settlement-adult-content-ban-

copyright takedown notices disproportionately deter women and younger people from counternotifying.[26]

Relatedly, it is possible to use the system to harass other users and burden platforms by filing notices and appealing the denial of notices despite the supposed limits on bad faith. Even with legitimate complaints about removals, there will be variances in who feels entitled to contest the decision and who can afford to pay the initial fee and wait to be reimbursed. Such resources will not be universally or equitably available. The system can easily be weaponized by online misogynists who already coordinate attempts to get content from sex-positive feminists removed or demonetized.[27] We've already seen someone willing to spend $44 billion to get the moderation he wants,[28] and although that's an outlier, there is a demonstrated willingness to use procedural mechanisms to harass.

One result is that providers may be incentivized to cut back on moderation of lawful but awful content, the expenses of which can be avoided by not prohibiting it in the terms of service or not identifying violations, in favor of moderating only putatively illegal content.[29] But forcing providers to focus on decisions about, for example, what claims about politicians are false and which are merely rhetorical political speech is likely to prove unsatisfactory; the difficulty of those decisions suggests that increased focus may not help without a full-on judicial apparatus.

Relatedly, the expansiveness of DSA remedies may water down their availability in practice. Reviewers or dispute resolution providers may sit in front of computers all day, technically giving human review to automated violation detection but in practice just agreeing that the computer found what it found. ProPublica has found similar practices with respect to putatively

algorithmic-bias-lgbtq (discussing Tumblr's adult content ban, whose implementation allegedly disparately impacted LGBTQ users).

26. *See* Jonathon W. Penney, *Privacy and Legal Automation: The DMCA as a Case Study*, 22 STAN. TECH. L. REV. 412, 450 (2019) (finding gendered reactions to DMCA takedown notices; women were more likely to feel chilled in speech; younger respondents were also more likely to be chilled than older ones).

27. *See* Samantha Cole, *#ThotAudit Is Compiling Massive Databases of Sex Workers and Reporting Them to PayPal*, VICE (Dec. 4, 2018), https://www.vice.com/en/article/gy7wyw/thotaudit-databases-of-sex-workers-and-reporting-them-to-paypal.

28. *See* Caleb Ecarma, *We're Officially in the Elon Musk Era of Content Moderation*, VANITY FAIR (Nov. 21, 2022), https://www.vanityfair.com/news/2022/11/elon-musk-twitter-content-moderation.

29. *See* Ben Horton, *The Hydraulics of Intermediary Liability Regulation*, 70 CLEV. ST. L. REV. 201, 205, 234 (2022) (explaining that profit-driven firms will respond to greater intermediary liability by diverting resources from moderating "lawful but awful" content to focusing on allegedly illegal content).

mandatory human doctor review of insurance denials at certain U.S. insurance companies.[30]

And, of course, the usual anticompetitive problems of mandating one-size-fits-all due process are present in the DSA: full due process for every moderation decision benefits larger companies and hinders new market entrants by increasing the costs of growth or capping their growth potential. Such a system may also encourage designs that steer users away from complaining, like BeReal's intense focus on selfies or TikTok's continuous flow system that emphasizes showing users more like what they've already seen and liked—if someone is reporting large amounts of content, perhaps they should just not be shown that kind of content anymore. It is hard to predict the effects, other than to note that they are not obviously going to be in the direction of high-quality, truthful, and useful content.

Likewise, the DSA's existing provisions for excluding services that are only ancillary to some other kind of product—like comments sections on newspaper sites, for example[31]—are partial at best, since it will often be unclear what regulators will consider to be merely ancillary.[32] And the exclusion of ancillary services enhances, rather than limits, the problem of design incentives. It will be much easier to launch a new Netflix competitor than a new Facebook competitor as a result. Notably, even Meta hesitated to launch its new Threads app in the EU, pending a better understanding of the rules.[33]

IV. COPYRIGHT-SPECIFIC RULES

The DSA is not the only major European intervention into content moderation. It was enacted soon after the European Union also required countries to make new rules about copyright online. These copyright-specific rules are subject to the same basic problem. Based on complaints that were largely about YouTube or at least about major streaming sites, the European Union demanded changes from the internet as a whole.[34] But Ravelry—a site

30. Patrick Rucker, *Maya Miller & David Armstrong. How Cigna Saves Millions by Having Its Doctors Reject Claims Without Reading Them*, PROPUBLICA (Mar. 25, 2023), https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims.

31. Digital Services Act Recital 13, 2022 O.J. (L 277).

32. Even the comments sections are apparently subject to review for whether they are really ancillary. *Id.* ("For example, the comments section in an online newspaper *could* constitute such a feature, where it is *clear* that it is ancillary to the main service represented by the publication of news under the editorial responsibility of the publisher.") (emphasis added).

33. Makena Kelly, *Here's Why Threads Is Delayed in Europe*, VERGE (July 10, 2023), https://www.theverge.com/23789754/threads-meta-twitter-eu-dma-digital-markets.

34. *See generally* GLYN MOODY, WALLED CULTURE: HOW BIG CONTENT USES TECHNOLOGY AND THE LAW TO LOCK DOWN CULTURE AND KEEP CREATORS POOR 117–

focused on the fiber arts—is not YouTube. The cost-benefit analysis of copyright filtering is very different for a site that is for uploading patterns and pictures of knitting projects than for a site that is not subject-specific.

Sites like the Archive of Our Own receive very few valid copyright claims, whether considered as a percentage of works uploaded, time period, or any other metric, and so the relative burden of requiring YouTube-like filtering and licensing is both higher and less justified.[35] The differences are not just between websites, but between types of works. Negotiating with photographers for licensing is very different than negotiating with music labels, but the European Union's framework requires attempts to license from organizations representing copyright owners of all kinds. It assumes that the licensing bodies will be functioning pretty much the same no matter what type of work is involved.

It is possible that the new framework may be flexible enough to allow a service to decide that it doesn't have enough of a problem with a particular kind of content to require licensing negotiations, but only if the European authorities agree that the service is a "good guy."[36] And it's worth noting, since both Ravelry and the Archive of Our Own are heavily used by women and nonbinary people, that the concept of a "good guy" is likely both gendered and racially coded, which raises concerns about its application.

V. TOWARDS TRUE PROPORTIONALITY IN REGULATION

Ultimately, proportionality is much harder to achieve than just saying "we are regulating more than Google, and we will make special provisions for startups." To an American, the claim that the DSA has lots of checks and balances seems in tension with the claim made at the symposium, both by supporters and critics, that the DSA looks for good guys and bad guys. This is a system that works only if its subjects have very high levels of trust that the definitions of good and bad guys will be shared.

---

42 (2022) (describing Article 17, the struggle to implement it, and its practical filtering mandate).

35. *See generally* Jennifer M. Urban, Joe Karaganis & Brianna L. Schofield, *Notice and Takedown in Everyday Practice*, UC BERKELEY PUBLIC LAW RESEARCH PAPER 2755628 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628 (explaining the substantial divide between large sites that face high volumes of infringement claims and sites that don't).

36. The concept that regulators would accept mistakes by "good guys" was important to many of the defenses, and explanations, of the DSA offered at the symposium for which this contribution was prepared. *See* Rebecca Tushnet, *27th Annual BTLJ-BCLT Symposium: From the DMCA to the DSA—A Transatlantic Dialogue on Online Platform Liability and Copyright Law* (Apr. 7, 2023), https://tushnet.blogspot.com/2023/04/27th-annual-btlj-bclt-symposium-from.html (summarizing comments of João Quintais).

Regulators who are concerned with targeting specific behaviors, rather than just decreasing the number of online services, should make extensive use of test suites. Daphne Keller of Stanford and Mike Masnick of Techdirt proposed this two years ago.[37] Because regulators write with the giant names they know in mind, they tend to assume that all services have those same features and problems—they just add TikTok to their consideration set along with Google and Facebook. But Ravelry has very different problems than Facebook or even Reddit. There are many other examples of services that many people use, but not in the same way they use Facebook or Google: Zoom, Shopify, Patreon, Reddit, Yelp, Substack, Stack Overflow, Bumble (or your own favorite dating site), Ravelry (or your own favorite hobby-specific site), Bandcamp, LibraryThing, Archive of Our Own, and Etsy. They are used, and abused, in ways that don't match up with the DSA's assumptions.

A test suite can reveal regulators' assumptions about how online services work in ways that clarify regulatory goals and make them more achievable. If a relevant service doesn't have the features that regulators assumed all services did—for example, it doesn't track its users well enough to give reliable estimates about their numbers—then regulators have options. They can either exclude such services (because without tracking, they can't be manipulating user data in worrisome ways) or provide alternative rules. Wikipedia was big enough to make it into the DSA discussions, but most others weren't. The other, less "charismatic" platforms who weren't considered may be burdened most because they haven't built the automated systems and data collection for reporting purposes that the DSA essentially requires. Not only may those systems be unnecessary for particular sites, but many of them are now required to do things that Facebook and Google weren't *able* to do until they were much, much bigger.

VI. CONCLUSION: TAKING THE MULTITUDES INTO ACCOUNT

Although enforcement discretion can moderate the effects of impossible regulatory demands, discretion has its own dangers. Clearer recognition of the inevitable ambiguities and errors entailed by platform regulation can improve system design more than regulators' ad hoc consent to a failure to achieve the

37. @daphnehk, TWITTER (Feb 22, 2021, 5:53 AM) https://twitter.com/daphnehk/status/1363849276690849800; Mike Masnick, *The Internet Is Not Just Facebook, Google & Twitter: Creating A 'Test Suite' For Your Great Idea to Regulate the Internet*, TECHDIRT (Mar. 18, 2021), https://www.techdirt.com/2021/03/18/internet-is-not-just-facebook-google-twitter-creating-test-suite-your-great-idea-to-regulate-internet.

unachievable—and certainly more than that lenience alone.[38] Ordering websites to do things they can't, and then excusing them if they seem nice enough, risks both arbitrariness and non-arbitrary discrimination against politically unpopular sites or users, especially in an age of democratic retrenchment.

The more complex the regulation, the more regulatory interactions need to be managed. Thinking about fifty or so different models of online services and considering how and indeed whether they should be part of this regulatory system could have substantially improved the DSA. Not all processes should be the same, just like not all websites should be the same, unless we want our only options to be Meta and YouTube.

38. It's true that many prevention mandates could be achieved by platforms going out of business—no social media, no social media disinformation—but neither the underlying problems (disinformation now spreading by email and word of mouth) nor the goals of regulation (better functioning social media) would thereby be achieved, so I am comfortable with the claim that full compliance is regularly going to be impossible.

HOW THE EUROPEAN UNION OUTSOURCES THE TASK OF HUMAN RIGHTS PROTECTION TO PLATFORMS AND USERS: THE CASE OF USER-GENERATED CONTENT MONETIZATION

Martin Senftleben,[†] João Pedro Quintais[††] & Arlette Meiring[†††]

ABSTRACT

With the shift from the traditional safe harbor for hosting to statutory content filtering and licensing obligations, the 2019 E.U. Directive on Copyright in the Digital Single Market (CDSMD) has substantially curtailed the freedom of users to upload and share their content creations online. Seeking to avoid overbroad inroads into freedom of expression, E.U. law obliges online platforms and the creative industry to consider human rights when coordinating their content filtering actions. Platforms must also establish complaint and redress procedures for users. Organizing stakeholder dialogues, the European Commission will seek to identify best practices. These "safety valves" in the legislative package, however, are mere fig leaves. Instead of safeguarding human rights, the E.U. legislature outsources human rights obligations to the platform industry. At the same time, the burden of policing content moderation systems is imposed on users who are unlikely to bring complaints in each individual case. The new legislative design in the European Union is likely to "conceal" human rights violations instead of bringing them to light. Nonetheless, the Digital Services Act (DSA) rests on the same problematic approach.

Against this background, we discuss the weakening—and potential loss—of fundamental freedoms because of the departure from the traditional notice-and-takedown approach in the European Union and the reliance on platform and user action to prevent human rights violations. Our analysis adds a new element to the ongoing debate on content licensing and filtering. Namely, we focus on how E.U. law has largely left the private power of platforms untouched to determine the "house rules" that govern the (algorithmic) monetization of detected matches between protected works and content uploads. Addressing the "legal vacuum" in the field of content monetization, we explore outsourcing and concealment risks in this unregulated space. Focusing on large-scale platforms for user-generated content, such as YouTube, Instagram and TikTok, two normative problems come to the fore: (1) the fact that rightholders, when opting for monetization, de facto monetize not only their own works

---

DOI: https://doi.org/10.15779/Z381G0HW20

† Professor of Intellectual Property Law and Director, Institute for Information Law (IViR), University of Amsterdam; Of Counsel, Bird & Bird, The Hague, The Netherlands.

†† Assistant Professor, Institute for Information Law (IViR), University of Amsterdam, The Netherlands. João Pedro Quintais's research in this Article is part of the VENI Project "Responsible Algorithms: How to Safeguard Freedom of Expression Online" funded by the Dutch Research Council (grant number: VI.Veni.201R.036).

††† Junior Researcher, Institute for Information Law (IViR), University of Amsterdam, The Netherlands.

but also the creative input of users; and (2) the fact that user creativity remains unremunerated as long as the monetization option is only available to rightholders. As a result of this configuration, the monetization mechanism disregards users' right to (intellectual) property and discriminates against user creativity. In this light, we discuss whether the DSA provisions that seek to ensure transparency of content moderation actions and terms and conditions offer useful sources of information that could empower users. We further raise the question whether the detailed regulation of platform actions in the DSA may resolve the described human rights dilemmas to some extent.

TABLE OF CONTENTS

I. INTRODUCTION

User-generated content (UGC)[1] is a core element of many internet platforms. With the opportunity to upload photos, films, music and texts, formerly passive users have become active contributors to (audio-)visual content portals, wikis, online marketplaces, discussion and news fora, social networking sites, virtual worlds, and academic paper repositories. Internet users upload a myriad of literary and artistic works every day.[2] A delicate question arising from this user involvement concerns copyright infringement. UGC may consist of self-created works and public domain material. However, it may also include unauthorized takings of third-party material that enjoys copyright protection. As UGC has become a mass phenomenon and a key factor in the evolution of the modern, participative web,[3] this problem raises complex issues and requires the reconciliation of human rights[4] ranging from the right to property,[5] to freedom of expression and information,[6] and freedom

1. For a definition and description of central UGC features, see SACHA WUNSCH-VINCENT & GRAHAM VICKERY, WORKING PARTICIPATIVE WEB: USER-CREATED CONTENT 8–12 (2007), https://www.oecd.org/sti/38393115.pdf.

2. For example, statistics relating to the online platform YouTube report over one billion users uploading 300 hours of video content every minute. *Cf. About Youtube*, YOUTUBE, https://www.youtube.com/intl/en-GB/yt/about/press/ (last visited Sept. 6, 2023); *Youtube Company Statistics*, STATISTIC BRAIN RSCH. INST., https://www.statisticbrain.com/youtube-statistics/ (last visited Sept. 6, 2023).

3. WUNSCH-VINCENT & VICKERY, *supra* note 1, 8–22.

4. In this Article, the terms human rights and fundamental rights are used interchangeably.

5. In the EU, the fundamental right to property enshrined in Article 17 of the Charter of Fundamental Rights of the European Union (2000/C 364/01) (CFR) explicitly refers to intellectual property in paragraph 2.

6. Article 11 CFR; Article 10 EUR. CONV. ON H.R. *Cf.* Martin Senftleben, *User-Generated Content – Towards a New Use Privilege in EU Copyright Law*, *in* RESEARCH HANDBOOK ON IP AND DIGITAL TECHNOLOGIES 136, 148–52 (Tanya Aplyn ed., 2020), https://papers.ssrn.com/abstract=3325017.

to conduct a business.[7] Users, platform providers, and copyright holders are central stakeholders.[8]

In line with the approach taken in the U.S. Digital Millennium Copyright Act (DMCA),[9] E.U. legislation in the field of ecommerce traditionally shielded UGC platforms from liability for copyright infringement by offering a liability exemption or "safe harbor" for hosting services. To qualify for the safe harbor, a hosting platform, provided that it was not actively involved in the posting of content, was only obliged to take immediate action and remove content when a rightholder informed the platform provider in a sufficiently precise and substantiated manner about infringing content.[10] The safe harbor system was

7. Article 16 CFR. *Cf.* CJEU, 16 February 2012, case C-360/10, Sabam/Netlog, ¶ 51. *Cf.* Christophe Geiger & Bernd Justin Jütte, *Platform Liability Under Art. 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match*, 70 GRUR INT'L 517 (2021); Martin Senftleben & Christina Angelopoulos, *The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the E-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market* 16–20 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3717022.

8. As to the debate on user-generated content and the need for the reconciliation of divergent interests in this area, see Martin Senftleben, *Breathing Space for Cloud-Based Business Models – Exploring the Matrix of Copyright Limitations, Safe Harbours and Injunctions*, 4 J. INTELL. PROP., INFO. TECH. & E-COMMERCE L. 87, 87–90 (2013); Steven D. Jamar, *Crafting Copyright Law to Encourage and Protect User-Generated Content in the Internet Social Networking Context*, 19 WIDENER L.J. 843 (2010); Natali Helberger, Lucie Guibault, E.H. Janssen, N.A.N.M. van Ejik, Christina Angelopoulus & Joris van Hoboken, *Legal Aspects of User Created Content*, 19 WIDENER L.J. 843 (2020); Mary W. S. Wong, *"Transformative" User-Generated Content in Copyright Law: Infringing Derivative Works or Fair Use?*, 11 VAND. J. ENT. & TECH. L. 1075 (2021); Edward Lee, *Warming Up to User-Generated Content*, 5 U. ILL. L. REV. 1459 (2008); Branwen Buckley, *SueTube: Web 2.0 and Copyright Infringement*, 31 COLUM. J.L. & ARTS 235 (2008); Tom W. Bell, *The Specter of Copyism v. Blockheaded Authors: How User-Generated Content Affects Copyright Policy*, 10 VAND. J. ENT. & TECH. L. 841 (2008); Steven Hechter, *User-Generated Content and the Future of Copyright: Part One – Investiture of Ownership*,10 VAND. J. ENT. & TECH. L. 863 (2008); Greg Lastowka, *User-Generated Content and Virtual Worlds*, 10 VAND. J. ENT. & TECH. L. 893 (2008).

9. *Cf.* Miquel Peguera, *The DMCA Safe Harbour and Their European Counterparts: A Comparative Analysis of Some Common Problems*, 32 COLUM. J.L. & ARTS 481 (2009). More recently, see Folkert Wilman, *The EU's System of Knowledge-Based Liability for Hosting Service Providers in Respect Of Illegal User Content – Between The E-Commerce Directive and the Digital Services Act*, 12 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. (2021), http://www.jipitec.eu/issues/jipitec-12-3-2021/5343.

10. Article 6(1) of 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), *Official Journal of the European Union* 2022 L 277, 1, and, previously, Article 14(1) of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce), *Official Journal of the European Communities* 2000 L 178, 1. *Cf.* CJEU, 23 March 2010, case C-236/08, Google and Google France, ¶¶ 114–18; CJEU, 12 July 2011, case C-324/09, L'Oréal v. eBay, ¶¶ 120–22. For commentary, see S. Kulk, *Internet Intermediaries and Copyright Law – Towards a*

based on the assumption that a general monitoring obligation would be too heavy a burden for platform providers and undesirable as a matter of public policy.[11] Without a safe harbor, the liability risk would thwart the creation of internet platforms depending on third-party content and frustrate the development of ecommerce.[12]

However, in preparing an update of E.U. copyright legislation and a departure from the notice-and-takedown consensus, the European Commission stated that the hosting safe harbor allowed UGC platforms to generate income without sharing the profits with producers of creative content.[13] In line with this "value gap" argument, the Commission's proposal for new copyright legislation—the template for Article 17 of the E.U. Directive on Copyright in the Digital Single Market (CDSMD, or "CDSM Directive")[14]—sought to render the liability shield inapplicable to copyrighted works.[15] Article 17 has been described as the "monster provision" of the

Future-Proof EU Legal Framework, Utrecht: University of Utrecht 2018; Martin Senftleben, *Breathing Space for Cloud-Based Business Models: Exploring the Matrix of Copyright Limitations, Safe Harbours and Injunctions*, 4 J INTELL. PROP. INFO. TECH. & ELEC. COM. L. 87, 87–103 (2013); Peguera, *supra* note 9; CHRISTINA ANGELOPOULOS, EUROPEAN INTERMEDIARY LIABILITY IN COPYRIGHT: A TORT-BASED ANALYSIS (2016); MARTIN HUSOVEC, INJUNCTIONS AGAINST INTERMEDIARIES IN THE EUROPEAN UNION: ACCOUNTABLE BUT NOT LIABLE? (2017).

11. *See* SENFTLEBEN & ANGELOPOULOS, *supra* note 7, at 16–20.

12. Article 15(1) of Directive 2000/31/EC of 8 June 2000 (E-Commerce Directive).

13. *See* European Commission, 9 December 2015, *Towards A Modern, More European Copyright Framework*, Doc. COM (2015) 626 final, at 9–10.

14. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, *Official Journal of the European Communities* 2019 L 130, 92 (CDSM Directive or CDSMD).

15. European Commission, *Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market*, Art. 13, COM (2016) 593 final (Sept. 14, 2016). Prior to this formal proposal of copyright legislation seeking to neutralize the safe harbour for hosting, the French High Council for Literary and Artistic Property had published a research paper prepared by Professor Pierre Sirinelli, Josée-Anne Benazeraf and Alexandra Bensamoun on November 3, 2015. The researchers had been asked to propose changes to current E.U. legislation "enabling the effective enforcement of copyright and related rights in the digital environment, particularly on platforms which disseminate protected content." They arrived at the conclusion that a provision should be added to current E.U. copyright legislation making it clear that "information society service providers that give access to the public to copyright works and/or subject-matter, including through the use of automated tools, do not benefit from the limitation set out [in the safe harbour for hosting of the E-Commerce Directive 2000/31/EC]." *See* High Council for Literary and Artistic Property of the French Ministry of Culture and Communication, 3 November 2015, *Mission to Link Directives 2000/31 and 2001/29 – Report and Proposals*, p. 11.

CDSM Directive, "both by its size and hazardousness."[16] Despite its relatively young age, it has already triggered abundant commentary.[17] The provision has also been subject to an interpretative Guidance by the European Commission ("Commission Guidance" or "Guidance"),[18] and survived an action for annulment with the Court of Justice of the European Union (CJEU).[19]

The regulatory strategy underlying Article 17 of the CDSMD is simple: deprived of the safe harbor for hosting and exposed to direct liability for infringing user uploads, platform providers will have to embark on UGC licensing and filtering.[20] In the final text of the Directive, the E.U. legislature applied this approach to a specific type of online platforms: online content-

16. Séverine Dusollier, *The 2019 Directive on Copyright In The Digital Single Market: Some Progress, A Few Bad Choices, And An Overall Failed Ambition*, 57 COMMON MKT. L. REV. 979 (2020).

17. *See, e.g.*, Martin Senftleben, *Bermuda Triangle – Licensing, Filtering and Privileging User-Generated Content Under the New Directive on Copyright in the Digital Single Market*, 41 EUR. INTELL. PROP. REV. 480 (2019); Martin Husovec & João Pedro Quintais, *How to License Article 17? Exploring the Implementation Options for the New EU Rules on Content-Sharing Platforms under the Copyright in the Digital Single Market Directive*, 70 GRUR INT'L 325 (2021); Matthias Leistner, *European Copyright Licensing and Infringement Liability Under Art. 17 DSM-Directive Compared to Secondary Liability of Content Platforms in the U.S. – Can We Make the New European System a Global Opportunity Instead of a Local Challenge?*, 2 ZEITSCHRIFT FÜR GEISTIGES EIGENTUM/INTELL. PROP. J. 123 (2020); Christophe Geiger & Bernd Justin Jütte, *Towards a Virtuous Legal Framework for Content Moderation by Digital Platforms in the EU? The Commission's Guidance on Article 17 CDSM Directive in the Light of the YouTube/Cyando Judgement and the AG's Opinion in C-401/19*, 43 EUR. INTELL. PROP. REV. 625 (2021); Axel Metzger & Martin Senftleben, *Understanding Article 17 of the EU Directive on Copyright in the Digital Single Market - Central Features of the New Regulatory Approach to Online Content-Sharing Platforms*, 67 J. COPYRIGHT SOC'Y U.S.A. 279 (2020).

18. *See* Communication from the Commission to the European Parliament and the Council, *Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market*, COM/2021/288 final [hereinafter Guidance Art. 17 CDSMD].

19. Case C-401/19, Republic of Poland v European Parliament and Council of the European Union, 26.04.2022, ECLI:EU:C:2022:297. For commentary, see João Pedro Quintais, *Between Filters and Fundamental Rights: How the Court of Justice saved Article 17 in C-401/19 - Poland v. Parliament and Council*, VERFASSUNGSBLOG (2022), https://verfassungsblog.de/filters-poland/; Martin Husovec, *Mandatory Filtering Does Not Always Violate Freedom of Expression: Important Lessons From Poland v. Council and European Parliament*, 60 COMMON MKT. L. REV. 173 (2023); Bernd Justin Jutte & Giulia Priora, *On the Necessity of Filtering Online Content and Its Limitations: AG Saugmandsgaard Øe Outlines the Borders of Article 17 CDSM Directive*, KLUWER COPYRIGHT BLOG (July 20, 2021), http://copyrightblog.kluweriplaw.com/2021/07/20/on-the-necessity-of-filtering-online-content-and-its-limitations-ag-saugmandsgaard-oe-outlines-the-borders-of-article-17-cdsm-directive/.

20. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (CDSMD), O.J. (L 130), art. 17(3).

sharing service providers (OCSSPs).[21] OCSSPs are providers of an information society service. Furthermore, their main purpose is to store and give the public access to a large amount of protected content by its users, which they organize and promote for profit-making purposes. In assessing whether a platform qualifies as an OCSSP, it is important to examine a relevant service's substitution effects and to make a case-by-case assessment of their profit orientation.[22] Recital 62 of the CDSMD clarifies that the definition is intended to confine the application of Article 17 to online services that play an important role on the online content market "by competing with other online content services, such as online audio and video streaming services, for the same audiences."[23]

The scope of the OCSSP concept is further delineated by a non-exhaustive list of exclusions, i.e., electronic communication services (e.g., Skype), providers of cloud services (e.g., Dropbox), online marketplaces (e.g., eBay), not-for-profit online encyclopedias (e.g., Wikipedia), not-for-profit educational and scientific repositories (e.g., ArXiv.org), and open-source software developing and sharing platforms (e.g., GitHub).[24] Nonetheless, legal uncertainty remains. While it is safe to assume that certain large-scale platforms, especially platforms with video-sharing features (e.g., YouTube, Facebook, Instagram), are covered, others do not so easily fit the concept. The definition includes several open-ended concepts ("main purpose," "large amount," "profit-making purpose") that ultimately require a case-by-case assessment of what providers qualify as OCSSPs.[25]

In practice, the adoption of Article 17 of the CDSMD means that OCSSPs seeking to avoid liability must enter into agreements with copyright holders. The initial Commission proposal already contemplated that this regulatory approach would bring content filtering obligations, then referred to as "effective content recognition technologies."[26] If a platform provider does not manage to conclude sufficiently broad licensing agreements with rightholders, Article 17(4)(b) of the CDSMD offers the prospect of a reduction of the

21. For the definition of this type of online platforms, see *id.* art. 2(6) and the further guidance provided in Recitals 62 and 63.

22. *Id.* recitals 62 and 63. *Cf.* Metzger & Senftleben, *supra* note 17.

23. CDSMD recital 62, 2019 O.J. (L 130).

24. *Id.* art. 2(6).

25. *See* Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 4–5. For analysis and criticism, see JOÃO PEDRO QUINTAIS, PÉTER MEZEI, ISTVÁN HARKAI, JOÃO C. MAGALHÃES, CHRISTIAN KATZENBACH, SEBASTIAN FELIX SCHWEMER & THOMAS RIIS, COPYRIGHT CONTENT MODERATION IN THE EU: AN INTERDISCIPLINARY MAPPING ANALYSIS (2022).

26. *See Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market*, COM (206) 593 final (Sept. 14, 2016), art. 13(1).

liability risk in exchange for content filtering and other preventive measures. If the platform—despite best efforts[27]—has not received a license, it can avoid liability for unauthorized acts of communication to the public or making available to the public when it manages to demonstrate that it:

> made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rightholders have provided the service providers with the relevant and necessary information . . . .[28]

Despite the neutral wording, it is clear that "unavailability of specific works and other subject matter" requires the use of algorithmic filtering tools.[29] In the legislative process leading to this remarkable paradigm shift in the European Union, the human rights impact of the departure from the traditional notice-and-takedown model has not gone unnoticed.

Algorithmic content moderation—including the use of automated filtering tools to detect infringing content before it appears online—has a deep impact on the freedom of users to upload and share information. When an algorithmic content recognition system identifies protected source material matching a platform's reference files in a user upload, the system can be used to prevent content from appearing in the first place. Instead of presuming that UGC is lawful until proven infringing, the default position of automated filtering systems is that every upload is suspicious and that copyright owners are entitled to ex ante control over the sharing of information online.

The wording of Article 17 of the CDSMD itself shows that the new legislative design has given rise to concerns about overbroad inroads into human rights. Article 17(10) of the CDSMD stipulates that, in stakeholder dialogues seeking to identify best practices for the application of content moderation measures, "special account shall be taken, among other things, of the need to balance fundamental rights and of the use of exceptions and limitations."[30] After the adoption of the Directive, the preparation of the Digital Services Act (DSA)[31] offered further opportunities for the E.U. legislature to refine and stabilize its strategy for safeguarding human rights that

27. CDSMD art. 17(4)(a), 2019 O.J. (L 130).

28. *Id.*

29. CJEU, 26 April 2022, case C-401/19, Poland v Parliament and Council, where this assumption has been confirmed.

30. CDSMD art. 17(10), 2019 O.J. (L 130).

31. Regulation 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022, on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), O.J. (L 277) 1 (EU).

may be affected by algorithmic content filtering tools. Article 14 of the DSA—regulating terms and conditions of intermediary services ranging from mere conduit and caching to hosting services[32]—reflects central features of the E.U. strategy.[33] Article 14(1) of the DSA requires that providers of hosting services—the category covering UGC platforms—inform users about:

> any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review, as well as the rules of procedure of their internal complaint handling system.[34]

This information duty indicates that users are expected to play an active role in the preservation of their freedom of expression and information. Article 14(4) of the DSA complements this transparency measure with a fundamental rule that goes far beyond sufficiently clear and accessible information in the terms and conditions. Providers of intermediary services, including UGC platforms:[35]

> shall act in a diligent, objective and proportionate manner in applying and enforcing the restrictions [that they impose in relation to the use of their service in respect of information provided by the recipients of the service], with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.[36]

In other words: in the case of upload- and content-sharing restrictions following from the employment of content moderation tools, the UGC platform is bound to some (imprecise) extent[37] to safeguard the fundamental rights of users, including the freedom of expression and information. As a guiding principle, Article 14(4) of the DSA refers to the principle of

32. *See* Digital Services Act art. 3(g), 2022 O.J. (L 277) (defining "intermediary services").

33. For a detailed analysis of Article 14 DSA, see João Pedro Quintais, Naomi Appelman & Ronan Fahy, *Using Terms and Conditions to Apply Fundamental Rights to Content Moderation*, 24 GERMAN L.J. 881 (2023).

34. Digital Services Act art. 14(1), 2022 O.J. (L 277).

35. These providers are covered by the concept of "online platforms" in Article 3(i) DSA. *Id.* art. 3(i).

36. *Id.* art.14(4).

37. There is a complex discussion concerning the extent to which fundamental rights can have (indirect) horizontal effect, i.e., as between private parties (here: platform and user), and how Article 14 of the DSA changes pre-existing E.U. law in this respect. *See* Quintais, Appelman & Fahy, *supra* note 33.

proportionality[38] that plays a central role in the reconciliation of competing fundamental rights under Article 52(1) of the E.U. Charter of Fundamental Rights ("Charter" or CFR).[39]

At first glance, it seems plausible to impose on platforms the obligation to safeguard fundamental rights of users, since they are closest to users and arguably best equipped to deal quickly with complex issues arising from infringement on a case-by-case basis.[40] The crux of the approach chosen in Article 14(4) of the DSA, however, clearly comes to the fore when raising the question whether the possibility of imposing human rights obligations on internet service providers exempts the state power itself from the noble task of ensuring the observance of fundamental rights. Can the legislature legitimately "outsource" the obligation to safeguard fundamental rights, such as freedom of expression and information, to private parties? And can the legislature, when passing on that responsibility, confidently leave the task of defending the public interest in this sensitive area in the hands of companies belonging to the platform and creative industry and to the users who may not lodge complaints in each individual case?

We will discuss these outsourcing questions—and the risk of platforms and public authorities hiding behind a low number of user complaints—in Part II. We will then turn to human rights risks in a detailed case study focusing on UGC monetization in Part III. While largely underexplored, UGC monetization is highly relevant in practice. On platforms, it constitutes a much more popular moderation action than blocking. In addition, it is very significant from a human rights perspective. In the absence of appropriate regulation, the monetization practices of platforms and large-scale rightholders may make inroads into human rights. Part IV provides an overview of our findings and recommendations.

---

38. Digital Services Act art. 14(4), 2022 O.J. (L 277) (referring to "proportionate manner").

39. Charter of Fundamental Rights of the European Union, Official Journal of the European Communities 2000 C 364, 1. Article 52(1) CFR reads as follows: "Any limitation on the exercise of the rights and freedoms ecognized [sic] by this Charter must be provided for by law and respect the essence of those rights and freedoms. Subject to the principle of proportionality, limitations may be made only if they are necessary and genuinely meet objectives of general interest ecognized [sic] by the Union or the need to protect the rights and freedoms of others."

40. For earlier case law already pointing in this direction, see CJEU, 27 March 2014, case C-314/12, UPC Telekabel Wien, ¶¶ 55–56, where the Court stated that internet service providers had to safeguard the fundamental rights of users; *see also* Christophe Geiger & Elena Izyumenko, *The Role of Human Rights in Copyright Enforcement Online: Elaborating a Legal Framework for Website Blocking*, 32 AM. U. INT'L L. REV. 43 (2016).

II. THE NEW CONSTITUTIONALISM DILEMMA: OUTSOURCING AND CONCEALING

Legislation that applies outsourcing strategies refrains from providing concrete solutions for human rights tensions in the law itself. Instead, the legislature imposes the burden on private entities to safeguard human rights that may be affected by the legislative measure at issue, such as the content filtering obligation in Article 17(4) of the CDSMD. In the case of UGC, the addressees of this type of outsourcing legislation are online platforms—OCSSPs—that offer users a forum for uploading and sharing their creations. The decision to outsource the burden of human rights balancing can be seen as the result of the legislature's inability to keep pace with rapid technological developments. In the absence of sufficient expertise and insight to devise concrete rules for the reconciliation of competing human rights positions, the legislature resorts to general guidelines—in the European Union, typically inspired by the principle of proportionality—which private entities must observe when fulfilling their obligation to implement the legislative measure in a way that preserves the human rights at stake. In the following sections, we discuss the corrosive effect of this outsourcing strategy (in Section II.A) and focus on the inadequacy of complaint and redress mechanisms for users as tools to bring human rights deficits to light (in Section II.B). Instead, we argue, reliance on users will amplify the risk of human rights violations when low reported numbers of complaints are used strategically to declare automated content filtering unproblematic (in Section II.C).

A. OUTSOURCING OF HUMAN RIGHTS OBLIGATIONS IN THE EUROPEAN UNION: REGULATION OF CONTENT MODERATION

Discussing the increasing tendency to take refuge in human rights outsourcing, Tuomas Mylly has observed that "gradually, intermediaries and other key private entities become more independent regulators."[41] He describes central characteristics of this process as follows:

> Courts are starting to rely increasingly on private entities to balance and adjust rights on technological domains but seek to secure formal appeal rights for users. Similarly, when legislatures shift decision-making power to intermediaries, they try to maintain some of the safeguards of traditional law and write wish-lists for private regulators. The executive pushes private regulation further to compensate for its policy failures and enters—at the request of the

41. Tuomas Mylly, *The New Constitutional Architecture of Intellectual Property*, *in* GLOBAL INTELLECTUAL PROPERTY PROTECTION AND NEW CONSTITUTIONALISM – HEDGING EXCLUSIVE RIGHTS 50, 71 (Jonathan Griffiths & Tuomas Mylly eds., 2021).

> legislature—into regulatory conversations with private regulators to issue "guidance" in the spirit of co-regulation, thus establishing an enduring link to private regulators.[42]

Arguably, Article 17 of the CDSMD and Article 14 of the DSA offer prime examples of provisions that outsource human rights obligations to private entities with the features Mylly describes. As explained above, Article 14(4) of the DSA places an obligation on intermediaries to apply their terms and conditions to content moderation restrictions in "a diligent, objective and proportionate manner."[43] In addition to this reference to the principle of proportionality, the provision emphasizes that online platforms are bound to carry out such restrictions (which would include content filtering) with due regard to the fundamental rights of users, such as freedom of expression.[44]

As to copyright limitations that support the freedom of expression, more specific rules follow from Article 17(7) of the CDSMD: the cooperation between OCSSPs and the creative industry in content moderation[45] must not result in the blocking of non-infringing UGC, including situations where UGC falls within the scope of a copyright limitation. Confirming Mylly's prediction that the executive power will enter regulatory conversations with private entities to establish best practices and guiding principles, Article 17(10) of the CDSMD adds that the European Commission shall organize stakeholder dialogues to discuss best practices for the content filtering cooperation:

> The Commission shall, in consultation with online content-sharing service providers, rightholders, users' organisations and other relevant stakeholders, and taking into account the results of the stakeholder dialogues, issue guidance on the application of this Article, in particular regarding the [content moderation] cooperation referred to in paragraph 4.[46]

In the quest for best practices, Article 17(10) of the CDSMD specifically requires that stakeholder dialogues take "special account"[47] of the need to balance fundamental rights. As in Article 14(4) of the DSA, reference is thus made to human rights tensions, although in a different context. The private entities involved in Article 17(10) of the CDSMD—copyright holders and

42. *Id.*

43. Digital Services Act art. 14(4), 2022 O.J. (L 277). *Id.* art. 14(1) explicitly refers to content moderation measures.

44. *Id.* art. 14(4).

45. *See* the interplay of creative industry notifications and filtering measures applied by the platform industry that results from CDSMD art. 17(4)(b)(c), 2019 O.J. (L 130).

46. *Id.* art. 17(10).

47. *Id.*

OCSSPs—are expected to resolve these tensions in the light of the guidance evolving from the co-regulatory efforts of the European Commission. Evidently, industry "cooperation" is the kingpin of this outsourcing scheme. To shed light on the human rights implications of this regulatory approach, we inspect the interplay of licensing and filtering obligations in Article 17 of the CDSMD more closely (in Section II.A.1). On this basis, it is possible to assess the industry cooperation resulting from this content moderation scheme (as discussed in Section II.A.2) and address the tension between abstract diligence and proportionality obligations on the one hand, and concrete cost and efficiency considerations on the other (as discussed in Section II.A.3). Considering the practical impact of the outsourcing approach underlying Article 17 of the CDSMD, the conclusion seems inescapable that, despite all references to diligence and proportionality, there is a serious risk of encroachments upon fundamental rights (as discussed in Section II.A.4).

1. Interplay of Licensing and Filtering Obligations in Article 17 of CDSMD

At the root of the obligation to filter UGC—and industry cooperation in this area—lies the grant of a specific exclusive right in Article 17(1) of the CDSMD that leads to strict, primary liability of OCSSPs for infringing content that is uploaded by users:

> Member States shall provide that an online content sharing service provider performs an act of communication to the public or an act of making available to the public when it gives the public access to copyright protected works or other protected subject matter uploaded by its users.[48]

By clarifying that the activities of platform providers amount to communication or making available to the public, this provision collapses the traditional distinction between primary liability of users who upload infringing content, and secondary liability of online platforms that encourage or contribute to infringing activities. It no longer matters whether the provider of a platform had knowledge of infringement, encouraged infringing uploads, or failed to promptly remove infringing content after receiving a notification. Instead, the platform provider is *directly* and *primarily* liable for infringing content that arrives at the platform. To reduce the liability risk, the platform provider will have to obtain a license from the rightholder for UGC uploads.

The Commission Guidance on "best efforts"[49] to obtain an authorization in the sense of Article 17(4)(a) of the CDSMD suggests in this context that

48. *Id.* art.17(1).

49. All obligations of "best efforts" must be interpreted in accordance with the principle of proportionality and the factors described in Article 17(5) CDSMD: the type, the audience

platforms should make a case-by-case analysis of licensing options.[50] At a minimum, OCSSPs should proactively engage with easily identifiable and locatable rightholders, especially collecting societies.[51] The proportions of this obligation may differ according to the type of platform. For instance, smaller platforms may only need to engage with a limited number of easily identifiable rightholders. The Guidance also suggests that unreasonable rightholder refusals to license should release platforms from their obligation to seek authorization.[52] However, platforms must be able to provide evidence for this, which could be challenging in practice. In cases where a certain type of content is not prevalent on a platform, platforms may be exempted from the task of seeking licenses proactively. Nonetheless, platforms should engage with rightholders who offer them.[53]

While these guidelines may help OCSSPs operationalize licensing, the focus on easily identifiable rightholders confirms that, in practice, licensing will hardly ever cover all conceivable UGC that may arrive at a platform. As the Guidance notes, collecting societies have an established position in the European Union.[54] With broad mandates to administer the rights of copyright owners, especially authors,[55] they seem natural partners to OCSSPs in the development of umbrella licensing solutions. However, they would have to offer an all-embracing licensing deal covering protected content of both members and non-members. Otherwise, the licensing exercise would make little sense, as it would fail to cover all types of conceivable user uploads.

Considering experiences with licensing packages offered by collecting societies in the past, it seems safe to assume that an umbrella solution with these proportions is currently unavailable in many E.U. Member States. It remains to be seen whether harmonized rules on extended collective licensing[56]

---

and the size of the service and the type of protected content uploaded by their users; and the availability of suitable and effective means and their cost for OCSSPs. Furthermore, Article 17(6) CDSMD creates a mitigated regime for platforms that are "new service providers with small turnover and audience." *See* Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 16–17.

50. Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 8–10.

51. *Id. See* Metzger & Senftleben, *supra* note 17, at 287–91.

52. As to this point, see *id.* at 289 (drawing a line with the FRAND requirement in standard essential patent cases, in particular CJEU, 16 July 2015, case C-170/13, Huawei v ZTE, ¶¶ 63–69).

53. Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 8–10.

54. *Id.* at 6.

55. *Id.*

56. Article 12 CDSMD. For a more detailed discussion of this licensing approach in the context of Article 17 CDSMD, see Husovec & Quintais, *supra* note 17. As to the discussion of extended licensing solutions in the area of orphan works, see Stef van Gompel, *Unlocking the Potential of Pre-Existing Content: How to Address the Issue of Orphan Works in Europe?*, 38 INT'L REV. INTELL. PROP. & COMP. L. 669 (2007).

will finally pave the way for broader and more flexible licensing solutions. However, even if an OCSSP finds a collecting society willing to enter a UGC agreement with an umbrella effect, a core problem of European licenses remains: the collecting society landscape is highly fragmented. The UGC license available in one Member State may remain limited to the territory of that Member State. Pan-European licenses are the exception, not the rule. If a collecting society offers Pan-European licenses for digital use, these licenses will be confined to specific repertoire, in respect of which the collecting society has a cross-border entitlement.[57]

Problems also arise in the field of initiatives to obtain licenses directly from rightholders. In the music industry, the willingness to grant licenses covering a broad spectrum of musical works may be relatively high.[58] Existing services, such as Spotify, demonstrate the availability of far-reaching licenses that encompass recent music releases. In the film industry, however, the situation is markedly different.[59] Film studios use diverse strategies and distribution outlets that do not include UGC platforms. They are unlikely to sacrifice profitable exploitation avenues by permitting users to share audio-visual material on UGC platforms from day one of the theatrical release (or availability on paid streaming platforms). This would enable UGC platforms to enter direct competition with the primary exploitation strategy pursued by the film studio itself. If there is willingness to conclude UGC licensing agreements despite these concerns, film studios will only accept agreements with limited use permissions that do not jeopardize their own opportunities to exploit the film in several stages and uphold their current windowing system.

Hence, creative users seeking to contribute to the online marketplace of ideas are at the mercy of copyright holders. Without contractual permission covering their content uploads, they cannot exercise their freedom of expression. Inevitably, the licensing imperative chosen in Articles 17(1) and 17(4)(a) of the CDSMD culminates in the introduction of filtering tools. As copyright holders and collecting societies are unlikely to offer all-embracing umbrella licenses, OCSSPs must rely on algorithmic tools to ensure that

57. For a detailed analysis of current E.U. rights clearance challenges in the digital environment, see SEBASTIAN FELIX SCHWEMER, LICENSING AND ACCESS TO CONTENT IN THE EUROPEAN UNION: REGULATION BETWEEN COPYRIGHT AND COMPETITION LAW (2019). For a comprehensive overview of the collective rights management situation in Europe, see Oleksandr Bulayenko, Stef Van Gompel, Christian Handke, Roel Peeters, Joost Poort, João Pedro Quintais & David Regeczi, *Study on Emerging Issues on Collective Licensing Practices in the Digital Environment* (2021), https://papers.ssrn.com/abstract=3970490.

58. *See* Bulayenko et al., *supra* note 57.

59. Guidance Art. 17 CDSMD, *supra* note 18, at n 18, 6. ("Collective licensing . . . is little used in the film sector where direct licensing by film producers is more usual.").

content uploads do not overstep the limits of the use permissions they managed to obtain.[60] In *Poland v. Parliament and Council*, the CJEU explicitly confirmed that Article 17(4)(b) of the CDSMD mandates UGC platforms to carry out a preventive review of user uploads in circumstances where rightholders have provided "relevant and necessary information"[61] for the detection of protected works.[62] Depending on the scale of the task, the review of user uploads requires the employment of automatic recognition and filtering tools. The court noted that no party to the *Poland* case had been able to designate possible alternatives to automated filtering tools. Therefore, at least for the largest platforms (e.g., YouTube, Facebook, and Instagram), automated content filtering is necessary to comply with the best efforts filtering obligations arising from Article 17(4)(b) of the CDSMD.[63]

From the perspectives of freedom of expression and freedom of information, the amalgam of licensing and filtering obligations in Article 17(4) of the CDSMD is highly problematic.[64] In accordance with the contours of the licensing deals which UGC platforms managed to obtain, algorithmic enforcement measures will curtail the freedom of users to participate actively in the creation of online content. The fundamental rights tension caused by this regulatory approach is evident. In decisions rendered prior to the adoption of Article 17, the CJEU has stated explicitly that in transposing E.U. directives and implementing transposing measures:

> Member States must . . . take care to rely on an interpretation of the directives which allows a fair balance to be struck between the various fundamental rights protected by the Community legal order.[65]

Interestingly, the application of filtering technology to a social media platform already occupied centre stage in *Sabam v. Netlog*. The case concerned Netlog's social networking platform, which offered every subscriber the opportunity to acquire a globally available "profile" space that could be filled with photos, texts, video clips etc.[66] Claiming that users made unauthorized use of music and films belonging to its repertoire, the collecting society Sabam sought to obtain an injunction obliging Netlog to install a system for filtering the

60. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council.
61. *Id.* ¶ 53.
62. *Id.*
63. *Id.*
64. For a more candid statement, see Senftleben, *supra* note 33, at 339–40.
65. CJEU, case C-275/06, Productores de Música de España (Promusicae) v. Telefónica de España SAU, ¶ 68.
66. CJEU, 16 February 2012, case C-360/10, Sabam v. Netlog, ¶¶ 16–18.

information uploaded to Netlog's servers. As a preventive measure and at Netlog's expense, this system would apply indiscriminately to all users for an unlimited period and would be capable of identifying electronic files containing music and films from the Sabam repertoire. In case of a match, the system would prevent relevant files from being made available to the public.[67] The *Sabam v. Netlog* case offered the CJEU the chance to provide guidance on a filtering system such as those that are envisaged in Article 17(4)(b) of the CDSMD.[68]

However, in *Sabam v. Netlog*, the CJEU did not arrive at the conclusion that such a filtering system was permissible. Instead, the court saw a serious infringement of fundamental rights. The court took the explicit recognition of intellectual property as a fundamental right in Article 17(2) of the CFR as a starting point. At the same time, the court recognized that intellectual property must be balanced against the protection of other fundamental rights and freedoms.[69] Weighing the right to intellectual property asserted by Sabam against Netlog's freedom to conduct a business,[70] the court observed that the filtering system would involve monitoring all or most of the information on Netlog's servers in the interests of copyright holders, would have no limitation in time, would be directed at all future infringements, and would be intended to protect not only existing but also future works.[71] Against this background, the CJEU concluded that the filtering system would encroach upon Netlog's freedom to conduct a business.[72]

The CJEU also found that the filtering system would violate the fundamental rights of Netlog's users. These fundamental rights included their right to the protection of their personal data and their freedom to receive or impart information, as safeguarded by Articles 8 and 11 of the CFR respectively.[73] The court recalled that the use of protected material in online communications may be lawful under statutory limitations of copyright in the Member States, and that some works may have already entered the public domain, or been made available for free by the authors concerned.[74] Filtering systems, however, may block communications using these lawful resources. Given this corrosive effect, the court concluded:

67. *Id.* ¶¶ 26, 36–37.

68. As to the different levels of content monitoring that can be derived from CJEU jurisprudence, see SENFTLEBEN & ANGELOPOULOS, *supra* note 7, at 7–16.

69. CJEU, 16 February 2012, case C-360/10, Sabam v. Netlog, ¶¶ 41–44.

70. EU Charter of Fundamental Rights art. 16 - Freedom to conduct a business.

71. CJEU, 16 February 2012, case C-360/10, Sabam v. Netlog, ¶ 45.

72. *Id.* ¶¶ 46–47.

73. *Id.* ¶¶ 48–50.

74. *Id.* ¶¶ 50.

> Consequently, it must be held that, in adopting the injunction requiring the hosting service provider to install the contested filtering system, the national court concerned would not be respecting the requirement that a fair balance be struck between the right to intellectual property, on the one hand, and the freedom to conduct business, the right to protection of personal data and the freedom to receive or impart information, on the other (see, by analogy, *Scarlet Extended*, paragraph 53).[75]

This case law confirms that the filtering obligations arising from Article 17(4) of the CDSMD are highly problematic. As a way out of the dilemma, the E.U. legislature walks the fine line of distinguishing between monitoring all UGC in search of a whole repertoire of works,[76] and monitoring all UGC in search of specific, pre-identified works.[77] *Sabam v. Netlog* concerned a filtering obligation targeting all types of UGC containing traces of works falling under the Sabam rights portfolio.[78] It seems that the drafters of Article 17(4)(b) of the CDSMD tried to avoid this prohibited general monitoring obligation, and thus escape the verdict of a violation of fundamental rights, by establishing the obligation to filter "specific works and other subject matter for which the rightholders have provided the service providers with the relevant and necessary information."[79]

2. Reliance on Industry Cooperation to Safeguard Fundamental Rights

At this point, the above-described element of industry "cooperation" enters the picture. Rightly understood, the content filtering system established in Article 17(4)(b) of the CDSMD relies on a joint effort of the creative industry and the online platform industry. To set the filtering machinery in motion, copyright holders must first provide OCSSPs with "relevant and necessary information"[80] with regard to those works which they want to ban from user uploads. The Commission Guidance states that information can be deemed "relevant" if it is, at a minimum, "accurate about the rights ownership of the particular work or subject matter in question."[81] Determining whether said information is "necessary" will depend on the technical measures

75. *Id.* ¶ 51.

76. *Id.* ¶¶ 26, 36–37.

77. *Cf.* SENFTLEBEN & ANGELOPOULOS, *supra* note 7, at 8–9.

78. CJEU, 16 February 2012, case C-360/10, Sabam v. Netlog, ¶ 26.

79. CDSMD art. 17(4)(b), 2019 O.J. (L 130). The intention to obviate the impression of a prohibited general monitoring obligation also lies at the core of Article 17(8) CDSMD. This provision declares that UGC licensing and filtering "shall not lead to any general monitoring obligation."

80. *Id.* art. 17(4)(b).

81. Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 14.

employed by platforms: the information provided by rightholders must enable the effective implementation of the platforms' solutions.[82]

Once relevant and necessary information on protected works is received, the OCSSP is obliged to include the information in the content moderation process and ensure the filtering—"unavailability"[83]—of content uploads that contain traces of the protected works. According to Article 17(7) of the CDSMD, it is this cooperation which must not result in the prevention of non-infringing UGC, including situations where UGC is covered by a copyright limitation. This cooperation is also the central item on the agenda of stakeholder dialogues which the Commission is expected to organize under Article 17(10) of the CDSMD to identify best practices.[84]

The fundamental problem of the whole cooperation concept, however, is the fact that, unlike public bodies and the judiciary, the central players in the cooperation scheme—the creative industry and the online platform industry—are private entities that are not intrinsically motivated to safeguard the public interest in the exercise and furtherance of fundamental rights and freedoms. Despite all invocations of diligence and proportionality[85]—"high industry standards of professional diligence" in Article 17(4)(b) of the CDSMD; "diligent, objective and proportionate" application in Article 14(4) of the DSA—the decision-making in the context of content filtering is likely more simple. Namely, the moment the balancing of competing human rights positions is left to industry cooperation, economic cost and efficiency considerations are likely to occupy center stage. Arguably, these considerations will often prevail over more abstract societal objectives, such as flourishing freedom of expression and information.

A closer look at the different stages of industry cooperation resulting from the regulatory model of Article 17 of the CDSMD confirms that concerns about human rights deficits are not unfounded. As explained, the first step in the content moderation process is the notification of relevant and necessary information relating to "specific works and other subject matter"[86] by copyright holders. In the light of case law precedents, in particular *Sabam v.*

82. Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 14 (providing examples related with "fingerprinting" and "metadata-based solutions").

83. CDSMD art. 17(4)(b), 2019 O.J. (L 130).

84. *Id.* art. 17(10) (stating that the Commission will issue guidance on the application of Article 17, "in particular regarding the cooperation referred to in paragraph 4").

85. *See* the references to "high industry standards of professional diligence" in CDSMD art. 17(4)(b), 2019 O.J. (L 130); "diligent, objective and proportionate" application in Article 14(4) DSA.

86. CDSMD art. 17(4)(b), 2019 O.J. (L 130).

Netlog,[87] use of the word "specific" can be understood to reflect the legislator's hope that copyright holders will only notify individually selected works. For instance, copyright holders could limit use of the notification system to those works that constitute cornerstones of their current exploitation strategy. The principle of proportionality and high standards of professional diligence also point in the direction of a cautious approach that confines work notifications to those repertoire elements that are "specific" in the sense that they generate a copyright holder's lion's share of revenue.[88] In line with this approach, other elements of the work catalogue could be kept available for creative remix activities of users. This, in turn, would reduce the risk of overbroad inroads into freedom of expression and information.

In practice, however, rightholders are highly unlikely to adopt this cautious approach. The legal basis for requiring a focus on individually selected works lies in the legislator's use of the expression "best efforts to ensure the unavailability of *specific* works and other subject matter"[89] in Article 17(4)(b) of the CDSMD. Proportionality and diligence considerations only form the broader context in which this specificity requirement is embedded. Strictly speaking, however, the reference to "best efforts to ensure the unavailability"[90] shows that the requirement of "high industry standards of professional diligence"[91] concerns the filtering step taken by a platform to ensure the unavailability of notified works, not the primary notification sent by copyright holders.

Just like the requirement of "high industry standards of professional diligence," the imperative of "diligent, objective and proportionate" application and enforcement of content restrictions in Article 14(4) of the DSA relates to platform content moderation measures that restrict user freedoms, not the rightholder notification system that sets the filtering process in motion. The success of the risk-reduction strategy surrounding the word "specific" in Article 17(4)(b) of the CDSMD and the words "diligent, objective and proportionate" in Article 14(4) of the DSA is thus doubtful. In the cooperation with OCSSPs, nothing seems to prevent the creative industry

87. CJEU, 16 February 2012, case C-360/10, Sabam v. Netlog, ¶ 51.

88. *See* Martin Senftleben, *How to Overcome the Normal Exploitation Obstacle: Opt-Out Formalities, Embargo Periods, and the International Three-Step Test*, 1 BERKELEY TECH. L.J. 1 (2014); *see also* Christophe Geiger, Daniel J. Gervais & Martin Senftleben, *The Three-Step-Test Revisited: How to Use the Test's Flexibility in National Copyright Law*, 29 AM. U. INT'L L. REV. 581 (2014); Daniel Gervais, *Towards a New Core International Copyright Norm: The Reverse Three-Step Test*, 9 MARQ. INTELL. PROP. L. REV. 1 (2005).

89. CDSMD art. 17(4)(b), 2019 O.J. (L 130) (emphasis added).

90. *Id.*

91. *Id.*

from sending copyright notifications that cover every element of long and impressive work catalogues. Platforms may thus receive long lists of all works which copyright holders have in their repertoire. Adding up all "specific works and other subject matter" included in these notifications, it could well be that Article 17(4)(b) of the CDSMD culminates in a filtering obligation that is very similar to the filtering measures which the CJEU prohibited in *Sabam v. Netlog*.[92] The risk of encroachments upon human rights is evident.

3. *Diligence and Proportionality Viewed Through the Prism of Cost and Efficiency Considerations*

Turning to the second step in the content moderation process—the act of filtering carried out by OCSSPs to prevent the availability of notified works on UGC platforms—it is noteworthy that proportionality and diligence obligations are directly applicable. As explained, the requirements of "high industry standards of professional diligence"[93] and "diligent, objective and proportionate"[94] application only form the broader context surrounding the notification of specific works by rightholders. When it comes to the content moderation process as such, however, these rules impact the activities of OCSSPs directly: the UGC filtering process must be implemented in a way that complies with these diligence and proportionality requirements.

The Commission Guidance clarifies in this respect that compliance with "high industry standards of professional diligence" must be evaluated against "available industry practices on the market,"[95] including technological solutions. Platforms have discretion only in selecting from existing solutions on the market.[96] In discussing prevailing market practices, the Guidance highlights content recognition based on fingerprinting[97] as the primary example, whilst acknowledging that this is not the norm for smaller

92. Senftleben, *supra* note 17, at 483–84.
93. Article 17(4)(b) CDSMD.
94. Article 14(4) DSA.
95. Guidance Art. 17 CDSMD, *supra* note 18, at n. 21, 12.
96. *Id.*
97. A fingerprint is a digital representation of media content, and contains all visual and/or audio information of the content. For a technical description, see, e.g., EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE, AUTOMATED CONTENT RECOGNITION: DISCUSSION PAPER. PHASE 1, EXISTING TECHNOLOGIES AND THEIR IMPACT ON IP (2020), https://data.europa.eu/doi/10.2814/52085 (last visited Sept. 4, 2021); JEAN-PHILIPPE MOCHON & SYLVAIN HUMBERT, CSPLA, A mission on the tools for the recognition of content protected by online sharing platforms: state of the art and proposals (2020), https://www.culture.gouv.fr/en/Sites-thematiques/Propriete-litteraire-et-artistique/Conseil-superieur-de-la-propriete-litteraire-et-artistique/Travaux/Missions/Mission-du-CSPLA-sur-les-outils-de-reconnaissance-des-contenus-proteges-par-les-plateformes-de-partage-en-ligne-etat-de-l-art-et-propositions.

platforms.[98] Other technologies include hashing, watermarking, the use of metadata, and keyword search.[99] These solutions may be developed in-house, as in the case of YouTube's Content ID and Meta's Rights Manager. OCSSPs may also procure them from third-party providers, such as Audible Magic or Pex.

As to the practical outcome of UGC filtering in the light of these diligence and proportionality requirements, however, it is to be recalled that OCSSPs will likely align the concrete implementation of content moderation systems with cost and efficiency considerations. Abstract commandments, such as the instruction to act in accordance with "high standards of professional diligence"[100] and in a "proportionate manner in applying and enforcing [UGC upload] restrictions"[101] can hardly be deemed capable of superseding concrete commercial cost and efficiency necessities. Tuomas Mylly accurately characterizes litanies of diligence and proportionality requirements as "wish-lists for private regulators."[102] On its merits, the legislature whitewashes statutory content filtering obligations by adding a diligence and proportionality gloss to reassure itself that the drastic measure will be implemented with sufficient care and caution to avoid the erosion of human rights. The success of this ingredient of the outsourcing recipe is doubtful. In reality, the subordination of industry decisions to diligence and proportionality imperatives—the acceptance of more costs and less profits to reduce the corrosive effect on freedom of expression and information—would come as a surprise. Instead, OCSSPs can be expected to be rational in the sense that they seek to achieve content filtering at minimal costs.[103]

Hence, there is no guarantee that industry cooperation in the field of UGC will lead to the adoption of the most sophisticated filtering systems with the highest potential to avoid unjustified removals of content mash-ups and remixes. A test of proportionality is unlikely to occupy centre stage unless the least intrusive measure also constitutes the least costly measure. A test of professional diligence is unlikely to lead to the adoption of a more costly and less intrusive content moderation system unless additional revenues accruing from enhanced popularity among users offsets the extra financial investment.

In addition, the E.U. legislation sends mixed signals. Article 17(5) of the CDSMD provides guidelines for the assessment of the proportionality of

98. Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 12.
99. *Id.* at n. 18, 12–13.
100. CDSMD art. 17(4)(b), 2019 O.J. (L 130).
101. Digital Services Act art. 14(4), 2022 O.J. (L 277).
102. Mylly, *supra* note 41, at 71.
103. Senftleben, *supra* note 17, at 484.

filtering obligations. The relevant factors listed in the provision, however, focus on "the type, the audience and the size of the service," "the type of works or other subject matter," and "the availability of suitable and effective means and their cost for service providers."[104] Hence, cost and efficiency factors have made their way into the proportionality assessment scheme. Paradoxically, it is conceivable that these factors encourage the adoption of cheap and unsophisticated filtering tools that lead to excessive content blocking. An assessment of liability risks also confirms that excessive filtering risks must be taken seriously. A UGC platform seeking to minimize the risk of liability is likely to succumb to the temptation of overblocking.[105] Filtering more than necessary is less risky than filtering only clear-cut cases of infringement. After all, the described primary, direct liability for infringing user uploads which follows from Article 17(1) of the CDSMD is hanging above the head of OCSSPs like the sword of Damocles.

The second step of the industry cooperation concept underlying Article 17 of the CDSMD is therefore at least as problematic as comprehensive notifications of entire work catalogues. The OCSSP obligation to embark on content filtering to police the borders of use permissions and prevent content availability in the absence of licenses raises serious concerns about interferences with human rights, particularly the freedom of expression and information.

4. Considerable Risk of Encroachments Upon Fundamental Rights

Surveying the described human rights risks that arise from the industry cooperation scheme in Article 17 of the CDSMD, the conclusion is inescapable that, despite all invocations of diligence and proportionality as mitigating factors, the outsourcing strategy underlying the E.U. regulation of content moderation in the CDSM Directive and the DSA is highly

104. CDSMD art. 17(5), 2019 O.J. (L 130).

105. *See* Maayan Perel (Filmar) & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STANFORD TECH. L. REV. 473, 490–91 (2016). For empirical studies pointing towards overblocking, see Sharon Bar-Ziv & Niva Elkin-Koren, *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown*, 50 CONN. L. REV. 37 (2017) ("Overall, the N&TD regime has become fertile ground for illegitimate censorship and removal of potentially legitimate materials."); Jennifer M. Urban, Joe Karaganis & Brianna Schofield, *Notice and Takedown: Online Service Provider And Rightsholder Accounts Of Everyday Practicenotice and Takedown In Everyday Practice*, 64 J. COPYRIGHT SOC'Y 371, 372 (2017) ("About 30% of takedown requests were potentially problematic. In one in twenty-five cases, targeted content did not match the identified infringed work, suggesting that 4.5 million requests in the entire six-month data set were fundamentally flawed. Another 19% of the requests raised questions about whether they had sufficiently identified the allegedly infringed work or the allegedly infringing material").

problematic. Instead of safeguarding human rights, the regulatory approach is likely to culminate in human rights violations. Against this background, it is important to analyse mechanisms that could bring human rights deficits to light and remedy shortcomings. Complaint and redress mechanisms for users may play an important role in this respect. We turn to these tools in the following section.

B. CONCEALING HUMAN RIGHTS DEFICITS CAUSED BY RELIANCE ON INDUSTRY COOPERATION

As explained above, UGC platforms are obliged by the DSA to make information on content moderation "policies, procedures, measures and tools" available to users.[106] This must be done in "clear, plain, intelligible, user-friendly and unambiguous language."[107] Moreover, the information must be publicly available in an easily accessible and machine-readable format.[108] These information and transparency obligations can be regarded as exponents of a broader human rights preservation strategy.[109] The broader pattern comes to the fore when the information flow generated in Article 14(1) of the DSA is placed in the context of the complaint and redress mechanism for unjustified content filtering that forms a building block of Article 17 of the CDSMD. Article 17(9) of the CDSMD requires that OCSSPs put in place:

> an effective and expeditious complaint and redress mechanism that is available to users of their services in the event of disputes over the disabling of access to, or the removal of, works or other subject matter uploaded by them.[110]

To connect the dots between Article 14(1) of the DSA and Article 17(9) of the CDSMD, it is particularly important to recognize that the OCSSP liability regime established in Article 17 of the CDSMD constitutes a specific subsystem of platform regulation which *complements* the platform regimes in the DSA. According to Article 2(4)(b) of the DSA, the DSA rules are without prejudice to the rules laid down by "Union law on copyright and related rights."[111] The Explanatory Memorandum accompanying the initial DSA

106. Digital Services Act art. 14(1), 2022 O.J. (L 277). Further information and transparency obligations are listed elsewhere in Article 14, namely in paras (2), (3), (5), and (6).

107. *Id.* art. 14(1).

108. *Id.*

109. Examples can be found in the GDPR and Terrorist Content Regulation.

110. CDSMD art. 17(9), 2019 O.J. (L 130).

111. Digital Services Act art. 2(4)(b), 2022 O.J. (L 277). For an extensive analysis on this topic, see Alexander Peukert, Martin Husovec, Martin Kretschmer, Péter Mezei & João Pedro Quintais, *European Copyright Society – Comment on Copyright and the Digital Services Act Proposal*, 53 IIC-INTERNATIONAL REV. OF INTELL. PROP. & COMPETITION L. 358 (2022); João Pedro

Proposal explained the interplay between the DSA and more specific regimes, such as E.U. copyright law, as follows:

> The proposed Regulation complements existing sector-specific legislation and does not affect the application of existing EU laws regulating certain aspects of the provision of information society services, which apply as *lex specialis*.[112]

In this vein, Recital 11 of the DSA states that the OCSSP liability regime in Article 17 of the CDSMD establishes specific rules and procedures that should remain unaffected by DSA rules. Insofar as the CDSM Directive does not contain specific rules, however, the DSA rules are fully applicable. The two sets of legislation—the CDSMD and the DSA—thus complement each other.[113]

Regarding the role of users in the human rights arena, this complementary character yields important insights: the legislature has confidently left the identification and correction of excessive content blocking to users. A relatively low number of user complaints, however, may be misinterpreted as an indication that content filtering hardly ever encroaches upon freedom of expression and information even though limited user activism may be due to overly slow and cumbersome procedures (as discussed in Section II.B.1). Instead of addressing this problematic concealment mechanism, the CJEU has confirmed the validity of the content moderation rules laid down in Article 17 of the CDSMD.[114] The court even qualified elements of the problematic outsourcing and concealment strategy as valid safeguards against the erosion of freedom of expression and information.[115] Instead of uncovering human rights risks, the court preferred to condone and stabilize the system (as discussed in Section II.B.2). Under these circumstances, only legislative countermeasures taken by E.U. Member States (as discussed in Section II.B.3) and content moderation assessments in audit reports (as discussed in Section II.B.4) give some hope that violations of human rights may finally be prevented despite the corrosive outsourcing and concealment scheme

Quintais & Sebastian Felix Schwemer, *The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?*, 13 EUR. J. RISK REGULATION 191 (2022).

112. European Commission, 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, Explanatory Memorandum, 4.

113. Quintais & Schwemer, *supra* note 111. *See infra* Section III.A contrasting the legal regimes applicable to OCSSPs and non-OCSSPs.

114. *See infra* Section II.B.2.

115. *Id.*

underlying the regulation of content moderation in the European Union (as discussed in Section III.B.5).

1. Reliance on User Complaints as Part of a Concealment Strategy

Article 17(9) of the CDSMD and Article 14(1) of the DSA both identify users as the primary addressees of information about content moderation systems and potential countermeasures.[116] This regulatory model is not new. In *UPC Telekabel Wien*, the CJEU sought to ensure that, in the case of website blocking measures, the national courts in E.U. Member States would be able to carry out a judicial review. This, however, was only conceivable if a challenge was brought against the blocking measure implemented by an internet service provider:

> Accordingly, in order to prevent the fundamental rights recognised by EU law from precluding the adoption of an injunction such as that at issue in the main proceedings, the national procedural rules must provide a possibility for internet users to assert their rights before the court once the implementing measures taken by the internet service provider are known.[117]

Therefore, the rights assertion option for users served the ultimate purpose of paving the way for judicial review. In Article 17(9) of the CDSMD, this pattern reappears. Users can avail themselves of the option to instigate complaint and redress procedures at platform level and, ultimately, go to court.[118] The DSA also contains specific user complaint and redress rights. Complementing Article 17(9) of the CDSMD,[119] Article 20 of the DSA sets forth detailed rules for internal complaint handling on online platforms. Article 54 of the DSA confirms with regard to DSA obligations that users are entitled to compensation for any damage or loss they suffered due to an infringement of DSA obligations. As noted, one of these obligations follows from Article 14(4) of the DSA. This provision obliges platforms to apply content moderation measures in a proportionate manner—with due regard to freedom of expression and information. In addition, Article 86(1) of the DSA affords users the opportunity to mandate a non-profit body, organization, or

116. Regarding Article 14(1) DSA, see *supra* Section II.A.

117. CJEU, 27 March 2014, case C-314/12, UPC Telekabel Wien, ¶ 57.

118. CDSMD art. 17(9), 2019 O.J. (L 130). ("[W]ithout prejudice to the rights of users to have recourse to efficient judicial remedies. In particular, Member States shall ensure that users have access to a court or another relevant judicial authority to assert the use of an exception or limitation to copyright and related rights.").

119. As to the complementary character of Article 20 DSA, see Article 2(4)(b) and Recital 11 DSA. For an extensive analysis of the combined application of CDSMD and DSA provisions, see Quintais & Schwemer, *supra* note 111; Peukert et al., *supra* note 111.

association to exercise their complaint, redress, and compensation rights on their behalf.[120]

However, the broad reliance placed on user activism is surprising. Evidence from the application of the DMCA counter-notice system in the United States[121] shows clearly that users are unlikely to file complaints in the first place.[122] This is confirmed by data from recent transparency reports from the largest platforms.[123] If users must wait relatively long for a result, it is foreseeable that a complaint-and-redress mechanism that depends on user initiatives is incapable of safeguarding freedom of expression and information. Moreover, an overly cumbersome complaint-and-redress mechanism may thwart user initiatives from the outset. While it cannot be ruled out that some users will exhaust the full arsenal of complaint, redress, and compensation options, it seems unrealistic to assume that user-complaint mechanisms have the potential of revealing the full spectrum and impact of free expression restrictions that result from automated content moderation systems. User complaints are unlikely to provide a complete picture.

In the context of UGC, it must also be considered that it is often crucial to react quickly to current news and film, book, and music releases. If the complaint and redress mechanism finally yields the insight that a lawful content remix or mash-up had been unjustifiably blocked, the window of relevance for the affected quotation or parody may already have passed.[124] From this perspective, the elastic timeframe for complaint handling—"shall be processed

120. The provision requires that according to their statutes, these non-profit institutions must have a legitimate interest in safeguarding DSA rights and obligations.

121. As to this feature of the notice-and-takedown system in U.S. copyright law, see Peguera, *supra* note 9, at 481.

122. *See* Jennifer M. Urban & Laura Quilter, *Efficient Process or "Chilling Effects"? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA COMPUT. & HIGH TECH. L.J. 621 (2006) (showing that 30% of DMCA takedown notices were legally dubious, and that 57% of DMCA notices were filed against competitors). While the DMCA offers the opportunity to file counter-notices and rebut unjustified takedown requests, Urban and Quilter find that instances in which this mechanism is used are relatively rare. *Cf.* the critical comments on the methodology used for the study and a potential self-selection bias arising from the way in which the analyzed notices have been collected by Frederick W. Mostert and Martin B. Schwimmer, *Notice and Takedown for Trademarks*, 101 TRADEMARK REP. 249, 259–60 (2011).

123. *See infra* Section III.B.1.

124. Apart from the time aspect, complaint systems may also be implemented in a way that discourages widespread use. *Cf.* Perel & Elkin-Koren, *supra* note 105, at 507–8, 514. In addition, the question arises whether users filing complaints are exposed to copyright infringement claims in case the user-generated quotation, parody or pastiche at issue (which the user believes to be legitimate) finally proves to amount to copyright infringement. *Cf.* Niva Elkin-Koren, *Fair Use by Design*, 64 UCLA L. REV. 1092 (2017).

without undue delay"[125]—gives rise to concerns. This standard differs markedly from an obligation to let blocked content reappear promptly. As Article 17(9) of the CDSMD also requires human review, it could delay a final decision on the infringing nature of content. Considering these features, the complaint-and-redress option may appear unattractive to users.[126]

Instead of adequately addressing concerns about human rights deficits, reliance on user complaints, thus, adds another risk factor. The complaint-and-redress mechanism may allow authorities to hide behind a lack of user activism, even if this is caused by the cumbersome or slow nature of the process. Relatively few user complaints may be misinterpreted as evidence that no overblocking occurs, keeping human rights deficits under the radar. The oversimplified equation "no user complaint = no human rights problem" offers the opportunity to dress up an overly restrictive content moderation system as a success, and to disguise encroachments upon freedom of expression and information.

The outsourcing problem described in the preceding section—inappropriate reliance on OCSSPs and copyright holders as human rights guardians—is thus aggravated by overreliance on complaint and redress mechanisms that users are unlikely to embrace in the first place. By leaving the responsibility to safeguard freedom of expression to users, the legislator cultivates a culture of "concealing" human rights deficits. Even if users lodge a complaint, it must be considered that any redress remains an ex post measure. That is to say, a remedy that reinstates freedom of expression and information only after harm is done, namely harm in the form of unjustified content blocking and UGC impoverishment. The E.U. approach is thus deficient for at least two reasons: the outsourcing of the obligation to safeguard human rights to online platforms and the reliance on user activism to bring human rights violations to light.

2. *Confirmation of the Outsourcing and Concealment Strategy in CJEU Jurisprudence*

This outcome of the risk assessment raises the additional question whether other institutions in the platform governance arena could fulfil the role of human rights guardians more reliably. The judiciary seems the logical candidate. Interestingly, the CJEU already had the opportunity to discuss violations of freedom of expression and information that may arise from content moderation under Article 17 of the CDSMD. In *Poland v. Parliament and Council*, the Republic of Poland had brought an annulment action arguing

125. CDSMD art. 17(9), 2019 O.J. (L 130).
126. *Cf.* Senftleben, *supra* note 17, at 484.

that OCSSPs were bound under Articles 17(4)(b) and (c) of the CDSMD to carry out preventive—ex ante—monitoring of all user uploads.[127] To fulfil this Herculean task, they had to employ automatic filtering tools. In Poland's view, E.U. legislation imposed this preventive monitoring obligation on OCSSPs "without providing safeguards to ensure that the right to freedom of expression and information is respected."[128] The contested provisions thus constituted a limitation on the exercise of the fundamental right to freedom of expression and information, which respected neither the essence of that right nor the principle of proportionality. Hence, the filtering obligations arising from Article 17 of the CDSMD could not be regarded as justified under Article 52(1) of the CFR.[129]

Discussing these annulment arguments, the CJEU pointed out that prior review and filtering of user uploads creates the risk of limiting a central avenue for the online dissemination of UGC. The filtering regime in Articles 17(4)(b) and (c) of the CDSMD imposes a restriction on the ability of users to exercise their right to freedom of expression and information which is guaranteed by Article 11 of the CFR and Article 10 of the European Convention on Human Rights (ECHR).[130] However, the court considered that such a limitation meets the requirements set forth in Article 52(1) of the CFR—mandating that any limitation on the exercise of the right to freedom of expression and information is legally established and preserves the essence of those freedoms.[131] The court was satisfied that the limitation arising from the filtering obligations in Article 17 of the CDSMD can be deemed justified in the light of the legitimate objective to ensure a high level of copyright protection to safeguard the right to intellectual property enshrined in Article 17(2) of the CFR.[132]

More specifically, the court identified no less than six freedom of expression safeguards in the regulatory design of Article 17 of the CDSMD which, in the court's view, give sufficient reassurance that freedom of expression and information will not be unduly curtailed. A key aspect in this assessment is the first point. The court assumed that the introduction of automated content filtering tools would not prevent users from uploading

127. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 24. For a more detailed discussion of the decision, see Martin Husovec, *Mandatory Filtering Does Not Always Violate Freedom of Expression: Important Lessons from Poland V. Council and European Parliament*, 60 COMMON MKT. L. REV. 173 (2023).

128. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 24.

129. *Id.* ¶ 24.

130. *Id.* ¶¶ 55, 58, 82.

131. *Id.* ¶ 63 (referring to the principle of proportionality).

132. *Id.* ¶ 69.

lawful content, including UGC containing traces of protected third-party material that was permissible under statutory exceptions to copyright.[133] In this context, the court recalled its earlier ruling in *Sabam v. Netlog* from which it followed that:

> a filtering system which might not distinguish adequately between unlawful content and lawful content, with the result that its introduction could lead to the blocking of lawful communications, would be incompatible with the right to freedom of expression and information, guaranteed in Article 11 of the Charter, and would not respect the fair balance between that right and the right to intellectual property.[134]

Hence, the court was confident that, in the light of its case law, OCSSPs would refrain from introducing content filtering measures unless these systems could reliably distinguish between lawful parody and infringing verbatim copying; in other words, unless they could leave lawful uploads unaffected.[135]

The court's second point addresses statutory exceptions to copyright more directly. In line with earlier decisions, the CJEU confirmed that copyright limitations supporting freedom of expression, such as the right of quotation and the exemption of parody, constitute "user rights."[136] To avoid the dismantling of these free expression strongholds, E.U. Member States have to ensure that automated filtering measures do not deprive users of their freedom to upload content created for the purposes of quotation, criticism, review, caricature, parody, or pastiche.[137] On this point the judgment endorsed, by

133. *Id.* ¶ 86.

134. *Id.* ¶ 86. *Cf.* CJEU, 16 February 2012, case C-360/10, Sabam v, Netlog, ¶¶ 50–51.

135. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 86.

136. *Id.* ¶¶ 87–88; CJEU, 29 July 2019, case C-516/17, Spiegel Online, ¶¶ 50–54; CJEU, 29 July 2019, case C‑469/17, Funke Medien NRW, ¶ 65–70; *see* Christophe Geiger & Elena Izyumenko, *The Constitutionalization of Intellectual Property Law in the EU and the Funke Medien, Pelham and Spiegel Online Decisions of the CJEU: Progress, but Still Some Way to Go!*, 51 IIC – INT'L REV. INTELL. PROP. & COMPETITION L. 282, 292–98 (2020); TANYA APLIN & LIONEL BENTLY, GLOBAL MANDATORY FAIR USE: THE NATURE AND SCOPE OF THE RIGHT TO QUOTE COPYRIGHT WORKS 75–84 (2020). For a recent discussion, see also Tito Rendas, *Are Copyright-Permitted Uses 'Exceptions', 'Limitations' or 'User Rights'? the Special Case of Article 17 CDSM Directive*, 17 J. INTELL. PROP. L. & PRAC. 54 (2022).

137. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 87. With regard to the particular importance of the inclusion of the open-ended concept of "pastiche," see Martin Senftleben, *Institutionalized Algorithmic Enforcement—The Pros and Cons of the EU Approach to UGC Platform Liability*, 14 FIU L. REV. 299, 320–27 (2020); JOÃO PEDRO QUINTAIS, COPYRIGHT IN THE AGE OF ONLINE ACCESS: ALTERNATIVE COMPENSATION SYSTEMS IN EU LAW 235–37 (2017); Senftleben, *supra* note 8, at 145–62; Emily Hudson, *The Pastiche Exception in Copyright Law: A Case of Mashed-Up Drafting?*, 4 INTELL. PROP. Q. 346, 348–

reference, the Advocate General Opinion stating that filters "must not have the objective or the effect of preventing such legitimate uses," and that providers must "consider the collateral effect of the filtering measures they implement," as well as "take into account, *ex ante*, respect for users' rights."[138]

As a third aspect that mitigates the corrosive effect of Articles 17(4)(b) and (c) of the CDSMD on freedom of expression and information, the court pointed out that the filtering machinery was only set in motion on the condition that rightholders provide platforms with the "relevant and necessary information"[139] concerning protected works that should not become available on the UGC platform. In the absence of such information, OCSSPs would not be led to make content unavailable.[140]

The fourth point highlighted by the court was the clarification in Article 17(8) of the CDSMD that no general monitoring obligation was intended.[141] The fifth point was the complaint-and-redress mechanism allowing users to bring unjustified content blocking to the attention of the platform provider.[142] Finally, the court recalled that Article 17(10) of the CDSMD tasks the European Commission with organizing stakeholder dialogues to ensure a uniform mode of OCSSP/rightholder-cooperation across Member States and establish best filtering practices in the light of industry standards of professional diligence.[143]

Qualifying all six aspects as valid safeguards against an erosion of freedom of expression and information, the court concluded that the design of Article 17 of the CDSMD includes appropriate countermeasures to survive the

52, 362–64; Florian Pötzlberger, *Pastiche 2.0: Remixing im Lichte des Unionsrechts*, GEWERBLICHER RECHTSSCHUTZ & URHEBERRECHT 675, 681 (2018).

138. Opinion of Advocate General Saugmandsgaard Øe, 15 July 2021, case C-401/19, Poland v. Parliament and Council, ¶ 193.

139. CDSMD art. 17(4)(b), 2019 O.J. (L 130).

140. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 89.

141. *Id.* ¶ 90; *see* CDSMD art. 17(8), 2019 O.J. (L 130); Digital Services Act recital 30, Art. 8, 2022 O.J. (L 277). *Cf.* SENFTLEBEN & ANGELOPOULOS, *supra* note 7.

142. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 94; *see* CDSMD art. 17(9), 2019 O.J. (L 130).

143. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶¶ 96–97. As to existing best practices guidelines, see Guidance Art. 17 CDSMD, *supra* note 18, at n. 21. For an early analysis of the Guidance Art. 17 CDSM Directive, see João Pedro Quintais, *Between Filters and Fundamental Rights: How the Court of Justice saved Article 17 in C-401/19 - Poland v. Parliament and Council*, VERFASSUNGSBLOG (2022), https://verfassungsblog.de/filters-poland/; Bernd Justin Jutte & Giulia Priora, *On the necessity of filtering online content and its limitations: AG Saugmandsgaard Øe outlines the borders of Article 17 CDSM Directive*, KLUWER COPYRIGHT BLOG (2021), http://copyrightblog.kluweriplaw.com/2021/07/20/on-the-necessity-of-filtering-online-content-and-its-limitations-ag-saugmandsgaard-oe-outlines-the-borders-of-article-17-cdsm-directive/.

annulment action brought by the Republic of Poland.[144] Still, the court cautioned E.U. Member States, as well as their authorities and courts, that when transposing and applying Article 17 of the CDSMD, they have to do so in a fundamental rights-compliant manner.[145]

Undoubtedly, the *Poland* decision is a milestone that contains several important clarifications. In particular, the court stated unequivocally that for an automated content filtering system to be deemed permissible, it must be capable of distinguishing lawful from unlawful content.[146] The court pointed out that OCSSPs cannot be required to prevent the uploading and making available of content which, in order to be found unlawful, requires an independent copyright assessment, including on the scope of statutory exceptions.[147] Hence, it could not be ruled out that, in cases raising complex copyright questions, rightholders can only avoid the availability of unauthorized content by sending a robustly substantiated notification—providing "sufficient information to enable the [OCSSP] to satisfy itself, without a detailed legal examination, that the communication of the content at issue is illegal and that removing that content is compatible with freedom of expression and information."[148] In light of previous case law and the current market and technological reality, the *Poland* decision can be understood to establish that only content that is "obviously" or "manifestly" infringing (and content that is "equivalent" to these evident risk categories) may be subject to content filtering measures with an effect ex ante—in the sense of preventing the appearance on the online platform from the outset.[149]

However, in light of the above-described human rights risks arising from the outsourcing and concealment strategy underlying Article 17 of the CDSMD, the *Poland* ruling is disappointing. A critical assessment of the regulatory scheme is missing. The court did not unmask the human rights risks that, as explained in the preceding section, are inherent in the heavy reliance on industry cooperation. The court also refrained from reflecting on human rights risks that could arise from the ineffectiveness of complaint and redress mechanisms for users. Instead of exposing the outsourcing and concealment strategy and addressing human rights deficits, the court rubberstamped both the broader regulatory design and its individual elements. By singling out no

144. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 98.
145. *Id.* ¶ 99.
146. *Id.* ¶ 86.
147. *Id.* ¶ 90. *Cf.* CJEU, 3 October 2019, case C-18/18, Glawischnig-Piesczek, ¶¶ 41–46.
148. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 92; CJEU, 22 June 2021, YouTube and Cyando, C-682/18 and C-683/18, ¶ 116.
149. Concluding similarly, see QUINTAIS ET AL., *supra* note 25.

less than six aspects of Article 17 of the CDSMD and declaring them valid safeguards against violations of freedom of expression and information, the court readily accepted several aspects of the Article 17 scheme that create the outsourcing and concealment risks discussed above.

This central problem of uncritical rubberstamping in the *Poland* decision clearly comes to the fore when the six free expression safeguards are re-evaluated in the light of the above-described outsourcing and concealment risks. Regarding the necessity of distinguishing between lawful and unlawful content uploads,[150] a platform reality check is sought in vain in the judgment. From a legal-theoretical perspective, the CJEU's assumption that filtering systems must not be applied if they cannot reliably distinguish permitted transformative uses from infringing verbatim copying may be correct. But this view does not account for the lack of incentives for platforms to refrain from the employment of unsophisticated overblocking systems in practice. The court does not even reflect on the fact that, instead of discouraging the use of excessive filtering machines, the direct liability risk evolving from Article 17(1) of the CDSMD provides a strong impulse to implement automated filtering systems, regardless of their capacity to distinguish between lawful and unlawful content.

Overblocking allows platforms to escape direct liability and avoid lengthy and costly lawsuits. The only risk from excessive filtering is that platforms must deal with user complaints which are unlikely to come in large numbers. Practically speaking, the implementation of an underblocking approach to safeguard freedom of expression is unlikely. In its pure universe of legal-theoretical assumptions, the court may assume that content filtering will only occur when automated systems can separate the wheat from the chaff. To whitewash the Article 17 approach based on such unrealistic assumptions, however, creates a human rights risk of its own.

The inclusion of rightholder notifications in the list of effective free expression safeguards, the third safeguard recognized by the court, also creates a human rights risk. As noted above, nothing prevents copyright owners from notifying long lists—entire catalogues—of protected works as reference files. Adding up all repertoire notifications, it seems naïve to assume that the notification mechanism in Article 17(4)(b) will not lead to a filtering volume that is comparable with the general filtering obligation which the court prohibited in *Sabam v. Netlog*.[151] From this perspective, the ban on general filtering obligations in Article 17(8) (the fourth safeguard identified by the

150. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 86.
151. *Id.* ¶ 86; CJEU, 16 February 2012, case C-360/10, Sabam v. Netlog, ¶¶ 50–51.

court) can also be unmasked as mere cosmetics. The fifth safeguard which the court accepted is the complaint-and-redress mechanism that causes the corrosive concealment risk described above. The sixth and final safeguard—stakeholder dialogues seeking to establish best practices—is a toothless tiger. Article 17(10) is silent on measures which the Commission could take to enforce the best practices guidelines following from meetings with stakeholders. It remains unclear why the court is willing to accept this as a valid free expression safeguard.

On balance, the court has missed an important opportunity to reveal and address human rights risks that arise from outsourcing and concealment elements of Article 17 of the CDSMD. As a reference point for its assessment of human rights risks, the court has chosen the most favorable interpretation of Article 17 features. It has assumed that platforms would only employ moderation systems capable of adequately distinguishing between lawful and unlawful content. The court qualified the rightholder obligation to provide information on protected works as a limiting factor that could reduce the impact of the content filtering machinery, etc. Adopting this approach, the court refused to consider the practical reality of industry cooperation. It also overlooked the impact of the overblocking incentive resulting from the risk of direct liability for infringing UGC. As a result, the court has made itself an accomplice in the outsourcing and concealment strategy that puts freedom of expression and information at risk.

3. Member State Legislation Seeking to Safeguard Transformative UGC

The foregoing critique of the six free expression safeguards which the CJEU identified in its *Poland* decision did not address the second point made by the court: the obligation placed on E.U. Member States to ensure that transformative UGC—consisting of quotations, parodies, pastiches, etc.—survives the blocking by automated content filtering systems.[152] The reason for this omission is simple: in contrast to other aspects which the court discussed, this element appears as a valid safety valve that could effectively safeguard freedom of expression and information in practice. This insight, however, does not change the critical assessment of the *Poland* judgment. With regards to outsourcing and concealment, the decision remains a missed opportunity to address and minimize human rights risks.

As to the valid second point in the *Poland* phalanx of free expression safeguards—the obligation to preserve copyright limitations for creative remix

152. *Id.* ¶¶ 87–88.

activities—Article 17(7) of the CDSMD plays a central role.[153] The provision leaves no doubt that E.U. Member States are expected to ensure that automated content filtering does not submerge areas of freedom that support the creation and dissemination of transformative user productions that are uploaded to UGC platforms. The second paragraph of Article 17(7) reads as follows:

> Member States shall ensure that users in each Member State are able to rely on any of the following existing exceptions or limitations when uploading and making available content generated by users on online content-sharing services:
>
> (a) quotation, criticism, review;
>
> (b) use for the purpose of caricature, parody or pastiche.[154]

The formulation "shall not result in the prevention" and "shall ensure that users . . . are able" give copyright limitations for "quotation, criticism, review" and "caricature, parody or pastiche" an elevated status. In Article 5(3)(d) and (k) of the Information Society Directive 2001/29/EC ("InfoSoc Directive"),[155] these use privileges were only listed as limitation prototypes which E.U. Member States are free to introduce (or maintain) at the national level. The adoption of a quotation right[156] and an exemption of caricature, parody, or pastiche[157] remained optional. Article 17(7) of the CDSMD, however, transforms these use privileges into mandatory breathing space for transformative UGC, at least in the specific context of OCSSP content moderation.[158]

---

153. *See* Senftleben, *supra* note 17, at 485–90; P. BERNT HUGENHOLTZ & MARTIN SENFTLEBEN, FAIR USE IN EUROPE: IN SEARCH OF FLEXIBILITIES 29–30 (2011).

154. CDSMD art. 17(7), 2019 O.J. (L 130).

155. 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 (*Official Journal of the European Communities* 2001 L 167, 10).

156. 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 (InfoSoc Directive), 2001 O.J. (L 167), art. 5(3)(d).

157. *Id.* art. 5(3)(k).

158. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 87. *Cf.* João Pedro Quintais, Giancarlo Frosio, Stef van Gompel, P. Bernt Hugenholtz, Martin Husovec, Bernd Justin Jütte & Martin Senftleben, *Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive: Recommendations from European Academics*, 10 J. INTELL. PROP. INFO. TECH. & E-COMMERCE L., 278–279 (2020). As to the influence of freedom of speech guarantees on copyright law in the EU, see CJEU, 1 December 2011, case C-145/10, Painer, ¶ 132; CJEU, 3 September 2014, case C-201/13, Deckmyn, ¶ 26 *see also* CJEU, 29 July 2019, case C-476/17, Pelham, ¶ 32, 37 and 59. *Cf.* MARTIN SENFTLEBEN, THE COPYRIGHT / TRADEMARK INTERFACE: HOW THE EXPANSION OF TRADEMARK

Under Article 17(7) of the CDSMD, E.U. Member States are the guardians of these user rights. This regulatory decision comes as a welcome surprise. In contrast to the prevailing preference for solutions based on outsourcing (passing on human rights responsibilities to private entities) and concealment (relying on user complaints to remedy human rights deficits), Article 17(7) entrusts the Member States with the important task of guaranteeing ("shall ensure") that, despite content filtering on platforms, users can share creations made for the purposes of "quotation, criticism, review" and "caricature, parody or pastiche."

In this regard, the *Poland* decision adds an important nuance. Namely, the CJEU qualified the complaint and redress mechanisms mandated by Article 17(9) of the CDSMD as *additional* safeguards against content overblocking:

> the first and second subparagraphs of Article 17(9) of Directive 2019/790 introduce several procedural safeguards, which are additional to those provided for in Article 17(7) and (8) of that directive, and which protect the right to freedom of expression and information of users of online content-sharing services in cases where, notwithstanding the safeguards laid down in those latter provisions, the providers of those services nonetheless erroneously or unjustifiably block lawful content.[159]

Hence, user complaint mechanisms evolving from Article 17(9) only constitute additional ex post measures. As they allow corrections of wrong filtering decisions only after the harm has occurred, they can hardly be considered sufficient per se. First and foremost, it is necessary to have ex ante mechanisms in place that allow permissible content uploads—quotations, parodies, pastiches, etc.—to survive automated content scrutiny. This is an important guideline for E.U. Member States. Implementing Article 17, they must ensure that UGC containing quotations, criticism, review, caricatures, parodies, or pastiches[160] appear directly on the platform.

In addition to limiting the scope of permissible filtering to "manifestly infringing" or "equivalent" content (discussed above), this goal can be achieved in practice by introducing mandatory flagging options for users. To ensure ex ante content availability—without exposure to filtering—domestic legislation in E.U. Member States can enable users to mark quotations,

PROTECTION IS STIFLING CULTURAL CREATIVITY 26–47, 280–83, 357–73 (2020); Christophe Geiger & Elena Izyumenko, *Freedom of Expression as an External Limitation to Copyright Law in the EU: The Advocate General of the CJEU Shows the Way*, 41 EUR. INTELL. PROP. REV. 131, 133–36 (2019).

159. *Id.* ¶ 93.

160. CDSMD art. 17(7), 2019 O.J. (L 130).

parodies, pastiches, etc. as permissible content uploads and oblige OCSSPs to make these uploads directly available on their platforms. An example of national legislation following this approach can be found in Germany.[161] Seeking to avoid disproportionate UGC blocking, Section 9(1) of the German Act on the Copyright Liability of Online Content Sharing Service Providers imposes a "must-carry" obligation on OCSSPs regarding "uses presumably authorised by law."[162] In practice, this means that the platform provider must communicate UGC in this category to the public until a potential complaint procedure establishes that the content infringes copyright. Under Section 11(1) of the German legislation, the OCSSP is also bound to "enable the user to flag the use as authorised by law pursuant to section 5."[163] Section 5(1) clarifies in this context that quotations, caricatures, parodies, pastiches, etc. are forms of use that are authorized by law. Finally, Section 9(2) stipulates that UGC is rebuttably presumed to fall within the privileged must-carry category when it:

> (1) contains less than half of a work or several works by third parties,
>
> (2) combines the part or parts of a work referred to in no. 1 with other content, and
>
> (3) uses the works of third parties only to a minor extent (section 10) or is flagged as legally authorised (section 11) . . . .[164]

Section 9(2) also clarifies that images may be used in this context in their entirety in accordance with Sections 10 and 11 of the German legislation.

As already indicated, Member State legislation of this kind is of particular importance. It provides an essential counterbalance to the predominant outsourcing and concealment logic underlying Article 17 of the CDSMD. As it puts the responsibility back into the hands of the State, the "shall" obligation in Article 17(7) can be qualified as the most promising safeguard against inroads into freedom of expression and information. The Member State responsibility following from this obligation constitutes the only "real" human rights safeguard binding state power directly instead of shifting the responsibility to a private party.

Alarmingly, however, the central importance of the State responsibility arising from Article 17(7) seems to have escaped the attention of most E.U.

---

161. *See* §§ 11(1), no. 1 and 3, 9(1) and (2), and 5(1) of the German Act on the Copyright Liability of Online Content Sharing Service Providers, https://www.gesetze-im-internet.de/englisch_urhdag/index.html (providing an English translation).

162. *Id.* § 9(1).

163. *Id.* § 11(1).

164. *Id.* § 9(2).

Member States. The German implementation model has not become widespread. Instead, many Member States opted for a national transposition that does not offer users specific legal tools, such as statutory flagging options, to benefit from the exemption of quotations, parodies, pastiches, etc.[165] The Netherlands, for instance, gave preference to a literal implementation of Article 17. Effective ex ante mechanisms—capable of placing quotations, parodies, pastiches, etc. beyond the reach of content filtering systems from the outset—are sought in vain. Instead, the Dutch legislator places reliance on complaint and redress mechanisms even though this legal instrument only allows users to take measures ex post: after quotations, parodies, pastiches, etc. have been filtered out and the UGC spectrum has been impoverished.[166] Considering the *Poland* decision, it is doubtful that this implementation approach is adequate. As explained, the CJEU characterized ex post complaint and redress mechanisms as additional safeguards that supplement—but cannot replace—ex ante safeguards, such as the statutory flagging options in Germany and legislation that sets clear limits to the scope of permissible filtering.[167]

4. European Commission Taking Action on the Basis of Audit Reports

As many E.U. Member States seem reluctant to translate their human rights responsibility under Article 17(7) of the CDSMD into statutory ex ante mechanisms that immunize quotations, parodies, pastiches, etc. from content filtering measures, it is important to look beyond the rules in the CDSM Directive. As noted, it is possible to factor DSA provisions into the equation when the CDSM Directive does not contain more specific rules.[168] A legal tool in the DSA that does not appear in the CDSM Directive is the possibility for the executive power to exercise control over content moderation systems based on audit reports. This alternative redress avenue for public authorities seeking to fulfil a watchdog function ex officio has been developed in Article 37 of the DSA.

165. For studies of national implementations of Article 17, see QUINTAIS ET AL., *supra* note 25; CHRISTINA ANGELOPOULOS, ARTICLES 15 & 17 OF THE DIRECTIVE ON COPYRIGHT IN THE DIGITAL SINGLE MARKET COMPARATIVE NATIONAL IMPLEMENTATION REPORT (2022).

166. Article 29c(7) of the Dutch Copyright Act (*Auteurswet*), https://wetten.overheid.nl/BWBR0001886/2022-10-01.

167. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶ 93. As to the German legislation, see the description above and German Act on the Copyright Liability of Online Content Sharing Service Providers, §§ 11(1), no. 1 and 3, 9(1) and (2), 5(1).

168. Digital Services Act recital 11, art. 2(4)(b) 2022 O.J. (L 277).

With respect to very large online platforms (VLOPs)[169] and very large online search engines (VLOSEs),[170] Article 37(1) of the DSA orders annual audits to assess compliance, among other things, with the due diligence obligations set forth in Chapter III of the DSA.[171] Interestingly, one of the obligations laid down therein concerns the "diligent, objective and proportionate"[172] application of content moderation systems in line with Article 14(4) of the DSA. Supplementing the complaint and redress system of Article 17(9) of the CDSMD that depends on user initiatives, Article 37 of the DSA may thus offer an important alternative basis that allows the executive power to prevent human rights violations.

Article 37(3) of the DSA ensures that organizations establishing the audit report are independent from the VLOPs and VLOSEs under examination. In particular, it prevents organizations from performing an audit when they have a conflict of interest with the VLOP or VLOSE concerned, or with a legal person connected to that service provider. The audit report must contain an opinion—in the categories "positive," "positive with comments," and "negative"—on whether the VLOP or VLOSE has complied with the obligations and commitments under Chapter III of the DSA, including the above-described human rights and proportionality obligations laid down in Article 14(1) and (4) of the DSA.[173] If the audit opinion is not "positive," auditors are bound to include operational recommendations and specify the measures necessary to achieve compliance. They must also recommend a timeframe for achieving compliance.[174] In such a case, the VLOP or VLOSE concerned must adopt, within one month from receiving the recommendations, an audit implementation report. If the provider does not intend to implement the operational recommendations, it must give reasons

169. In accordance with Article 33(1) DSA, an online platform is qualified as a VLOP when it has a number of average monthly active service recipients in the European Union that is equal to, or higher than, 45 million, and has been designated as a VLOP by the European Commission pursuant to Article 33(4) DSA.

170. In accordance with Article 33(1) DSA, a search engine is qualified as a VLOSE when it has a number of average monthly active service recipients in the European Union that is equal to, or higher than, 45 million, and has been designated as a VLOSE by the European Commission pursuant to Article 33(4) DSA.

171. *See Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines*, EUR. COMM'N (Apr. 25, 2023), https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413 (listing the first set of VLOPs and VLOSEs designated by the Commission).

172. Digital Services Act art. 14(4), 2022 O.J. (L 277).

173. *Id.* art. 37(4)(g).

174. *Id.* art. 37(4)(h).

for not doing so and set out alternative measures that it has taken to address the instances of non-compliance identified in the audit report.[175]

As to the role of the European Commission, Article 42(4) of the DSA is of particular importance. This provision obliges VLOPs and VLOSEs to transmit audit reports and audit implementation reports to the Commission without undue delay. If, based on this information, the Commission suspects a VLOP or VLOSE of infringing Article 14 of the DSA, it can initiate proceedings pursuant to Article 66(1) of the DSA. It may request further information, conduct interviews, and inspect premises to learn more about the suspected infringement.[176] In case of a "risk of serious damage for the recipients of the service," Article 70(1) of the DSA entitles the Commission to order interim measures on the basis of a prima facie finding of infringement. If the Commission finally establishes non-compliance with "the relevant provisions of this Regulation"—including the human rights safeguards in Article 14(4) of the DSA—in a decision pursuant to Article 73(1) of the DSA, it may impose fines of up to six percent of the VLOP's or VLOSE's total worldwide annual turnover in the preceding financial year.[177] For the imposition of fines, Article 74(1) of the DSA requires a finding that the service provider under examination has infringed Article 14(4) of the DSA intentionally or negligently.

Considering this cascade of possible Commission actions, the potential of the audit mechanism in Article 37 of the DSA must not be underestimated. The audit system may be an important addition to the canon of norms in the CDSM Directive and, in particular, a promising counterbalance to outsourcing/concealment risks arising from the regulatory design of Article 17 of the CDSMD. Like the Member State legislation discussed in the preceding section, Commission interventions evolving from the problem analysis in an audit report are welcome departures from the strategy to pass on human rights responsibilities to platforms or users. Namely, the state power itself—in this case the Commission as the executive body of an international intergovernmental organization—remains directly responsible for detecting and remedying human rights deficits.

A potential blind spot of the described audit cascade leading to investigations, however, is this: in order to offer sufficient starting points for Commission action, audit reports addressing content moderation systems must go beyond a general problem analysis. The audit opinion must convincingly discuss a platform's failure to satisfy human rights obligations

175. *Id.* art. 37(6).
176. *Id.* arts. 67–69.
177. *Id.* art. 74(1).

evolving from Article 14(4) of the DSA. It must contain a concrete assessment of the risk of human rights violations and a sufficient substantiation of that risk. Hence, the Commission must ensure sufficient focus on the detailed examination of human rights deficits. It should adopt a delegated act based on Article 37(7) of the DSA that creates clarity about the necessity to devote particular attention to human rights questions in audit reports and seek all information necessary for a proper assessment of human rights risks.[178]

C. OUTSOURCING AND CONCEALMENT STRATEGY PUTTING HUMAN RIGHTS AT RISK

On balance, the closer inspection of content moderation rules in the CDSM Directive and the DSA confirms a worrying tendency of reliance on industry cooperation and user activism to safeguard human rights. Both exceptions to the rule of outsourcing to private entities—the transformative use safeguard in Article 17(7) of the CDSMD and the audit system evolving from Article 37 of the DSA—are currently underdeveloped. E.U. Member States have not consistently taken specific legislative action to protect transformative UGC from content filtering measures. The success of the DSA cascade of interventions—from audit reports to non-compliance decisions and fines[179]—is unclear. Therefore, it would be premature to sound the all-clear based on these opportunities to engage state power itself in initiatives to uphold human rights.

While this outcome of the foregoing analysis already darkens the horizon, the discussion of human rights risks arising from outsourcing and concealment strategies would be incomplete without shedding light on how these strategies work out in practice. When content moderation systems detect traces of protected third-party material in UGC, the most common rightholder reaction is not the blocking of the content at issue. Instead, rightholders often opt for "monetization"—the opportunity to garner advertising revenue that accrues from the continued online availability of UGC. Surprisingly, the monetization mechanism largely remains uncharted territory in both the CDSM Directive and the DSA. Hence, the question arises whether human rights risks emerging from inappropriate outsourcing and concealment schemes are particularly strong in this area. We turn to this issue in the following chapter.

178. At time of writing, the Commission has presented its Delegated Regulation on Independent Audits for public feedback. *See Digital Services Act: Delegated Regulation on Independent Audits Now Available for Public Feedback*, EUR. COMM'N (May 5, 2023), https://digital-strategy.ec.europa.eu/en/news/digital-services-act-delegated-regulation-independent-audits-now-available-public-feedback.

179. Digital Services Act arts. 66–74, 2022 O.J. (L 277).

III. CASE STUDY: ALGORITHMIC MONETIZATION OF USER-GENERATED CONTENT

UGC monetization is a largely underexplored but a highly relevant copyright content moderation action in practice.[180] Although not initially obvious, monetization has important human rights dimensions, and therefore offers a good case study of regulatory outsourcing and concealment tendencies. Transparency reports from the largest platforms suggest that monetization is a popular—perhaps the most popular—moderation action taken by rightholders that have access to platforms' content recognition tools. Despite this, the CDSM Directive largely ignores the topic, and the DSA only tackles it at a superficial level, mostly by outsourcing its regulation to private parties (as discussed in Section III.A). This regulatory design has enabled the emergence of copyright management systems and practices that allow platforms and the largest rightholders to dictate the terms of this crucial form of exploitation of copyrighted content. The workings of these systems are mostly concealed behind complex terms and conditions and opaque algorithmic systems.

Our analysis of the visible parts of these mechanisms, however, suggests that they work in ways that are partly contrary to E.U. copyright law, and mostly to the detriment of individual UGC creators (as discussed in Section III.B). The combination of a lax regulatory framework and the resulting monetization practices leads to a host of problems, including the lack of a proper legal basis for monetization of transformative UGC by third-party rightholders, the lack of remuneration for user creativity, and the misappropriation of monetary rewards by larger copyright holders. These problems translate into three human rights deficits: (1) the appropriation and exploitation of transformative UGC based on the exclusive rights of third parties while failing to notice copyright limitations that support freedom of expression; (2) the violation of copyright of UGC creators even though this copyright, just like the copyright of larger rightholders, falls under the fundamental right to property; and (3) the discriminatory treatment of these creators as compared to larger rightholders (as discussed in Section III.C).

180. A notable exception in scholarship is Henning Grosse Ruse-Khan, *Automated Copyright Enforcement Online: From Blocking to Monetization of User-Generated Content*, UNIV. OF CAMBRIDGE FACULTY OF L. RSCH. PAPER NO. 8/2020 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3565071.

A. UGC MONETIZATION BETWEEN E.U. COPYRIGHT LAW AND THE DSA

This Section highlights how monetization of UGC by platforms, despite its economic significance, remains a relatively unregulated space in E.U. copyright law. We start by conceptualizing monetization as a type of content moderation action from a legal perspective (in Section III.A.1). We then explain how the E.U. copyright *acquis* does not directly regulate UGC monetization in the CDSM Directive, leaving the matter mostly to private ordering of platforms and their users (as discussed in Section III.A.2). Finally, after clarifying how the DSA applies to copyright hosting platforms (OCSSPs or not), we explore whether the DSA places any constraints on this private ordering (as discussed in Section III.A.3). Our conclusion is that both the CDSMD and the DSA mostly outsource the regulation of UGC monetization to private parties. Section III.B then examines this outsourcing exercise in practice. The last analysis, in Section III.C, discusses the human rights deficits it creates for users and the public.

1. Monetization as Content Moderation

To understand how the monetization of UGC enjoying copyright protection ("copyrighted UGC") is regulated in E.U. law, it is helpful to place that action within the broader context of "content moderation." The concept of content moderation is for the most part not clearly defined in literature and is used to describe a wide spectrum of platform activities. Some authors view content moderation as a broad set of governance mechanisms that facilitate cooperation and prevent abuse,[181] while others describe it as the organized practice of screening UGC to determine its appropriateness to a particular context or set of constraints.[182] The term has also been defined as the set of practices that online platforms use to screen, rank, filter, and block UGC,[183] or as the detection, assessment, and intervention taken on content or behavior deemed unacceptable by platforms or other information intermediaries.[184]

181. James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42 (2015); *see also* Giovanni De Gregorio, *Democratising Online Content Moderation: A Constitutional Framework*, 36 COMPUT. L. & SEC. REV. 105374, 2 (2020) (building on Grimmelmann's definition).

182. SARAH T. ROBERTS, CONTENT MODERATION (2017), https://escholarship.org/uc/item/7371c1hf. *Cf.* Sarah Myers West, *Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms*, 20 NEW MEDIA & SOC'Y 4366 (2018); ROBYN CAPLAN, CONTENT OR CONTEXT MODERATION? (2018), https://datasociety.net/library/content-or-context-moderation/.

183. Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 41 (2020).

184. Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich & Sarah Myers

Some authors offer narrower definitions linked to the technical action taken by a service provider, viewing content moderation systems as those that classify UGC using either matching or prediction, resulting in a decision and subsequent governance outcome, such as content removal, geo-blocking, or account takedown.[185] Finally, some authors define content moderation from the perspective of the remedies associated with it, including those against individual content items or against an online account, those consisting of visibility restrictions, and those imposing financial consequences, among others.[186]

Existing scholarly analysis of content moderation has been carried out in the absence of legal definitions of the concept in both U.S. and E.U. law. In a significant legal innovation, the DSA now advances a legal definition of "content moderation" as

> the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account.[187]

The definition covers, firstly, activities by various types of intermediaries across the technology "stack"; not only online platforms, but also providers of other types of "intermediary services," such as "mere conduit" and "caching,"[188] as well as—in theory—"online search engines."[189] Secondly, content moderation involves actions taken with the specific purpose of

West, *Expanding the Debate About Content Moderation: Scholarly Research Agendas for the Coming Policy Debates*, 9 INTERNET POL'Y REV. 1 (2020).

185. Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 BIG DATA & SOC'Y 1 (2020). This definition would exclude recommender systems, norms, design decisions, and architectures.

186. Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021).

187. Digital Services Act art. 3(t), 2022 O.J. (L 277).

188. Defined in Article 3(g) DSA. In the engineering community, networks often are described in layers, which each relate to a separate functional level of the network. *Cf., e.g.*, *OSI Model*, WIKIPEDIA, https://en.wikipedia.org/w/index.php?title=OSI_model&oldid=1072600519 (describing the Open Systems Interconnection model).

189. We say in theory because although the definition of "intermediary services" in Article 3(g) DSA does not list "online search engines," the definition of the latter in Digital Services Act art 3(j), 2022 O.J. (L 277) does mention that they are a type of intermediary service.

detecting, identifying, and addressing "illegal content"[190] or information that is incompatible with the terms and conditions of intermediary service providers. Such content (or part of it) is sometimes referred to as "harmful" or "lawful but awful."[191] Thirdly, the content in question must be provided by the "recipients of the service," i.e., originate from the user rather than the provider itself.[192] "Online platforms"[193]—the type of intermediary we are interested in—mainly involve content uploaded by users that we here refer to as UGC.

The DSA's definition is not exhaustive. It encompasses a general clause and various types of examples. The general clause states that content moderation encompasses measures that impact the availability, visibility, and accessibility of illegal content or information. Subsequently, two sets of examples of such measures are provided. The first set of measures pertains to *content or information*, such as the demotion, demonetization, disabling access, and removal thereof. Whereas content-level measures of disabling access and removal are restrictions on availability or accessibility, measures such as demotion and demonetization are closer to restrictions on visibility and therefore closer to what has been referred to in scholarship and practice as "shadow banning."[194] The second set exemplifies measures that pertain to the *user or account*, such as the termination or suspension of the user's account, i.e., temporary or permanent "de-platforming."[195] The following figure provides a schematic overview of the definition.

190. Defined in Digital Services Act art. 3(h), 2022 O.J. (L 277).

191. *See, e.g.*, Eric Goldman & Jess Miers, *Online Account Terminations/Content Removals and the Benefits of Internet Services Enforcing Their House Rules*, 1 J. FREE SPEECH L. 191, 194 (2021).

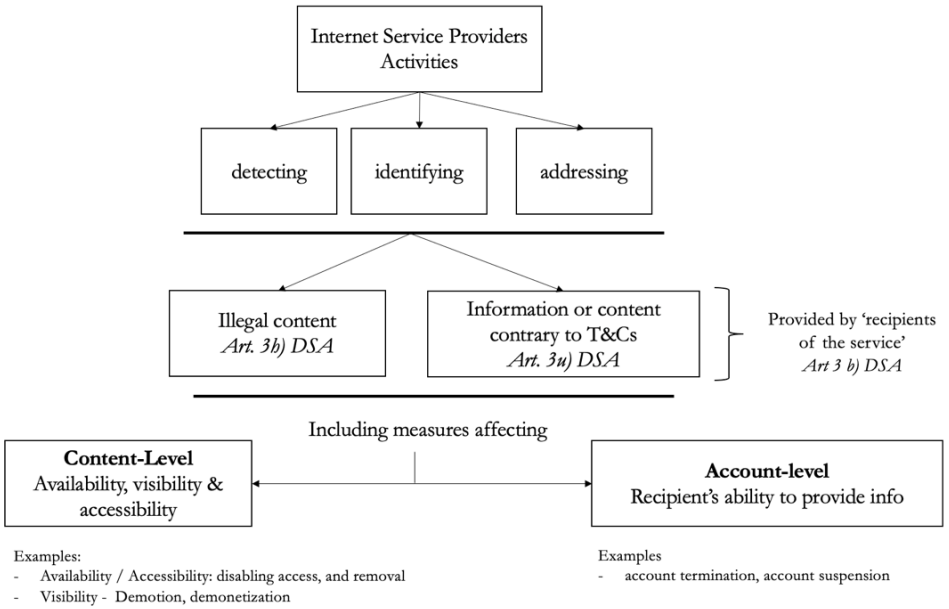
192. Defined in Digital Services Act art. 3(b), 2022 O.J. (L 277); *see also* Article 3(p), (q) (defining, respectively, "active recipient of an online platform" and "active recipient of an online search engine").

193. Defined in *id.* art. 3(i), 2022 O.J. (L 277).

194. West, *supra* note 182; Kelley Cotter, *"Shadowbanning is Not A Thing": Black Box Gaslighting and the Power to Independently Know and Credibly Critique Algorithms*, INFO., COMM'CN & SOC'Y 1 (2021); Laura Savolainen, *The Shadow Banning Controversy: Perceived Governance and Algorithmic Folklore*, 44 MEDIA, CULTURE & SOC'Y 1091 (2022); Paddy Leerssen, *An End to Shadow Banning? Transparency Rights in the Digital Services Act Between Content Moderation and Curation*, 48 COMPUT. L. & SEC. REV. 105790 (2023). The DSA describes visibility restrictions in Recital 55 as those that "may consist in demotion in ranking or in recommender systems, as well as in limiting accessibility by one or more recipients of the service or blocking the user from an online community without the user being aware ('shadow banning')."

195. On the concept of "de-platforming," see Shagun Jhaver, Christian Boylston, Diyi Yang & Amy Bruckman, *Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter*, 5 PROC. ACM HUM.-COMPUT. INTERACT. 381 (2021); Richard Rogers, *Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media*, 35 EUR. J. COMMC'N 213 (2020); Helen Innes & Martin Innes, *De-Platforming Disinformation: Conspiracy Theories and Their Control*, INFO., COMMC'N & SOC'Y 1 (2021).

Figure 1. Content Moderation DSA Definition Diagram



The conceptual framework provided by this definition is useful to examine the regulation of copyrighted UGC. With some degree of certainty, it helps to map out what types of content moderation actions are currently regulated in E.U. copyright law, what actions are in a legal grey area, and what actions are wholly unregulated. As we shall see, this determination is also crucial to identify which parts of the more general complementary DSA framework may apply to copyright content moderation actions by platforms in addition to the specific rules of the copyright *acquis*. Our main argument in the following analysis is that monetization of copyrighted UGC on online platforms is largely unregulated by E.U. copyright law. Considering the content moderation concept underlying the DSA, however, it can be said that it falls within the scope of the DSA's content moderation rules.

2. E.U. Copyright Law and Monetization

We have discussed *supra* the basic workings of Article 17 of the CDSMD, as well as related issues of outsourcing (in Section II.A) and concealment (in Section II.B), which lead to human rights deficits. Here, we merely wish to point out that from a copyright content moderation perspective, Article 17 mainly regulates filtering, blocking and takedown actions for UGC in relation to which copyright owners have provided "relevant and necessary

information" or a "sufficiently substantiated notice."[196] It does not contain specific rules on other types of content moderation actions, such as restrictions on visibility of content or monetization of UGC that is not covered by licensing deals.[197]

Then, the question that arises is whether *other* provisions in the CDSM Directive could apply to the monetization of UGC. In this respect, two provisions—on fair remuneration and transparency[198]—can theoretically be considered, although neither is a particularly good fit for UGC monetization.

The first provision is Article 18 of the CDSMD, which establishes the principle that creators (authors and performers) who license their works or subject matter must receive appropriate and proportionate remuneration.[199] In the context of OCSSPs, one could imagine that this principle could protect UGC creators against abusive remuneration practices by platforms.[200] However, whereas such a principle clearly applies to remuneration paid to creators in the context of licensing deals evolving from Article 17(1) of the CDSMD, it is harder to see how it can be operationalized vis-à-vis unlicensed content that is subsequently monetized through advertisement on the platform. This type of use and remuneration was not envisioned in the preparatory works of the Directive or, to the best of our knowledge, discussed during the legislative process. Before Article 17 was adopted, this practice also largely occurred in the shadow of the hosting safe harbor, in a context where platforms that complied with a notice-and-takedown regime were not directly liable for the UGC they hosted.

196. CDSMD art. 17(4)(b),(c), 2019 O.J. (L 130).

197. *See* QUINTAIS ET AL., *supra* note 25 (reaching a similar conclusion).

198. CDSMD arts. 18, 19, 2019 O.J. (L 130). These provisions may apply in the context of UGC since users-creators enter into non-exclusive license agreements with platforms to exploit uploaded content on their services, including for monetization purposes. *See* João Pedro Quintais, Giovanni De Gregorio & João C. Magalhães, *How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and its Implications*, 48 COMPUT. L. & SEC. REV. 105792 (2023).

199. Member States have discretion on the mechanism to adopt when implementing this principle, as long as it complies with E.U. law. *See* CDSMD recital 73 2019 O.J. (L 130) (a lump sum payment may amount to proportionate remuneration "but it should not be the rule").

200. The application of Art. 18 CDSMD to OCSSPs has been confirmed by Commissioner Thierry Breton in response to a parliamentary question by a MEP. *See Appropriate and Proportionate Remuneration for All Online Services*, EUROPEAN PARLIAMENT (Dec. 5, 2021), https://www.europarl.europa.eu/doceo/document/E-9-2021-002618_EN.html; *E-002618/2021; Answer given by Mr. Breton on behalf of the European Commission*, EUROPEAN PARLIAMENT (Sept. 9, 2021), https://www.europarl.europa.eu/doceo/document/E-9-2021-002618-ASW_EN.html.

The second provision is Article 19 of the CDSMD, which imposes a transparency obligation on licensees or transferees of works or performances to provide creators with detailed information on the exploitation of their creations, including modes of exploitation, revenues generated, and remuneration due.[201] The obligation must account for the specificities of each sector and must be fulfilled on a regular basis. It is unclear whether and to what extent this transparency obligation applies to the context of UGC monetization (outside traditional licensing deals) and, if it does, whether current information practices of OCSSPs conform with this requirement.[202]

In conclusion, just like Article 17, the rules in Articles 18 and 19 of the CDSMD only minimally restrict the autonomy of platforms to establish their own internal governance policies in relation to UGC monetization. The principle of appropriate and proportionate remuneration is too broadly defined to effectively constrain a platform's remuneration and monetization policies towards users. Additionally, it is uncertain whether the requirement for transparency has a significant impact on a platform's current reporting practices to individual creators. In other words, when it comes to moderation actions related to monetization of (unlicensed) content, E.U. copyright law affords platforms a broad autonomy space. Given the significant power imbalance between platforms and users,[203] the question arises whether outside the copyright *acquis*, it is possible to find relevant provisions that constrain the monetization of copyrighted UGC, particularly in the DSA. To answer this question, we first clarify the DSA's application to copyright platforms and then examine its specific rules on monetization restrictions.

---

201. *See* Séverine Dusollier, *The 2019 Directive on Copyright in the Digital Single Market: Some Progress, A Few Bad Choices, and an Overall Failed Ambition*, COMMON MKT. L. REV., 979, 1023–24 (2020).

202. YouTube does seem to provide creators who participate in the YouTube Partner Program with information on the modes of exploitation of their videos, the revenues generated by their videos, and the remuneration due via creators' respective YouTube Studio accounts. In this account, creators can select 'Analytics' to see revenue reports related to their earnings. The percentage of the gross revenue generated with the videos that is paid to the creator (revenue share) is outlined in the creator's partner agreement with YouTube. The revenue share depends on the terms of the 'module' selected by the creator (e.g., 'Watch Page Monetization Module' and 'Shorts Monetization Module.' *See Check your YouTube revenue*, YOUTUBE, https://support.google.com/youtube/answer/9314488; *YouTube partner earnings overview*, YOUTUBE, https://support.google.com/youtube/answer/72902?hl=en&ref_topic=9257988#zippy=%2Cwhere-can-i-see-my-earnings (last visited Mar. 28, 2023).

203. Quintais, Appelman, & Fahy, *supra* note 37; Quintais, De Gregorio, & Magalhães, *supra* note 198.

3. Digital Services Act and Monetization

The DSA applies to online platforms that host copyrighted content. But it applies differently depending on whether a platform qualifies as an OCSSP or not.[204] Whereas some large-scale platforms, especially those with video-sharing features such as YouTube, Facebook, and Instagram, clearly qualify as OCSSPs, others are explicitly excluded from that category due to the carve-outs in Article 2(6) of the CDSMD.[205] But a grey area subsists, caused by the fact that the legal definition of OCSSPs relies on open-ended concepts, such as "main purpose," "large amount," and "profit-making purpose," necessitating a case-by-case assessment of whether providers meet these requirements.[206] In addition, the extent to which a platform is covered by the definition may remain unclear. This is because a provider may offer multiple services; thus, a service-by-service analysis is necessary to determine whether a provider qualifies as an OCSSP.[207] Consequently, as regards copyright liability for the content it hosts, the same provider may be subject to Article 17 of the CDSMD for certain services and the more general copyright liability regime, following from acts of communication to the public in the sense of Article 3 of the InfoSoc Directive, for others.

Regarding copyright liability falling outside the scope of Article 17 CDSMD, the general safe harbor system in the DSA remains applicable, including the safe harbor for hosting in Article 6 of the DSA (more on this distinction below). Considering there are numerous platforms that host copyrighted content, as well as other types of content, while providing different services, it is a complex task to determine liability regimes and respective content moderation obligations.[208] In addition to the liability rules, whether a copyright-hosting platform qualifies as an OCSSP or not, it will be subject to the DSA's due diligence obligations for online platforms or VLOPs, albeit to different degrees.[209] Consequently, it is important to explore which rules, if any, the DSA might contain that supplement E.U. copyright law concerning monetization.

The DSA's definition of content moderation explicitly refers to demonetization as a content-level restriction by intermediaries.[210] Monetization is conceptualized as an action of obtaining monetary payment or

204. *See* Quintais & Schwemer, *supra* note 111.
205. *See supra* Part I.
206. Guidance Art. 17 CDSMD, *supra* note 18, at n. 18, 3–5.
207. *Id.* at n. 18, 5.
208. QUINTAIS ET AL., *supra* note 25.
209. *See* Quintais & Schwemer, *supra* note 111; Peukert et al., *supra* note 111.
210. *See supra* Figure 1.

revenue through advertisement of "information" (in our case, copyrighted UGC) provided by the user. This activity can be restricted by suspending or terminating the monetary payment or revenue associated to that information.[211] The question then is what types of obligations are imposed on UGC platforms relating to demonetization.[212]

First, Article 17 of the DSA obliges providers of hosting services to accompany each content moderation action affecting individual recipients of the service with statements of reasons. Following Article 17(1)(b) of the DSA, such statements of reasons are also required for decisions involving the "suspension, termination or other restriction of monetary payments," i.e., demonetization actions, on the ground that the information provided by the user is illegal content (here: copyright-infringing UGC) or incompatible with the provider's terms and conditions.[213] A statement of reasons must contain detailed information on the action taken[214] and fulfil two core functions: (1) to *notify* users of any sanctions relating to their content, and (2) to *explain* why they were imposed.[215] This is especially important considering research that shows how demonetization actions are challenging to observe in practice.[216]

Second, Article 20 of the DSA requires online platform service providers to provide recipients of their services with access to an effective internal complaint-handling system that enables them to lodge complaints against decisions taken by the provider of the online platform, including against "decisions whether or not to suspend, terminate or otherwise restrict the ability to monetize information provided by the recipient."[217] As noted, for OCSSPs it is unclear to what extent this provision would apply to conventional moderation actions, such as blocking or removal of content, which are already regulated in Article 17(4) of the CDSMD. However, Article 20 of the DSA should clearly apply to OCSSPs and non-OCSSPs alike in respect of *complaints against demonetization decisions*. In fact, Article 20 offers a promising array of tools. Access to complaint-handling systems should be available for at least six months. Complaints should be easy to submit and supported with sufficient

211. Digital Services Act recital 55, 2022 O.J. (L 277).

212. As it was sufficiently discussed above, we will not further address here the DSA's cornerstone provision on terms and conditions (Article 14), which applies also to demonetization as a type of content moderation restriction.

213. Digital Services Act art. 17(2)(b), 2022 O.J. (L 277).

214. *Id.* art. 17(3).

215. Leerssen, *supra* note 194, at 7.

216. Robyn Caplan & Tarleton Gillespie, *Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy*, April-June SOCIAL MEDIA + SOC'Y 1 (2020); *see also* Leerssen, *supra* note 194 (noting the "importance of notice policies for unobservable remedies such as demonetization").

217. Digital Services Act art. 20(1)(c), 2022 O.J. (L 277).

evidence. Complaints must be handled promptly, fairly, and diligently. Platforms should reverse decisions without undue delay if the complaint sufficiently establishes that the reported information is not illegal or incompatible with the platforms' terms and conditions. Complainants should be promptly informed of decisions, given options for out-of-court resolution, and provided with other avenues for redress. Qualified staff should oversee complaint decisions, avoiding sole reliance on automated decision making.

Article 21 of the DSA then allows users affected by a platform's decision to select any certified out-of-court dispute settlement body to resolve disputes relating to those decisions. Without going into detail on the certification process, it is noteworthy that platforms must bear all the fees charged by the out-of-court dispute settlement body if a user prevails in the dispute. Conversely, should the platform prevail, the user does not have to reimburse any of the platforms' fees or expenses, unless the user manifestly acted in bad faith.[218]

To be sure, these are helpful provisions. Clear and specific information about the reasons why monetary payments related to UGC have been restricted and information about redress possibilities theoretically offer users the opportunity to effectively take action against demonetization. Likewise, greater clarity and detail regarding in-platform and out-of-court dispute settlement regarding demonetization are positive, especially when accompanied by favorable rules on costs. However, the outsourcing and concealment criticism developed above regarding complaint and redress mechanisms applies with equal force here. As we show in Section III.B, most monetization claims by rightholders, e.g., in YouTube's Content ID tool, are not contested by users despite the availability of complaint and redress mechanisms. Even if Article 17 of the DSA improves the *quality* of information surrounding a monetization restriction for the affected user, it does not change the ex post nature of the mechanism. Similar arguments can be made for Articles 20 and 21 of the DSA. What remains to be seen is whether these provisions will have any meaningful impact on the behavior of affected users, absent more fundamental regulation of copyright monetization. One possible approach could be to impose ex ante restrictions on the ability of rightholders with access to content recognition tools to claim monetization in the first place.

In short, the DSA's approach to monetization operates at the level of transparency and ex post safeguards. These features mirror to a large extent what we have called a human rights "outsourcing" approach in our analysis

218. *Id.* art. 21(5).

above. The practice of UGC monetization, discussed below, suggests that this approach is problematic and leads to the concealment of human rights deficits.

B. THE PRACTICE OF UGC MONETIZATION

This Section explores the systems and tools developed by platforms for the moderation of copyrighted content through three case studies of large-scale platforms that qualify as OCSSPs: YouTube, Meta's Facebook and Instagram, and TikTok. We identify, describe, and examine the functionality of the systems and tools that these platforms make available to rightholders, including options to track usage, block, and monetize protected content. Our data is drawn from publicly available information pages on the platforms' websites, the platforms' own copyright transparency reports covering the first half of 2022, recordings of the 2019–2020 Commission Stakeholder Dialogue, and existing literature. Our analysis highlights the relative importance of monetization as a content moderation action, the way in which existing moderation rules and systems favor monetization by legacy enterprise rightholders, and three significant human rights deficits arising therefrom.

1. YouTube

Launched in 2005 and soon thereafter acquired by Google, the online video-sharing website YouTube is one of the longest running and most popular online platforms among creators and internet users worldwide.[219] As part of its free service, YouTube allows users to upload, watch, like and share videos. To upload content, users simply sign in into their account, upload a video file, enter the necessary details (e.g., title, description, licensing information) and add special elements such as subtitles. With every upload, YouTube's content recognition tools screen and check the file for copyrighted third-party materials and, if the user participates in the YouTube Partner program (see below), a check is made also for advertising suitability, after which the user can choose settings for monetization.[220]

YouTube offers various tools to rightholders to "protect and manage"[221] copyrighted content on the platform. The platform's Copyright Management

219. In the European Union alone, YouTube counts 401.7 million monthly active users. *See* GOOGLE, INFORMATION ABOUT MONTHLY ACTIVE RECIPIENTS UNDER THE DIGITAL SERVICES ACT (EU) (2022), https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-24_2022-7-1_2022-12-31_en_v1.pdf.

220. *See Upload YouTube Videos*, YOUTUBE, https://support.google.com/youtube/answer/57407?hl=en&co=GENIE.Platform%3DDesktop&oco=0#zippy=%2Cdetails%2Cmonetization%2Cad-suitability%2Cvideo-elements%2Cchecks (last visited Mar. 2, 2023).

221. *Overview of Copyright Management Tools*, YOUTUBE, https://support.google.com/youtube/answer/9245819?hl=en#zippy=%2Ccopyright-takedown-webform%2Ccontent-id%2Ccopyright-match-tool (last visited Mar. 2, 2023).

Suite consists of three main products: the Webform, the Copyright Match Tool and the Content ID system.[222] Each product targets different types of rightholders, depending on both the *scale* of the rightholders' content management needs and the rightholders' *capabilities* (i.e., knowledge, resources) to manage copyright.[223]

Webform is a simple tool through which any user holding copyright can manually request the removal of their copyrighted content[224] from the platform. Its functionality is therefore that of a traditional notice-and-takedown system. The Webform is specifically meant to accommodate "those with infrequent [copyright protection] needs"[225] and is open to everyone, i.e., more than 2 billion channels worldwide.[226]

The Copyright Match Tool is a more sophisticated product based on Content ID matching technology (see below). The tool automatically scans new user uploads for matches with existing protected content on the platform. Contrary to Webform, Copyright Match is not open to everyone. Those eligible for the use of this tool are, primarily, channels and other creators that are enrolled in the YouTube Partner Program (a program through which selected creators get access to resources to monetize their content)[227] and channels that have filled out YouTube's copyright management tools application form and thereby shown a need for an advanced rights management tool.[228] Since October 2021, the tool has also become available to YouTube users who submitted valid/approved Webform removal requests and indicated in the Webform that they would like YouTube to prevent the future upload of (any copies of) the reported video.[229] In that capacity, the Copyright Match Tool has the affordance of a notice-and-staydown (NSD)

222. YOUTUBE, COPYRIGHT TRANSPARENCY REPORT (2022), https://services.google.com/fh/files/misc/hytw_copyright_transparency_report.pdf?hl=en.

223. *Id.* at 1, 4; *see also* YouTube Presentation, European Commission, Recording of the Third Meeting of the Stakeholder Dialogue on Article 17 of the Directive on Copyright in the Digital Single Market (Nov. 25, 2019) (including a presentation by YouTube) [hereinafter Article 17 Dialogue Recording, YouTube Presentation].

224. While YouTube's business model is built around audiovisual content, Webform can also be used to remove other copyright-protected works from the platform, including audiobooks, ebooks and still images. Article 17 Dialogue Recording, YouTube Presentation, *supra* note 223.

225. YOUTUBE, COPYRIGHT TRANSPARENCY REPORT, *supra* note 222, at 4.

226. *Id.*

227. *How to Make Money on YouTube*, YOUTUBE, https://www.youtube.com/creators/how-things-work/video-monetization/ (last visited Oct. 17, 2023).

228. YOUTUBE, COPYRIGHT TRANSPARENCY REPORT, *supra* note 222, at 1–2.

229. *Id.*

system. Taken together, more than two million channels have access to Copyright Match.[230]

When the scanning tool finds a match, it provides the rightholder with information on the total views of the user upload, the channel it was uploaded to, the percentage of protected content that was used as well as some screenshots of the video. The system also indicates whether the user upload has different video but the same audio, and whether there is only a partial match, such as in the case of sampling. In this interface, rightholders are given three options: (1) do nothing and leave the video up; (2) file a removal request and ask YouTube to automatically prevent the upload of copies in the future; or (3) contact the uploader.[231] Like the Webform, the Copyright Match tool does not afford rightholders the option to monetize the matched content.

The last and most powerful tool within the Copyright Management Suite in terms of automation and available copyright enforcement actions, is Content ID. Since 2007, this system has enabled copyright holders to identify new user uploads that include materials they own, and to automatically initiate action based on self- and pre-specified rules dictating how to handle matched content. Content ID is specifically aimed at rightholders "with the most complex rights management needs, such as movie studios, record labels, and collecting societies."[232] To be approved for Content ID, rightholders must demonstrate a "need for [a] scaled tool," an "understanding of copyright" as well as the "resources to manage the complex automated matching system."[233] Smaller, independent creators may only indirectly access (features of) the system via intermediary service providers that manage rights through the system on behalf of others.[234] In the first half of 2022, approximately 9,000 (enterprise) partners had access to Content ID.[235]

To set up for Content ID, eligible rightholders must provide YouTube with extensive information. This includes: (1) reference files (e.g., audio, visual

230. *Id.*

231. *Id.* at 3 ("From this interface, users can choose to archive the match and leave the video up, file a takedown request (with the option to ask YouTube to automatically prevent copies), or contact the uploader").

232. *Id.*

233. *Id.* at 1.

234. *Id.* at 3. Such service providers may include multi-channel networks and other organizations. *See, e.g.*, *Services Directory*, YOUTUBE, https://servicesdirectory.withyoutube.com/directory/#?services=content-id-management (providing a list of service providers) (last visited Mar. 2, 2023).

235. YOUTUBE, COPYRIGHT TRANSPARENCY REPORT, *supra* note 222, at 1, 4.

or audiovisual)[236] that meet the content eligibility criteria;[237] (2) ownership information, e.g., on the territories in which the content is owned and how much of the content is owned; (3) metadata that describe the content, e.g., titles and industry identification numbers; and (4) the preferred copyright moderation or enforcement actions to be carried out in the event of a match detection between a user upload and the reference content ("match policies").[238] In our view, these information requirements are sufficient to meet the threshold of "relevant and necessary information" set out in Articles 17(4)(b) and (c) of the CDSMD.[239] As explained, the provision of such information triggers YouTube's best efforts obligations to deploy preventive measures to ensure the unavailability of the notified works on the platform and, where appropriate, to prevent the future upload of works for which rightholders in the past provided a valid notice for removal (i.e., the "stay-down" part of Article 17(4)(c) of the CDSMD).

After the upload of reference files and the specification of match policies, Content ID starts checking new uploads to the platform against such files. Matching videos are automatically claimed on behalf of the rightholder, upon which the preferred match policies are applied. There are three types of actions

236. *Using Content ID*, YOUTUBE, https://support.google.com/youtube/answer/3244015?hl=en (last visited Oct. 17, 2023).

237. For example, the rightholder must hold exclusive rights over the reference content for the territories ownership is claimed, the content must be sufficiently distinct (*e.g.*, no remasters) and each piece of intellectual property must have an individual reference (*e.g.*, complications, mashups and full albums cannot be filed as a reference), see *Content eligible for Content ID*, YOUTUBE, https://support.google.com/youtube/answer/2605065#zippy=%2Cexclusive-rights%2Cdistinct-reference-content%2Cindividual-references-for-each-piece-of-intellectual-property%2Coriginal-video-game-soundtrack-guidelines%2Ccontent-that-is-sold-or-licensed-at-scale-for-incorporation-into-other-works%2Casset-metadata-for-reference-content%2Cfingerprint-only-reference-content (last visited Mar. 2, 2023). The media files are translated by YouTube into unique digital 'fingerprints.' In exceptional cases, YouTube allows rightholders to fingerprint the media file on their own devices and provide YouTube the fingerprint instead of the original media file. *See* Article 17 Dialogue Recording, YouTube Presentation, *supra* note 223.

238. YOUTUBE, COPYRIGHT TRANSPARENCY REPORT, *supra* note 222, at 3; *see also* Article 17 Dialogue Recording, YouTube Presentation, *supra* note 223.

239. The second part of Article 17(4)(c) CDSM (*i.e.*, the 'notice-and-staydown' part) refers back to Article 17(4)(b) and thus to the provision of "relevant and necessary information" about specific works, which is needed to prevent future uploads; *see also* Guidance on Article 17 CDSMD (n. 21) ("When implementing Article 17(4)(c), the Member States need to clearly differentiate the type of information rightholders provide in a 'sufficiently substantiated notice' for the removal of content (the 'take down' part of (c)) from the 'relevant and necessary information' they provide for the purposes of preventing future uploads of notified works (the 'stay-down' part of (c), which refers back to (b))").

that can be applied to a Content ID claim. Rightholders can instruct the system to:

(1) *track* the matching content's viewership statistics ("leave-up-and-track");
(2) *block* the content from being viewed ("takedown-staydown");[240] or
(3) *monetize* the content by displaying advertisements with it ("leave-up-and-get-paid").[241]

Match policies may also include directions from the rightholder on when Content ID should claim a video before anything else. Rightholders can set certain parameters, telling the system to automatically claim videos based on for instance geography ("when the UGC is uploaded from a certain country"), moment of upload ("when the UGC is uploaded during a specific time window"), match type ("when the UGC matches audio only, video only, or both"), or match amount ("when the UGC contains more than X minutes or Y percent of the reference file").[242] The fact that YouTube seemingly allows rightholders to set the threshold for the length or percentage of the uploaded video that must match the reference file to activate a Content ID claim is problematic. This is because rightholders are afforded the opportunity to set the threshold for a pre-defined blocking action *below* the legal standard of "manifestly infringing" content—i.e., "identical or equivalent" content—which in our view can only be associated with a high matching percentage across different parameters.[243] In our view, this is inconsistent with the CJEU's judgment in the *Poland* case discussed above.[244]

Monetization of matching UGC via Content ID occurs by placing advertisements against the matched content. In principle, the rightholder

240. *See* GOOGLE, SECTION 512 STUDY: REQUEST FOR ADDITIONAL COMMENTS 3 (Feb. 21, 2017), https://www.regulations.gov/comment/COLC-2015-0013-92487.

241. *Id.*

242. *Upload and Match Policies*, YOUTUBE https://support.google.com/youtube/answer/107129 (last visited Oct. 17, 2023); *see also* Article 17 Dialogue Recording, YouTube Presentation, *supra* note 223.

243. Admittedly, European copyright law does not provide a fixed number for the percentage or length of a video that has to match the reference file to be considered "manifestly infringing." Some Member States, however, have independently introduced numeric thresholds in their national implementation laws, for instance to indicate which uses are presumed to be authorized by law (e.g., presumed to be a quotation, parody, pastiche, etc.). For instance, Section 9(2)(1) of the German Act on the Copyright Liability of Online Content Sharing Service Providers provides that UGC that contains *less than half* of a work or several works by third parties is presumably authorized by law. Moreover, according to Section 9(2)(3) jo. Section 10(1)–(2) of the same Act, uses *up to 15 seconds* of a cinematographic work or an audio track are deemed to be "minor and are therefore presumably authorized by law.

244. *See supra* Section II.B.2.

receives all advertising revenue[245] produced by the claimed video that the uploader or creator of that video would have obtained absent the claim. This does not rule out, however, the possibility for rightholders to voluntarily share advertising revenues with uploaders, for instance when the upload is a cover song video and a music publisher wants to encourage fans to make such cover songs.[246] If a video is monetized, but the uploader decides to dispute the Content ID claim, YouTube will temporarily hold the advertising revenue from the video. Once the dispute is resolved, the platform will release the revenue to the appropriate party.[247] It is important to note, however, that nearly all Content ID claims go undisputed. For instance, in the first half of 2022, only 0.5% of the 750 million Content ID claims were disputed by the user-uploader. This is largely consistent with existing studies that have reported a relatively low usage of counternotice mechanisms, confirming the above-described lack of effectiveness of ex post complaint-and-redress mechanisms as a means to safeguard users' rights.[248]

According to YouTube's own statistics, more than 98.9% of all copyright actions taken on YouTube arise from Content ID users. Of those actions, monetization is clearly the most popular policy applied to claims: in the first half of 2022, over 90% of all Content ID claims were reportedly monetized, which resulted in the payment of $7.5 billion to rightholders in advertising revenue.[249] What is remarkable about these numbers is that while monetization is the preferred moderation action via Content ID, discussion on the topic and

245. It is not entirely clear from public information whether the rightholder, when the monetization policy is applied, *at all times* receives the *entire* advertising revenue, or that this may vary depending on, for example, whether the rightholder has the rights to both the video and audio or to the audio or video only; whether the rightholder merely owns the rights in a specific territory; whether there is co-authorship; etc. Based on publicly available information, however, we assume that rightholders receive the entire ad advertising revenue, but this is to be confirmed in interviews.

246. *Monetizing Eligible Cover Videos*, YOUTUBE https://support.google.com/youtube/answer/3301938?hl=en (last visited Oct. 17, 2023).

247. *Monetization During Content ID Disputes*, YOUTUBE https://support.google.com/youtube/answer/7000961 (last visited Oct. 17, 2023).

248. Urban, Karaganis & Schofield, *supra* note 105; Bar-Ziv & Elkin-Koren, *supra* note 105. Article 17(9) mandates *ex post* complaint and redress mechanisms, which should however be complementary to *ex ante* safeguards, such as restrictions to the scope of permissible filtering. *See supra* Section III.B; Martin Senftleben, *The Meaning of "Additional" in the Poland ruling of the Court of Justice: Double Safeguards – Ex Ante Flagging and Ex Post Complaint Systems – are Indispensable*, KLUWER COPYRIGHT BLOG (June 1, 2022), http://copyrightblog.kluweriplaw.com/2022/06/01/the-meaning-of-additional-in-the-poland-ruling-of-the-court-of-justice-double-safeguards-ex-ante-flagging-and-ex-post-complaint-systems-are-indispensable/.

249. YOUTUBE, COPYRIGHT TRANSPARENCY REPORT, *supra* note 222, at 1, 3–4.

its regulation was largely absent during the entire legislative process leading to the adoption of Article 17 of the CDSMD, as well as the subsequent Commission Stakeholder Dialogue and Guidance (where it is not even mentioned) and national implementation processes. This is all the more impressive since YouTube was the poster child for the "value gap" narrative that supported legislative intervention.[250]

Table 1 provides a summary of the functionality and affordances of tools in YouTube's Copyright Management Suite.

Table 1. Affordances of YouTube's Copyright Management Suite tools

	Webform	**Copyright Match**	**Content ID**
Functionality	Notice-and-takedown	*Ex ante* filtering + notice-and-staydown	*Ex ante* filtering + notice-and-staydown + monetization
Beneficiaries	Rightholders with infrequent needs or only a few copyright-protected works on the platform	Rightholders who experience a higher amount of reposting of their copyright-protected content	Rightholders with the most complex copyright management needs (enterprises)
Eligibility and number of users/uses	Open to everyone; *more than two billion channels*	Open to participants in the YouTube Partner Program and rightholders who demonstrated a short history of takedowns; *more than two million channels*	Open to a select group of partners (large, knowledgeable and resourceful players); *more than nine thousand partners*
Automation level	Low	Medium	High
Available copyright management actions	Content removal	Archive match (leave video up); File removal request; or Contact uploader	Track (leave video up); Block; or Monetize

250. Annemarie Bridy, *The Price of Closing the "Value Gap": How the Music Industry Hacked EU Copyright Reform*, 22 VAND. J. ENT. & TECH. L. 323, 323–58 (2020).

2. Meta's Facebook and Instagram

Facebook and Instagram, two of the biggest social media platforms around the globe owned by Meta,[251] pivot on the sharing of photos and videos (Instagram) and all types of media (Facebook) between families and friends. Both platforms offer to rightholders an internally developed copyright management and protection tool called Rights Manager.[252] At the core of Rights Manager is technology that allows for the matching of video, audio, and image materials,[253] on the basis of which rightholders can identify potentially infringing content and take actions accordingly.

To be approved for Rights Manager, users must submit an application. The exact acceptance criteria are not public, but users are reportedly evaluated based on, inter alia, historical behavior (they must not have posted content without permission from the valid copyright holder in the past); catalogue size (there should be a substantial body of work that requires a scaled tool); content eligibility;[254] the likelihood of a mass audience and of infringing user uploads; and historical use of the existing notice-and-takedown system to such a volume that this is demonstrably insufficient for the scale of matching.[255] According to Meta, smaller-sized users are generally not accepted to Rights Manager, as their needs are deemed to be met by the publicly accessible Copyright Report Form, a notice-and-takedown system.[256] Once access to Rights Manager is granted, rightholders must provide Meta with information like that required by YouTube for Content ID. This includes (1) reference files (video, audio, or

251. By the end of 2022, the platforms had respectively 255 and 250 million average monthly active users in the European Union only. *See* META, DIGITAL SERVICES ACT – INFORMATION ON AVERAGE MONTHLY ACTIVE RECIPIENTS IN THE EUROPEAN UNION (2023), https://transparency.fb.com/sr/dsa-report-feb2023/.

252. *See Introducing Rights Manager*, META (Apr. 12, 2016), https://www.facebook.com/formedia/blog/introducing-rights-manager (since 2016).

253. *Rights Manager*, META, https://rightsmanager.fb.com/ (last visited Mar. 3, 2023).

254. The user must have exclusive rights to the content, the reference content must be sufficiently distinct from other reference files (e.g., no screenshots from videos) and each piece of intellectual property must have an individual reference (e.g., compilations, mashups, countdown lists and reaction videos cannot be filed as a reference but must be separated into individual components). *See Content Eligible for Reference Files*, META, https://www.facebook.com/business/help/389834765475043 (last visited Mar. 3, 2023).

255. *Rights Manager Eligibility*, META (Aug. 16, 2023), https://www.facebook.com/business/help/705604373650775?id=237023724106807; *see also* Meta (then Facebook) Presentation, European Commission, Recording of the Fourth Meeting of the Stakeholder Dialogue on Article 17 of the Directive on Copyright in the Digital Single Market (Dec. 16, 2019) [hereinafter Article 17 Dialogue Recording, Meta Presentation].

256. Article 17 Dialogue Recording, Meta Presentation, *supra* note 255.

images),[257] (2) ownership information, including information about accounts that are authorized to publish the content ("white-listing"), (3) match rules (i.e., rightholders can determine what constitutes a match by setting parameters for the matching threshold, for instance in terms of temporal or percentual overlap)[258] and (4) the preferred actions to be taken in the event of a match ("match actions"). As noted for YouTube's Content ID, the fact that rightholders can set a low threshold for the overlap to trigger a match action seems problematic from a E.U. law perspective, at least when the match action is set to the ex ante blocking of uploads.

There are six match actions that can be attached to a match rule. Of the six, rightholders can instruct the system to automatically:

(1) *monitor* matching content for insights;

(2) *apply an ownership link* (i.e., place a banner on the matching content which links to a destination the rightholder designates, thereby using the UGC as a promotional opportunity);

(3) *collect advertising revenue* (only when the content is eligible for monetization); or

(4) *block* matching content from being viewed;

Moreover, rightholders can choose to manually:

(5) *review* matches and decide what to do at a later time; or

(6) *submit a copyright takedown report* from within the Rights Manager interface ("Copyright Report Form").[259]

Similar to Content ID, monetization of UGC via Rights Manager can be realized by placing in-stream advertisements in a video. Importantly, the "Collect ad earnings" option is not always available since the matching content itself must be eligible for monetization in the first place. This means that the content must have been uploaded by pages (not profiles) that comply with Facebook's or Instagram's Partner Monetization Policies and Brand Safety Controls; be at least 1 minute in length; and be published from a page enabled for in-stream ads (on Facebook) or from an account enabled for monetization

257. Most rightholders upload 'playable content' to Rights Manager, which is then fingerprinted by the tool. Meta provides the option to ingest fingerprinted or hashed content into the tool to only the "most trusted rightholders." *See* Article 17 Dialogue Recording, Meta Presentation, *supra* note 255.

258. *Match Rules in Rights Manager*, META https://m.facebook.com/help/711479543236965 (last visited Oct. 17, 2023).

259. *Rights Manager*, META, https://rightsmanager.fb.com/ (last visited Oct. 17, 2023); *see also Copyright Report Form*, FACEBOOK, https://www.facebook.com/help/contact/1758255661104383 (last visited Oct. 17, 2023).

with in-stream video ads (on Instagram). Matches that do not qualify for the monetization action are automatically set to the "monitor" match action.[260]

The amount of advertising revenue rightholders can earn via Rights Manager depends on the allocation of copyright ownership in the content item at issue. According to Meta, if rightholders exclusively own both the rights to the video and audio of audiovisual works, they are eligible to collect the *entire* ad earnings that the uploader would have received.[261] However, if rightholders merely own the rights to the video or audio, but not to both, they may only collect *half* of the ad earnings. If rightholders own rights in a specific geographic territory, they can collect the ad earnings generated by *views in that territory*. Lastly, if multiple rightholders share the rights to a work, the earnings should be *divided* among them.[262]

According to Meta, the monetization option within Rights Manager serves as an "authorization system," through which rightholders can "authorize the content on the platform and receive compensation for it."[263] In other words, it works like a license for the platform to use the copyrighted content in exchange for the collection of the advertisement revenue that accrues from it.

To the best of our knowledge, there are no publicly available data on copyright enforcement actions executed via Rights Manager. During the Commission Stakeholder Dialogue meeting of 16 December 2019, Meta (then Facebook) noted that "over 99% of the matches . . . are allowed to remain on the platform,"[264] which implies that claimed content is largely monitored, ownership-linked or monetized. However, the relative popularity of monetization via Rights Manager is unknown.[265]

If a user-uploader believes a match action applied by the rightholder is invalid, they can submit a dispute with the respective platform. When a user-uploader and rightholder continue to disagree on the lawfulness of the upload and match action, even after multiple phases of review and appeal, and the rightholder still wishes to uphold the claim, the internal dispute process within Rights Manager ends, and the rightholder must either release the claim or submit a takedown request via the Copyright Report Form.[266] Table 2 provides

260. *Content Eligible for Collect Ad Earnings Match Action*, META, https://www.facebook.com/business/help/985332875266274?id=237023724106807 (last visited Oct. 17, 2023).

261. *Collect ad Earnings in Rights Manager*, META, https://www.facebook.com/business/help/891090414760198?id=237023724106807 (last visited Oct. 17, 2023).

262. *Id.*

263. Article 17 Dialogue Recording, Meta Presentation, *supra* note 255.

264. *Id.*

265. *See infra* Section III.A.

266. *Rights Manager*, META, https://rightsmanager.fb.com/ (last visited Oct. 17, 2023).

a summary of the functionality and affordances of Facebook and Instagram's copyrighted content moderation tools.

Table 2. Affordances of Facebook and Instagram's Rights Manager tool and Copyright Report Form

	Rights Manager	**Copyright Report Form**
Functionality	*Ex ante* filtering + notice-and-staydown + monetization	Notice-and-takedown
Beneficiaries	Trusted rightholders with a substantial body of protected content on the platform	Any copyright holder
Eligibility and number of users/uses	Open to rightholders who can prove good historical behavior, own a large body of protected content that is likely to be used in new upload, and have a demonstrated need for a scaled tool (based on history of takedowns); *number of users unknown.*	Open to everyone; in the first half of 2022, more than 1.2 million copyright reports were submitted on Facebook and 450,000 reports were filed on Instagram[267]
Automation level	High	Low
Available copyright management actions	Monitor (automatically) Apply ownership link (automatically) Monetization (automatically) Block (automatically) Review match (manually) File takedown-report (manually)	Content removal

3. TikTok

Within only a few years, the short-form video-sharing platform TikTok (2016) owned by ByteDance has become a true social media sensation.[268] Via an app, users can create, share, and discover short video clips (of up to 10 minutes), varying from the well-known "TikTok dance" videos to videos that revolve around food, lip-syncing and social media challenges.

267. *Intellectual Property Transparency Report H1 2022*, META, https://transparency.fb.com/data/intellectual-property/notice-and-takedown/facebook/ (last visited Oct. 17, 2023).

268. On February 16, 2023, TikTok reported it has over 150 million active monthly users in the EU. *See Investing for our 150m strong community in Europe*, TIKTOK (Feb. 16, 2023), https://newsroom.tiktok.com/en-eu/investing-for-our-150-m-strong-community-in-europe.

Contrary to YouTube and Facebook/Instagram, TikTok does not seem to offer rightholders a sophisticated in-house copyright management tool. According to the platform's Intellectual Property Policy, rightholders who want to act against the upload of copies of their works can file Copyright Infringement Reports to request the removal of allegedly infringing content.[269] Additionally, they can manually fill out a form through which they can provide "relevant and necessary information"[270] about themselves and their copyrighted works, upon receipt of which TikTok "will do its best to ensure that [the] copyright work is made unavailable on TikTok in the EU."[271]

Thus, TikTok's current IP policy does not mention the possibility for rightholders to monetize allegedly infringing UGC by claiming uploaders' advertising revenue. This is not surprising, as the monetization of videos through advertising in itself is a relatively new phenomenon on the platform. Only in May 2022, TikTok introduced 'Pulse,' the platform's first advertising revenue sharing program through which highly popular individual creators, public figures and media publishers with over 100,000 followers can receive part of the revenue earned from advertisements run on their content.[272] Several creators have indicated, however, that the ad revenue sharing initiative has not exactly proven financially attractive to them, with pay-outs often not exceeding ten dollars.[273] In May 2023, TikTok launched an extension to the program named "Pulse Premiere," allowing "premium publishing partners" in "lifestyle & education, sports, and entertainment categories" to monetize their content on TikTok through "a revenue-sharing model"[274] and reportedly offering the

269. *Intellectual Property Policy*, TIKTOK (June 7, 2021), https://www.tiktok.com/legal/page/global/copyright-policy/en. Uploaders, in turn, can file counter-notifications via a Counter Notification Form.

270. *Id.*

271. *Id.*

272. *TikTok Pulse: Bringing Brands Closer to Community and Entertainment*, TIKTOK (May 4, 2022), https://newsroom.tiktok.com/en-us/tiktok-pulse-is-bringing-brands-closer-to-community-and-entertainment.

273. *See, e.g.*, Dan Whateley, Tanya Chen & Marta Biino, *How Much Tiktok Has Paid 8 Creators For Views Through its New Ad-Revenue Sharing Program Pulse*, INSIDER (Dec. 15, 2022), https://www.businessinsider.com/tiktok-pulse-ad-revenue-share-payments-creators-2022-11?international=true&r=US&IR=T; *see also* Alexandra Sternlicht, *Creators Report Extremely Low Earnings from TikTok's Ad Revenue Sharing Initiative*, FORTUNE (Jan. 24, 2023), https://fortune.com/2023/01/23/creators-report-extremely-low-earnings-from-tiktoks-ad-revenue-sharing-initiative/.

274. *Pulse Premiere: Connecting Brands with Premium Publisher Content*, TIKTOK (May 3, 2023), https://newsroom.tiktok.com/en-us/pulse-premiere.

publishers a 50% split.[275] While the monetization of creators' *own* content through advertising is slowly gaining ground on TikTok, it remains to be seen whether the platform will start offering copyright holders the option to monetize allegedly infringing UGC on the platform as well.

Table 3 provides a summary of the functionality and affordances of TikTok's copyrighted content moderation tools, excluding the content monetization option, for which no verifiable public data are available.

Table 3. Affordances of TikTok's copyright management tools

	Copyright Infringement Report	**Filtering system (ensuring unavailability of content on TikTok European Union)**
Functionality	Notice-and-takedown	filtering + notice-and-staydown
Beneficiaries	Any copyright holder	Any copyright holder
Eligibility and number of users/uses	Open to everyone; in the first half of 2022, 94,427 copyright removal reports were submitted[276]	No eligibility requirements but request *Ex ante* is reviewed for "accuracy, validity, and completeness" *number of users unknown.*
Automation	Low	Low
Available copyright management actions	Content removal	Blocking (automatically)

4. Third-Party Providers of Content Recognition Tools

Importantly, OCSSPs sometimes deploy—in addition to their in-house solutions—content recognition technologies provided by third-party vendors or service providers. Examples of entities offering such services are Audible Magic (audio and video) and Pex (audio and video). An essential difference with the in-house tools is that rightholders register their content directly with the third-party provider and not with the platforms. Depending on the platforms that implement the respective third-party technology—and thereby bear the costs of the service—the registered reference files are continuously

275. Alexandra Bruell, *TikTok Is Launching Ad Product for Publishers and Giving Them 50% Cut*, WALL ST. J. (May 3, 2023), https://www.wsj.com/articles/tiktok-is-launching-an-ad-product-for-publishers-and-giving-them-a-50-cut-cff0c9a0.

276. *Intellectual Property Removal Requests Report January 1, 2022 – June 30, 2022*, TIKTOK (Nov. 29, 2022), https://www.tiktok.com/transparency/en/intellectual-property-removal-requests-2022-1/.

checked against new content uploaded by the users of these platforms. In that sense, the third-party providers function as intermediaries between rightholders and platforms.

Audible Magic offers a fingerprinting-based content identification tool for audio and video files, which has been implemented by various OCSSPs such as Tumblr, Twitch, and Vimeo.[277] Until recently, Facebook/Meta also used the company's technology to match audio files in tandem with its proprietary Rights Manager.[278] However, this partnership appears to have been terminated in 2022.[279] Rightholders can register their media assets in Audible's Authoritative Registry, which contains millions of digital fingerprints. During the registration process, rightholders provide the service with digital or physical copies of their works, basic information about the work (song titles, artist names, record labels, show titles, season and episode numbers), and designated business rules, i.e., the preferred actions to be taken in the event of a match. The business rules are equivalent to the match policies and match rules/actions in Content ID and Rights Manager respectively,[280] and may differ for each platform. Hence, the same copyrighted work could be blocked at one platform and allowed or monetized at others.[281] Platforms that license Audible's database and recognition technology continuously translate newly uploaded UGC into machine-readable data which are forwarded to Audible Magic in the form of identification requests. Audible Magic processes these identification requests and returns the match results and business rules associated with the matched works back to the platforms. Notably, Audible Magic merely communicates the business rules to platforms. Their application—i.e., the actual blocking or monetization of content—is carried out by the platforms.[282]

Another fingerprinting-based solution available on the market is the "Attribution Engine" developed by technology company Pex. Its workings are

277. *Customers and Partners*, AUDIBLE MAGIC, https://www.audiblemagic.com/customers-partners/ (last visited Mar. 9, 2023).

278. QUINTAIS ET AL., *supra* note 25, at 260–61, 265.

279. Facebook is not included in the list of customers and partners on the Audible Magic website. *See Customers and Partners*, *supra* note 277.

280. Audible Presentation, European Commission, Recording of the Fourth Meeting of the Stakeholder Dialogue on Article 17 of the Directive on Copyright in the Digital Single Market (Dec. 16, 2019) [hereinafter Article 17 Dialogue Recording, Audible Presentation]; *see also What are business rules?*, AUDIBLE MAGIC, https://support.audiblemagic.com/hc/en-us/articles/7576145385619-What-are-business-rules- (last visited Oct. 17, 2023).

281. Audible's slogan is, notably, "Accelerating the distribution *and monetization* of content through the digital media ecosystem." AUDIBLE MAGIC, https://www.audiblemagic.com/ (emphasis added).

282. *Id.*

similar to Audible Magic's system. Rightholders register their content (videos and sound recordings) in the Asset Registry and set their preferred licensing policies: monetize, block, apply a customized license, or apply a free license. Fingerprints of UGC uploaded to platforms are transmitted to the Attribution Engine. In the event of a match, the platform is informed of the rightholder and reference file in question as well as the pre-defined policy for that content.[283] Unlike Audible Magic, Pex also offers a dispute resolution service to platforms, which is incorporated in the Attribution Engine. When uploaders do not agree with the applied policy, they can raise a dispute. The platform, in turn, can register the dispute via an application programming interface (API) with the Attribution Engine, which will then verify the dispute and inform the rightholder. If the rightholder decides to stick to the claim and the uploader still disagrees with it, parties have the option to request neutral copyright experts appointed by the World Intellectual Property Organization (WIPO) Arbitration and Mediation Center to review the case. Platforms complying with the WIPO panel's decision are indemnified by Pex from any legal risk. When the dispute resolution process is exhausted, parties are free to resort to other legal remedies.[284]

In sum, by identifying allegedly infringing UGC and communicating rightholders' preferred copyright enforcement actions to OCSSPs, third-party content recognition tools complement in-house tools and enable rightholders to moderate their content and enforce their rights on online platforms. It must be emphasized, however, that third-party providers do *not* apply the pre-defined match policies in practice. This remains at the discretion of the OCSSPs and rightholders.

C. HUMAN RIGHTS DEFICITS

The case studies above indicate that UGC monetization is a common practice among copyright holders, and at least the most popular on YouTube.[285] However, there are still significant gaps in what is known about monetization on OCSSPs.

Of the four platforms discussed above, YouTube has arguably made the most information available about its systems and tools. But even in this case,

283. *Attribution Engine*, PEX, https://pex.com/products/attribution-engine/ (last visited Oct. 17, 2023).

284. *Id.*; *see also Pex Partners with World Intellectual Property Organization Arbitration and Mediation Center Providing First Neutral Copyright Dispute Resolution Procedure*, PEX (Sept. 30, 2021), https://pex.com/blog/pex-partners-with-world-intellectual-property-organization-arbitration-and-mediation-center-providing-first-neutral-copyright-dispute-resolution-procedure/.

285. *See supra* Section II.B; Grosse Ruse-Khan, *supra* note 180.

much remains unclear. First, the data published by YouTube on UGC monetization via Content ID is aggregated globally; the lack of country-by-country data makes legal and empirical analysis of this phenomenon challenging given the territorial nature of copyright. Second, there is no detailed information on how rightholders can set thresholds for matching content. In the European Union at least, this is crucial for the fundamental rights assessment of such systems. Third, beyond the option of setting thresholds for matching content, there is little to no information on how these systems account for the individual context surrounding uses that are permitted by law, such as use for the purposes of parody, caricature, and pastiche protected under Article 17(7) of the CDSMD.

As regards Meta's Rights Manager, little information is available on the use of the tool in practice. Meta's 2022 Transparency Report on Intellectual Property[286] addresses only its notice-and-takedown system ("Copyright Report Form") but lacks data on Rights Manager. It is therefore unclear how many rightholders are currently using the tool, how often rightholders on Facebook and Instagram (in the European Union and worldwide) opt for monetization, how often disputes are raised, and so on. It is also not clear from publicly available information what happens to the ad earnings during an ongoing dispute.

Equally remarkable is TikTok's general lack of information on the workings of its ex ante copyright filtering system and its new monetization program, which could potentially form the basis for UGC monetization by rightholders in addition to the platform's core business of UGC licensing agreements with rightholders.[287]

Considering the above, one important conclusion of our analysis is the need for heightened transparency and increased access to content moderation data held by platforms. This is crucial not only for researchers to study the activity of platforms in a relatively unregulated content moderation space (here: monetization of UGC), but also to enable evidence-based policy making in this area.[288] In this respect at least, the rules in the DSA on statement of

286. *Intellectual* Property, META (Aug. 16, 2023), https://transparency.fb.com/data/intellectual-property/ (providing data from the first half of 2022).

287. Grosse Ruse-Khan, *supra* note 180 (noting the focus of TikTok on the licensing approach).

288. *See* SEBASTIAN FELIX SCHWEMER, CHRISTIAN KATZENBACH, DARIA DERGACHEVA, THOMAS RIIS & JOÃO PEDRO QUINTAIS, IMPACT OF CONTENT MODERATION PRACTICES AND TECHNOLOGIES ON ACCESS AND DIVERSITY 67 (2023) (reaching a similar conclusion on the basis of legal and empirical research).

reasons should provide much needed data on the workings of monetization systems.[289]

Based on the information that *is* publicly available, however, we can identify at least three human rights issues that arise from UGC monetization as currently implemented by platforms and rightholders: (1) the misappropriation of transformative UGC by third-party rightholders even though that content falls within the scope of copyright limitations that support freedom of expression (as discussed in Section III.C.1); (2) the encroachment upon UGC creator copyright which falls under the fundamental right to property in line with Article 17(2) of the CFR (as discussed in Section III.C.2); and the discriminatory treatment of UGC creators (as discussed in Section III.C.3).

1. Misappropriation of Freedom of Expression Spaces

The first issue concerns freedom of expression. As explained in Section II.B.2, the CJEU confirmed in the *Poland* decision that copyright limitations supporting freedom of expression, such as the right of quotation and the exemption of parody, constitute "user rights."[290] Despite our criticism of the judgment, this aspect is a positive development (as explained in Section II.B.3). Article 17(7) of the CDSMD confirms the elevated status of copyright limitations serving quotations, parodies, pastiches, etc. Article 17(7), second paragraph, even imposes an obligation on E.U. Member States to immunize these areas of freedom from filtering measures that could prevent the online publication and sharing of transformative UGC (as explained in Section II.B.3). Placing these developments in the context of the broader discussion on public domain preservation,[291] it can be said that the CJEU and the E.U. legislature have created robust areas of freedom by taking steps to keep these forms of transformative use outside the scope of copyright exclusivity. Relying on Yochai Benkler's public domain concept that includes sufficiently clear, "easy" cases of permitted use, it can be added that the court and the E.U. legislature have made these user rights part of "the range of uses of

289. *See supra* Section III.A.3.b.

290. CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶¶ 87–88; CJEU, 29 July 2019, case C-516/17, Spiegel Online, ¶¶ 50–54; CJEU, 29 July 2019, case C-469/17, Funke Medien NRW, ¶¶ 65–70.

291. For an overview and core arguments, see SENFTLEBEN, *supra* note 158, at 26–47, 280–283, 357–373.

information that any person is privileged to make absent individualized facts that make a particular use by a particular person unprivileged."[292]

Against this background, the corrosive effect of UGC monetization in this freedom of expression space clearly comes to light. Considering the "user right"-status in CJEU jurisprudence and the confirmation of the crucial importance of this habitat for freedom of expression in Article 17(7) of the CDSMD, European Union and Member State authorities are expected to preserve this part of the public domain. Leaving the further regulation of this freedom of expression space to online platforms and rightholders in the creative industry, the State outsources a central aspect of the obligation to safeguard human rights. The result is as problematic as it is predictable: without any clear legal basis for the appropriation and exploitation of transformative UGC, existing moderation systems allow copyright holders with access to these systems to monetize transformative UGC and usurp this freedom of expression space.

Under Articles 5(3)(d) and (k) of the InfoSoc Directive and Article 17(7) of the CDSMD, quotations for criticism or review, as well as parodies, caricatures, and pastiches, require neither the authorization of the copyright holder nor the payment of remuneration. In *VG Wort*, the CJEU clarified that any authorization which a rightholder may give in such a case "is devoid of legal effects"[293] because the statutory exemption places the use beyond the reach of any license or authorization. Hence, the fact that rightholders opting for monetization tolerate the use does not justify the channeling of advertising revenue to the creative industry. This is a negative consequence of the reliance on industry cooperation that is a central characteristic of the outsourcing tendency discussed *supra* Section II.A.2.

As an author of this Article, Martin Senftleben, has argued elsewhere, national legislation in the European Union may maximize the freedom of expression space called into existence by Articles 5(3)(d) and (k) of the InfoSoc Directive and Article 17(7) of the CDSMD by giving the open-ended concept of "pastiche" a broad meaning and bringing a wide variety of UGC, including the whole spectrum of memes and mash-ups, within its scope.[294] In such a case, it would make sense to combine the broadening of the freedom of expression space with the payment of fair compensation to collecting societies

292. Yochai Benkler, *Free as the Air to Common Use: First Amendment Constraints on Enclosure of the Public Domain*, 74 NYU L. REV. 354, 362–63 (1999). For a more detailed discussion of this public domain concept, see SENFTLEBEN, *supra* note 158, at 282–88.

293. CJEU, 27 June 2013, joined cases C-457/11 to C-460/11, VG Wort, ¶ 37.

294. Senftleben, *supra* note 137, at 316–23.

for the incorporation of protected third-party materials.[295] In this vein, Section 5(2) of the German Act on the Copyright Liability of Online Content Sharing Service Providers creates a non-waivable right of authors to "appropriate remuneration" that can only be assigned in advance to a collecting society.[296] Evidently, this solution aims at the creation of an additional revenue stream that flows directly to individual creators and not to exploiters of their works in the creative industry.[297]

The UGC monetization mechanism on online platforms is markedly different. First, it is unlikely to maximize the freedom of expression space on the basis of a flexible, broad interpretation of the pastiche exemption. To the contrary, the direct liability risk arising from Article 17(1) of the CDSMD will most probably lead to a restrictive interpretation of user rights and overblocking (see *supra* Section II.A.3). This only maximizes monetization options for larger rightholders participating in the system. Second, the advertising revenue is not paid to collecting societies. Hence, individual creators of third-party material woven into UGC cannot directly benefit from the additional source of income, either because they are not eligible to access the monetization system or because the rights revenue may not flow downstream to them from the enterprise rightholders that claim it.[298] In a nutshell: as currently designed and implemented, UGC monetization via online platforms fails to preserve the freedom of expression space for quotations, parodies, pastiches, etc. Instead, it constitutes a problematic exponent of the outsourcing of human rights obligations to online platforms and large rightholders.

From a human rights perspective, the main deficit caused in this scenario is that moderation systems of the biggest platforms allow large rightholders (eligible to access such systems) to appropriate and exploit as their own transformative UGC that E.U. law explicitly exempts from their permission on freedom of expression grounds. Once again: Article 17(7) of the CDSMD requires national laws and competent authorities to ensure that user rights prevail on platforms over a number of notice-and action measures, especially ex ante filtering. In practice, however, by outsourcing this human rights

295. Senftleben, *supra* note 8, at 154–62.

296. *See* Section 5(2) of the German Act on the Copyright Liability of Online Content Sharing Service Providers of 31 May 2021 (Federal Law Gazette I, 1204 (1215)), https://www.gesetze-im-internet.de/englisch_urhdag/index.html.

297. As to this important aspect of remuneration via collecting societies, see Senftleben, *supra* note 17, at 487.

298. This could be the case, *e.g.*, if a music producer claims monetization on the use of a musical composition and sound recording used in UGC but does not share the rights revenue with the composer and/or performer.

responsibility to platforms and copyright holders, the E.U. (and national) legislators allow the development of complex and opaque moderation systems. Such systems bundle together moderation options covered by the E.U. copyright regulatory framework—such as to block and remove UGC—with options that are largely unregulated, like monetization. In doing so, they hide behind the veil of complexity and opaqueness workings of the systems that potentially violate human rights.

In this case, the inability of content recognition tools to identify contextual exceptions (parody, quotation, pastiche, etc.), enables that vast amounts of UGC are checked for non-contextual matches with reference files, empowering rightholders to claim that content as their own. Depending on how the thresholds are set in platforms' match policies, rules, or actions, rightholders will automatically have the option to monetize that content. Since monetization is a popular moderation option, it is likely that UGC protected by user rights is unlawfully monetized on a regular basis. This provides yet another argument to limit the permissibility of preventive filtering to narrowly defined instances of "manifestly infringing" content (as discussed in Section II.B.2).

UGC monetization also has a subtle side effect that can be placed in the context of the above-described concealment strategies (as discussed in Section II.B.1). If rightholders do not opt to block or remove content, the illusion is created that no freedom of expression harm occurs. After all, the content remains publicly available. However, by monetizing such content, rightholders de facto claim ownership and exclusivity over transformative content permitted by law and therefore remove it from the freedom of expression zone that enables follow-on creation. If monetization goes unchallenged, as it often does, the UGC in question is then considered for all practical purposes as exclusively owned and commercially exploited by the rightholder who claims it, irrespective of its legal qualification as a permitted free use belonging to the public domain. This human rights deficit, in turn, leads to another, as it thwarts the ability of users to potentially exercise their own copyright over the UGC they created.

2. *Encroachment Upon the Fundamental Right to Copyright of UGC Creators*

A second and closely related problem is that of unremunerated user creativity and misappropriation of monetary rewards by copyright holders that have access to platforms' content recognition tools. This problem leads to a

human rights deficit because it encroaches on the remunerative dimension of the user's fundamental right to property—in this case intellectual property.[299]

UGC may consist of self-created material, public domain material, authorized and unauthorized takings of copyrighted material, or a combination of the previous. Oftentimes, UGC is the product of a certain amount of *creative effort*, whether in creating the content from scratch or in adapting existing works to create a new one. Unlike in U.S. law, there is no harmonized concept of "derivative work" in E.U. law; there is also no harmonized right of adaptation but rather a broadly interpreted right of reproduction.[300] This means that although exceptions or limitations for transformative uses—which may qualify as defenses or user rights[301]—are to some extent harmonized in Article 17(7) of the CDSMD, national laws diverge in how they deal with different types of transformative works. Naturally, the general EU standard of originality that requires a work to be the "author's own intellectual creation"[302] applies to transformative or derivative works.[303] Thus, considering the low threshold for originality in E.U. law, it is likely that UGC will often qualify as copyrighted content, meaning that the user-uploader should be entitled to monetize it, or at least part thereof.

If the UGC in question does qualify as copyrighted content, two main scenarios arise in relation to content where traces of a third-party reference file are matched in platforms' moderation systems. In both scenarios, as we shall see, the fact that the law outsources the regulation of monetization to platforms and larger rightholders, and the opaque practices of these players in

299. In the EU, copyright as a type of intellectual property right is protected in Article 17(2) CFR. For an analysis of this provision and its interpretation by the CJEU, see Jonathan Griffiths & Luke Mcdonagh, *Fundamental Rights and European Intellectual Property Law - The Case of Art 17(2) of the EU Charter* (2011), https://papers.ssrn.com/abstract=1904507; Peter Oliver & Christopher Stothers, *Intellectual Property Under the Charter: Are the Court's Scales Properly Calibrated?*, 54 COMMON MKT. L. REV. 517 (2017), https://papers.ssrn.com/abstract=3042530.

300. Generally on this topic, see HUGENHOLTZ & SENFTLEBEN, *supra* note 153. *See* Martin Senftleben, *Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham*, 51 INT'L REV. INTELL. PROP. & COMP. L. 751, 758–60 (2020) (discussing the "backdoor" harmonization strategy developed by the CJEU).

301. *See* Rendas, *supra* note 136.

302. *See, e.g.*, P. Bernt Hugenholtz & João Pedro Quintais, *Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?*, 52 INT'L REV. INTELL. PROP. & COMP. L. 1190 (2021) (discussing the modern application of this standard).

303. National laws may contain rules on whether the use of a pre-existing work without permission in a transformative work constitutes copyright infringement. This topic is not harmonized under E.U. law. However, this aspect should not influence the qualification of the transformative work as a copyrighted "work," provided it meets the originality standard in E.U. law.

this regard, leads to an encroachment upon UGC creators' fundamental right to property, including copyright as a form of intellectual property.[304] This interference with the right to property deprives UGC creators of legitimate rights revenue relating to their own creative contributions.

In the first scenario, third-party material incorporated in the transformative UGC and identified via a platform's content recognition tools is *not* covered by a transformative use exception, like parody, caricature, or pastiche. In this case, application of the legal rules would leave the rightholder of the pre-existing work with two options: (1) accept to share the revenue with the user-uploader or (2) request the content's takedown on the grounds of unauthorized reproduction of the original work in part and its making available online. However, current practices and systems of platforms enable the rightholder to appropriate the entirety of the monetization revenue if they decide to leave the content up.

In the second scenario, the third-party material incorporated in the transformative work *is* covered by an exception or limitation, thus enabling the use of the third-party material in the UGC. However, the content moderation system—which is incapable of determining the contextual lawful use—nevertheless allows rightholders to monetize exclusively. In this case, subject to specific national rules (e.g., statutory licensing regimes for these uses), it follows from our analysis of E.U. law that user-uploaders should receive the entirety of the available monetization revenue for what is effectively a commercial exploitation of their works.

To the best of our knowledge, in neither scenario the content moderation systems we examined contemplate monetization for the UGC creator as default options in their matching options, rules or procedures. Rather, existing systems are designed to empower legacy enterprise rightholders, leaving follow-on users-creators only with the option to complain ex post. Systems like Content ID and Rights Manager enable (mostly large) rightholders to immediately apply a monetization action when segments of UGC match with reference files, without requiring the rightholders to prove copyright infringement. The burden to argue for the applicability of an exception falls on users-creators, who must file disputes and counter-notices and therefore "fight" for the monetization of their content, assuming they are eligible for monetization in the first place under the platforms' policies.[305] As the data examined above suggests, these counterclaims rarely occur.

304. E.U. Charter of Fundamental Rights art. 17(2).

305. *See* Quintais, De Gregorio & Magalhães, *supra* note 198.

In our view, it is doubtful whether the new rules in the CDSM Directive and the DSA will improve this. Rather, they appear to create a particular challenging landscape for users, due to the potential for obfuscation of the complaint procedure offered by the overlapping application of legal regimes that outsource these procedures to platforms. As noted, monetization restrictions may be appealed—both via in-platform procedures and certified out-of-court dispute settlement bodies—under Articles 20 and 21 of the DSA (as discussed in Section III.A.3.b). However, this means that for the same item of UGC in one platform, part of the complaint and redress mechanism is regulated by Article 17(9) of the CDSMD (for UGC that is blocked or removed), whereas the other part is governed by Article 20 of the DSA (for UGC that is monetized).

The result, then, is that through the strategies of outsourcing and concealing, the fundamental right of UGC creators to be remunerated for their works through monetization is mostly eliminated or reduced to a potentially ineffective right of complaint.

3. Unequal Treatment and Discrimination of UGC Creators

As Martin Husovec and João Pedro Quintais argue, the way the design of Article 17 of the CDSMD favors large-scale or enterprise rightholders gives rise to the question of whether it violates the principle of equal treatment in Article 20 of the Charter.[306] Although not regularly applied in the context of intellectual property rights, the principle of equal treatment is no stranger to E.U. copyright law and to the remuneration interests of authors. In the area of collective licensing and fair compensation, for instance, the CJEU has clarified that Member States cannot impose fair compensation rules that would unjustifiably "discriminate between the different categories of economic operators marketing comparable goods covered by the private copying exception or between the different categories of users of protected subject matter."[307]

Other judgements regarding private copying, such as *VG Wort* and *Microsoft Mobile Sales International*, have further detailed the interpretation of equal treatment.[308] These cases suggest that the principle of equal treatment

306. Martin Husovec & João Quintais, *Too Small to Matter? On the Copyright Directive's Bias in Favour of Big Right-Holders*, *in* GLOBAL INTELLECTUAL PROPERTY PROTECTION AND NEW CONSTITUTIONALISM: HEDGING EXCLUSIVE RIGHTS (Tuomas Mylly & Jonathan Griffiths eds., 2021).

307. Copydan Båndkopi v. Nokia Danmark A/S [2015] ECLI:EU:C:2015:144, ¶ 33.

308. *Verwertungsgesellschaft Wort (VG Wort) v. Kyocera and Others* and *Canon Deutschland GmbH* and *Fujitsu Technology Solutions GmbH* and *Hewlett-Packard GmbH v. Verwertungsgesellschaft Wort (VG Wort)* [2013] ECLI:EU:C:2013:426, especially ¶¶ 73–79; *Microsoft Mobile Sales International*

may require lawmakers to establish "objective and transparent criteria where private ordering (based on existing rules) cannot guarantee this."[309]

How does this translate to Article 17 of the CDSMD? The argument of unequal treatment is not easy to make, since in this instance the risk of discriminatory treatment is buried "under layers of technicalities of E.U. copyright law."[310]

The departure point for this argument is that the default liability and licensing obligation dimension of Article 17 of the CDSMD—including the best efforts obligation to license—provides an advantage to large rightholders. To avoid liability, OCSSPs must proactively approach rightholders to obtain an authorization. To save time and resources, OCSSPs are likely to focus on easily identifiable and locatable rightholders (as discussed in Section II.A.1). In contrast, small rightholders and individual creators are disadvantaged as they must find and contact OCSSPs themselves, monitor the market, review use of their works by third parties, and approach each platform separately.[311]

Moreover, the liability exemption mechanism in Articles 17(4)(b) and (c) of the CDSMD places individual UGC creators, who rely on transformative use exceptions, at a disadvantage in relation to larger rightholders. Through the outsourcing and concealment strategies we have described, the design and control of UGC moderation systems are left to the private ordering of platforms and larger rightholders. The way these systems and practices have developed enables rightholders to appropriate UGC via monetization tools, despite it not being regulated in Article 17. Individual creators, which in theory benefit from freedom to engage in transformative uses and are entitled to copyright protection for their creative contributions to UGC, are in practice left with ineffective ex post complaint and redress tools.

As recognized by the court in *Poland*, the justification for tolerating the encroachment of Article 17(4) of the CDSMD upon freedom of expression is the objective of ensuring a high level of protection for copyright holders under Article 17(2) of the Charter. The court relied specifically on the appropriateness and effectiveness of the liability exemption mechanism to achieve this aim.[312] However, this is only the case if the mechanism is

Oy and Others v. Ministero per i beni e le attività culturali (MiBAC) and Others [2016] ECLI:EU:C:2016:717, especially ¶¶ 44–50.

309. Husovec & Quintais, *supra* note 306.

310. *Id.*

311. *See supra* Section II.A.1; *see also* Guidance Art. 17 CDSMD, *supra* note 18, at n 21); QUINTAIS ET AL., *supra* note 25.

312. CJEU Poland, ¶¶ 69, 84.

implemented and interpreted in light of fundamental rights, especially the freedom of expression and other fundamental rights of users.

What appears to have eluded the court is how inroads into the freedom of expression and the right to copyright of UGC creators may lead to their unequal treatment as compared to large-scale rightholders. Following the case law cited above, it is difficult to see how this different treatment between categories of rightholders is non-discriminatory, objectively justified, or socially just.

Importantly, the different pre-existing resources of large rightholders compared to individual UGC creators do not justify the uneven situation described above. In our view, the advantageous position of the first does not arise solely because they are better positioned financially and infrastructurally to internalize and benefit from the complex regime of Article 17 of the CDSMD. Rather, the unequal treatment of creators is a result of the uneven allocation of liability and obligations in the legal regime, and how they are implemented—or taken advantage of—by powerful actors. Naturally, as the court recognized in *Microsoft Mobile Sales International*, Member States cannot rely on market players implementing the provisions to correct discriminatory treatment,[313] especially where these might not be aligned with their corporate interests.

In its non-copyright case law, the court has stated that the "legislature's exercise of its discretion must not produce results that are manifestly less appropriate than those that would be produced by other measures that were also suitable for those objectives."[314] It follows that when implementing and interpreting Article 17 of the CDSMD, national legislators should consider measures to ameliorate or solve the unequal treatment described above.[315] Our analysis already points towards some solutions in this respect: clarification of the scope of matching as applying only to "manifestly infringing" UGC, which should have downstream positive effects in reducing abuses of the "monetization" option (in the sense that less matches will lead to less monetization claims and UGC appropriation); additional transparency as regards monetization actions on platforms; consideration of collective licensing schemes with non-waivable remuneration rights for individual creators; better recognition of the legal position of UGC creators; and design

313. *Microsoft Mobile Sales International*, *supra* note 308, at ¶ 49. As the court notes, ignorance of a problem with private contracts provides "no guarantee" that two groups in comparable situations will eventually be treated equally.

314. CJEU, 16 December 2008, case C-127/07, Arcelor Atlantique and Lorraine and Others, ¶ 59.

315. *See* Husovec & Quintais, *supra* note 306.

changes to platforms' systems to enable them to effectively monetize their own creative contributions to transformative UGC.

IV. CONCLUSION

A closer inspection of content moderation rules in the CDSM Directive and the DSA confirms a worrying tendency of reliance on industry cooperation and user activism to safeguard human rights. Instead of putting responsibility for detecting and remedying human rights deficits in the hands of the state, the E.U. legislature prefers to outsource this responsibility to private entities, such as OCSSPs, and conceal potential violations by leaving countermeasures to users. The risk of human rights encroachments is compounded by the fact that, instead of exposing and discussing the corrosive effect of human rights outsourcing in Article 17 of the CDSMD, the CJEU has rubberstamped this regulatory approach in its *Poland* decision.

As a welcome departure from the court-approved outsourcing and concealment scheme, Article 17(7) of the CDSMD obliges Member States to ensure that transformative UGC, containing quotations, parodies, pastiches, etc., survives content filtering and can be uploaded to online platforms. In addition, audit reports evolving from Article 37 of the DSA can offer important information for the European Commission to identify and eliminate human rights violations. Both exceptions to the rule of outsourcing to private entities, however, are currently underdeveloped. E.U. Member States have not consistently taken specific legislative action to shield transformative UGC from content filtering measures. The success of the DSA cascade of interventions—from audit reports to non-compliance decisions and fines—is unclear.

A case study shedding light on the largely uncharted territory of UGC monetization—the most common rightholder reaction to the detection of traces of protected third-party material in UGC—confirms that outsourcing and concealment strategies put human rights at risk. E.U. law gives platforms far-reaching autonomy to establish their own governance policies in relation to UGC monetization. The DSA only provides certain ex post mechanisms, such as transparency obligations and complaint systems. In this unregulated space, much remains unclear. The lack of country-by-country data makes a legal and empirical analysis of monetization practices challenging. In particular, there is no detailed information on how rightholders can set thresholds for matching content. In the European Union, this is crucial information for the fundamental rights assessment of such systems. Finally, beyond the option of setting thresholds for matching content, there is little information on how these systems account for the individual context surrounding uses that are

permitted by law, such as use for the purposes of parody, caricature, and pastiche protected under Article 17(7) of the CDSMD.

From a human rights perspective, the main deficit caused in this opaque environment is that content moderation systems established by the biggest platforms allow larger rightholders to appropriate and exploit as their own transformative UGC that, under E.U. law, is explicitly exempt from their permission on freedom of expression grounds. Outsourcing the human rights responsibility to platforms and copyright holders, the E.U. (and national) legislators allow the development of complex moderation systems with monetization options that disregard freedom of expression spaces. The current regime further leads to an encroachment on UGC creators' fundamental right to (intellectual) property, as it deprives them of the opportunity to benefit from revenues generated by their transformative content. Finally, we argue that the existing framework leads to the unequal treatment of UGC creators. This results from the uneven allocation of liability and obligations, and their practical implementation by powerful platforms and large rightholders.

To minimize the corrosive effect of monetization systems on the fundamental rights of creative users, it is important to reduce the impact of content filtering mechanisms—and related monetization options for rightholders—from the outset. Considering the outcome of the *Poland* decision, a first step in this direction is the confinement of content filtering to "manifestly infringing" UGC.[316] Going beyond the *Poland* ruling, however, it should also be considered to introduce collective licensing schemes with non-waivable remuneration rights for individual UGC creators. In addition, it is important to redesign monetization systems and make them inclusive, in the sense of offering creative users monetization opportunities that are equivalent to those available to large rightholders.

316. *Cf.* CJEU, 26 April 2022, case C-401/19, Poland v. Parliament and Council, ¶¶ 84–86; QUINTAIS ET AL., *supra* note 25.

AN ECONOMIC MODEL OF ONLINE INTERMEDIARY LIABILITY

James Grimmelmann[†] *& Pengfei Zhang*[††]

ABSTRACT

Scholars have debated the costs and benefits of internet intermediary liability for decades. Many of their arguments rest on informal economic arguments about the effects of imposing different liability rules on online platforms. Some scholars argue that broad immunity is necessary to prevent overmoderation; others argue that liability is necessary to prevent undermoderation. These are economic questions, but they rarely receive economic answers.

In this Article, we seek to illuminate these debates by giving a formal economic model of online intermediary liability. The key features of our model are *externalities*, *imperfect information*, and *investigation costs*. A platform hosts user-submitted content, but it does not know which of that content is harmful to society and which is beneficial. Instead, the platform observes only the probability that each item is harmful. Based on that knowledge, it can choose to take the content down, leave the content up, or incur a cost to determine with certainty whether it is harmful. The platform's choice reflects the tradeoffs inherent in content moderation: between false positives and false negatives, and between scalable but more error-prone processes and more intensive but costly human review.

We analyze various plausible legal regimes, including strict liability, negligence, blanket immunity, conditional immunity, liability on notice, subsidies, and must carry, and we use the results of this analysis to describe current and proposed laws in the United States and European Union.

TABLE OF CONTENTS

DOI: https://doi.org/10.15779/Z38WP9T772

 The authors presented previous versions of this Article at the BTLJ-BCLT From the DMCA to the DSA symposium, the Freedom of Expression Scholars Conference, and the Cornell Tech Research Lab for Applied Law and Technology. They would like to thank the participants, Aislinn Black, Banks Miller, Elettra Bietti, Daphne Keller, Sanketh Menda, Clint Peinhardt, and Vitaly Shmatikov.

† Cornell Law School and Cornell Tech, Cornell University.

†† School of Economic, Political, and Policy Sciences, The University of Texas at Dallas.

I. INTRODUCTION

In the scholarly literature on intermediary liability, economic claims are common: for instance, that platform liability creates chilling effects;[1] that platforms do (or do not) have the right incentives to self-police;[2] and that platform liability creates a trade-off between protecting free speech and

1. *See, e.g.*, Felix Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293, 304 (2013).

2. *See, e.g.*, Danielle Keats Citron & Mary Anne Franks, *The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform*, 2020 U. CHI. LEGAL F. 45, 52.

deterring abusive speech.[3] They are mostly informal policy arguments, not testable propositions. It is not clear when two claims conflict, or when they can coexist. Indeed, it is often not even clear whether two authors are making the same claim or different claims.

We do not propose to resolve any of these disputes. Instead, we aim to clarify the terms of the debate. In this Article, we recast arguments about online intermediary liability into a common language—the language of microeconomics. We give an economic model of online intermediary liability, with equations and diagrams. We see six significant benefits to legal scholarship from having such a model—benefits that in turn can help lead to better and more appropriately calibrated intermediary-liability law.

First and most fundamentally, modeling promotes communication. A suitable model can serve as a common framework for scholars to compare and contrast arguments. Our taxonomy of liability regimes reduces the at-times bewildering array of arguments about the proper scope of intermediary liability into a (we hope) orderly structure that makes it straightforward to see how different claims relate.

Second, modeling promotes intuition. A good model can bring out the consequences of a course of conduct or make plain why parties behave the way that they do. There are several common patterns in intermediary-liability law that have simple and vivid expressions in our model.

Third, modeling promotes visualization. We have attempted to provide a simple and memorable visual shorthand for every moving part in our model and every interesting effect of a legal regime. For example, we hope that even if you take nothing else away from this Article, you will have a clear visual sense for why a platform might either overmoderate or undermoderate even in the absence of liability.

Fourth, modeling promotes rigor. The process of writing down a model forces one to make one's assumptions explicit. Reasoning through a model's consequences requires a close examination of each claimed effect. In the course of working through our own model, we learned a lot about how arguments for and against intermediary immunity work, and this Article conveys some of what we have learned.

Fifth, modeling promotes proof. Given a set of explicit assumptions, it is possible to show rigorously whether particular conclusions follow. For

3. *See, e.g.*, Daphne Keller, *Toward a Clearer Conversation About Platform Liability*, KNIGHT FIRST AMEND. INST. (Apr. 6, 2018), https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability.

example, we demonstrate that under our assumptions, strict liability consistently results in overmoderation. Of course, the real world is not required to comply with a proof about a model. But proofs like these make models more useful because they help pin down the predictions made by the model.

Sixth, modeling promotes empiricism. This is not an econometric Article; there are no datasets and very few numbers. But a model like ours helps identify the right econometric questions to ask. We hope that it provides a roadmap for future empirical work.

This Article has seven Parts, including this Introduction and a brief Conclusion. Part II surveys previous work in this space. Part III describes the model in formal detail. Parts IV and V analyze a variety of liability regimes in detail. Part VI shows how various current and proposed laws map on to those liability regimes. Part VII concludes and discusses possible extensions of the model.

II. BACKGROUND AND RELATED LITERATURE

At the most general level, this Article asks whether a regime of *online intermediary liability* or *online intermediary immunity* is economically optimal. A liability regime requires platforms to compensate the victims of harmful content posted by the platform's users. An immunity regime, by contrast, does not impose such liability on platforms—even when the users themselves might be held liable for posting the harmful content. There are numerous variations and hybrids of these two basic regimes, but this dichotomy is the fundamental legal and policy question of online intermediary law.

The literature on the economic analysis of law and the legal scholarship on intermediary liability are both immense. In the former, we draw particularly on the tradition of formalizing liability rules started by John Prather Brown,[4] and on the standard distinction between strict liability and negligence.[5] In this literature, the usual goal of the liability system is to promote efficiency, which means minimizing total social costs. The focus is on the effect of different liability rules on incentives for taking precautions to reduce risk. One key insight is that if only injurers influence risks, both strict liability and negligence can induce them to take optimal care.

4. John Prather Brown, *Toward an Economic Theory of Liability*, 2 J. LEGAL STUD. 323 (1973).

5. *See id.*; *see also* Steven Shavell, *Strict Liability Versus Negligence*, 9 J. LEGAL STUD. 1 (1980). *See generally* STEVEN SHAVELL, ECONOMIC ANALYSIS OF ACCIDENT LAW (2009) (summarizing literature).

The economic theory of liability underpins the product liability regime, where a manufacturer or seller might be held liable for harm caused by a defective or unsafe product.[6] This theory takes into account the relationship between liability, market price, and firms' profit-maximizing production. One of the key insights of this literature is that whether and how to impose liability depends on the characteristics of the product and the information available to consumers. For example, a strict liability rule would be more appropriate when it is difficult to test for product safety. In general, strict liability is more efficient than negligence, as it either results in prices that induce optimal production or induces consumers to purchase the optimal quantity. One of the key purposes of our model is to square this conclusion with the widespread argument in the legal literature that strict liability is particularly inappropriate for platforms.

The economic literature on online intermediary liability, however, is still in its infancy. The most detailed treatment is a student note by Matthew Schruers (now the president of the Computer and Communications Industry Association).[7] He analyzes four legal regimes—negligence, notice-based liability, strict liability, and immunity—in a model where the intermediary can vary its level of care. Although the moving parts in his model are different than ours, it is an illuminating analysis of the tradeoffs involved in imposing intermediary liability. His most trenchant insight is that notice reduces the effort required for the platform to achieve a given level of care,[8] an approach that informs our own information-based treatment of liability on notice. He also notes the essential parallel between liability on notice and strict liability and the chilling effect of strict liability on online speech.

In a more recent article, Xinyu Hua and Kathryn Spier extend the product-liability framework to a two-sided platform that enables interactions between sellers and buyers.[9] By either raising its prices or investing in screening, the

6. *See* Koichi Hamada, *Liability Rules and Income Distribution in Product Liability*, 66 AM. ECON. REV. 228 (1976); A. Mitchell Polinsky, *Strict Liability vs. Negligence in a Market Setting*, 70 AM. ECON. REV. (PAPERS & PROC.) 363 (1980); William M. Landes & Richard A. Posner, *A Positive Economic Analysis of Products Liability*, 14 J. LEGAL STUD. 535 (1985); A. Mitchell Polinsky & Steven Shavell, *The Uneasy Case for Product Liability*, 123 HARV. L. REV. 1437 (2010). *See generally* Andrew F. Daughety & Jennifer F. Reinganum, *Economic Analysis of Products Liability: Theory*, *in* RESEARCH HANDBOOK ON THE ECONOMICS OF TORTS 69 (Jennifer H. Arlen ed., 2013) (summarizing literature on economics of product liability).

7. Matthew Schruers, Note, *The History and Economics of ISP Liability for Third Party Content*, 88 VA. L. REV. 205 (2002).

8. *Id.* at 237–39.

9. Xinyu Hua & Kathryn E. Spier, *Holding Platforms Liable*, HKUST BUSINESS SCHOOL RESEARCH PAPER NO. 2021-048 (2022), https://ssrn.com/abstract=3985066 or http://dx.doi.org/10.2139/ssrn.3985066.

platform wants to keep safe sellers while deterring harmful sellers. They argue that whether to impose liability on the platform depends on whether the sellers are judgement-proof. If the sellers have deep pockets, then intermediary immunity is optimal. If the sellers are instead judgement-proof, then intermediary liability is necessary, and imposing residual liability on the platform improves social welfare.[10]

Intermediary immunity, as codified by § 230, has been credited for promoting the growth of the internet.[11] Many authors argue that § 230 was a response to concerns about the negative impact of lawsuits on online service providers, and that § 230 strikes a balance between free speech and safety.[12] Some scholars observe that platforms do moderate in the absence of any liability.[13] They argue that even intermediary immunity might lead to the over-removal of content by the platform (through user account termination, shadow banning, or "collateral censorship").[14]

At the same time, other scholars criticize the current shape of § 230.[15] Danielle Citron has argued that online platforms facilitate and amplify harassment and hate speech.[16] She and others argue that courts have given § 230 an overly broad interpretation and that this broad immunity provides excessively strong incentives to allow or encourage online materials to go unmoderated.[17] Consequently, the broad immunity fails to protect the victims

10. *See* Yassine Lefouili & Leonardo Madio, *The Economics of Platform Liability*, 53 EUR. J.L. & ECON. 319 (2022).

11. Anupam Chander, *How Law Made Silicon Valley*, 63 EMORY L.J. 639, 650–57 (2013).

12. *See, e.g.*, Paul Ehrlich, *Communications Decency Act 230*, 17 BERKELEY TECH. L.J. 401, 411–13 (2002); Cecilia Ziniti, *Optimal Liability System for Online Service Providers: How Zeran v. America Online Got It Right and Web 2.0 Proves It*, 23 BERKELEY TECH. L.J. 583, 584 (2008); Matt C. Sanchez, *The Web Difference: A Legal and Normative Rationale Against Liability for Online Reproduction of Third-Party Defamatory Content*, 22 HARV. J.L. & TECH. 301, 317–19 (2008); Jeff Kosseff, *Defending Section 230: The Value of Intermediary Immunity*, 15 J. TECH. L. & POL'Y 123, 144–45 (2010).

13. *See, e.g.*, Eric Goldman, *Online User Account Termination and 47 U.S.C. § 230(c)(2)*, 2 U.C. IRVINE L. REV. 659, 670–71 (2012).

14. *See, e.g.*, Wu, *supra* note 1, at 296–97.

15. *See generally* Joel R. Reidenberg, Jamela Debelak, Jordan Kovnot & Tiffany Miao, *Section 230 of the Communications Decency Act: A Survey of the Legal Literature and Reform Proposals*, FORDHAM L. LEGAL STUDIES RSCH. PAPER NO. 2046230 (2012), https://ssrn.com/abstract=2046230 (surveying scholarly analyses of Section 230 and proposed reforms).

16. DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 61 (2014).

17. *See id.*; *see also* Jennifer Benedict, *Deafening Silence: The Quest for a Remedy in Internet Defamation*, 39 CUMB. L. REV. 475, 493 (2008); Colby Ferris, *Communication Indecency: Why the Communications Decency Act, and the Judicial Interpretation of It, Has Led to a Lawless Internet in the Area of Defamation*, 14 BARRY L. REV. 123, 136 (2010).

of online abuse with no recourse against the platforms, whose profit maximizing business models facilitate the harmful activities.[18]

There is substantial literature advocating the reform of § 230, but scholars disagree on the appropriate form of intermediary liability. Some papers seek to refine the scope of intermediary immunity by taking a multi-factor approach. For example, some authors suggest that courts should consider the level of editorial control exercised by the platform and the harm caused by defamatory statements when determining immunity in each case (though it is not clear how the court should weigh these different factors).[19] Other authors discuss strict liability for certain kinds of harms;[20] in particular, Nancy Kim suggests treating intermediary liability like product liability.[21] And some authors suggest applying criminal liability to sites that have been facilitating and profiting from illegal activities.[22]

Other scholars propose forms of conditional immunity. Danielle Citron and Benjamin Wittes have argued that the platforms should enjoy immunity only if they can prove that they took reasonable efforts to address online abuse, and lawmakers should specify the obligations for this duty of care.[23] Doug Lichtman and Eric Posner suggest a conditional immunity rule in which ISPs would be held liable for infringing content only if they fail to implement reasonable measures to prevent or deter infringement.[24] Caitlin Hall proposes

18. *See, e.g.*, Danielle K. Citron & Benjamin Wittes, *The Problem Isn't Just Backpage: Revising Section 230 Immunity*, 2 GEO. L. TECH. REV. 453 (2018); Ann Bartow, *Internet Defamation as Profit Center: The Monetization of Online Harassment*, 32 HARV. J.L. & GENDER 383 (2009).

19. *See, e.g.*, Jae Hong Lee, *Batzel v. Smith & Barrett v. Rosenthal Defamation Liability for Third-Party Content on the Internet*, 19 BERKELEY TECH. L.J. 469, 493 (2004); Vanessa S. Browne-Barbour, *Losing Their License to Libel: Revisiting § 230 Immunity*, 30 BERKELEY TECH. L.J. 1505, 1553–56 (2015).

20. *See, e.g.*, Mark MacCarthy, *What Payment Intermediaries Are Doing About Online Liability and Why It Matters*, 25 BERKELEY TECH. L.J. 1037, 1043 (2010) (discussing the payments industry as a source of moderation that responds to legal incentives).

21. Nancy S. Kim, *Imposing Tort Liability on Websites for Cyberharassment*, 118 YALE L.J. POCKET PART 115, 117 (2008).

22. *See* Shahrzad T. Radbod, *Craigslist—A Case for Criminal Liability for Online Service Providers*, 25 BERKELEY TECH. L.J. 597 (2010).

23. *See* Danielle Keats Citron, *How to Fix Section 230*, 103 B.U. L. REV. (forthcoming 2023); Citron & Wittes, *supra* note 18.

24. *See* Doug Lichtman & Eric A. Posner, *Holding Internet Service Providers Accountable*, 14 SUP. CT. ECON. REV. 221, 251–54 (2006). Lichtman and Posner focus on ISPs' role in "the creation and propagation of worms, viruses, and other forms of malicious computer code." *Id.* at 221. This is still a form of content moderation; the content at issue is harmful to computer systems.

an immunity conditioned on the display of rating labels that alert internet users of the credibility of information posted on the sites.[25]

Other authors compare the different immunity regimes of § 230 and § 512.[26] Those who highlight what they see as the advantages of the DMCA-style regime note the ability of the victims to prevent further harm through a notice-and-takedown system, the coordination of internet service providers (ISPs) in removing harmful materials, and the administrative ease of transition. On the other hand, those who worry about applying a notice-and-takedown system to all user-generated content highlight the possible chilling effects on free speech. One important line of work discusses graduated-response, in which ISPs issue warnings and impose penalties on users who engage in infringing behaviors and use the termination of service as a threat to induce lawful behavior.[27]

Other authors compare the intermediary liability regime in the European Union with that in the United States. Daphne Keller examines the relationship and the tension between online platforms' liability in Europe (e.g., the E-Commerce Directive) and the European Union's General Data Protection Regulation (GDPR).[28] Miriam Buiten, Alexandre De Streel, and Martin Peitz examine the European Union's Digital Single Market strategy, and they argue that the current E.U. liability framework is inadequate for dealing with the challenges of online content moderation.[29] They claim that the absence of "Good Samaritan" protection in the E.U. e-Commerce Directive creates

25. Caitlin Hall, Note, *A Regulatory Proposal for Digital Defamation: Conditioning § 230 Safe Harbor on the Provision of a Site "Rating,"* STAN. TECH. L. REV. N1 (2008).

26. *See* Mark A. Lemley, *Rationalizing Internet Safe Harbors*, 6 J. ON TELECOMM. & HIGH TECH. L. 101, 102–04 (2007); Sarah Duran, *Hear No Evil, See No Evil, Spread No Evil: Creating a Unified Legislative Approach to Internet Service Provider Immunity*, 12 U. BALT. INTELL. PROP. L.J. 115, 118–33 (2004); Olivera Medenica & Kaiser Wahab, *Does Liability Enhance Credibility: Lessons from the DMCA Applied to Online Defamation*, 25 CARDOZO ARTS & ENT. L.J. 237, 256–62 (2007); Cyrus Sarosh Jan Manekshaw, *Liability of ISPs: Immunity from Liability Under the Digital Millennium Copyright Act and the Communications Decency Act*, 10 COMPUT. L. REV. & TECH. J. 101, 110–32 (2005); Jonathan Band & Matthew Schruers, *Safe Harbors Against the Liability Hurricane: The Communications Decency Act and the Digital Millennium Copyright Act*, 20 CARDOZO ARTS & ENT. L.J. 295, 296–319 (2002).

27. *See generally* Peter K. Yu, *The Graduated Response*, 62 FLA. L. REV. 1373 (2010); Annemarie Bridy, *Graduated Response American Style: "Six Strikes" Measured Against Five Norms*, 23 FORDHAM INTELL. PROP., MEDIA & ENT. L.J. 1 (2012); Rebecca Giblin, *Evaluating Graduated Response*, 37 COLUM. J.L. & ARTS 147 (2014).

28. Daphne Keller, *The Right Tools: Europe's Intermediary Liability Laws and the EU General Data Protection Regulation*, 33 BERKELEY TECH. L.J. 287, 351–61 (2018).

29. Miriam C. Buiten, Alexandre De Streel & Martin Peitz, *Rethinking Liability Rules for Online Hosting Platforms*, 28 INT'L J.L. & INFO. TECH. 139 (2020).

perverse incentives for platforms not to monitor online activity, thus undermining self-regulation.[30]

There is also a strong empirical literature on content moderation.[31] Collectively, this research suggests that platforms tend to have a bias towards over-removal.[32] In 2005, Jennifer Urban and Laura Quilter presented the first set of descriptive statistics on the notice-and-takedown process under DMCA § 512.[33] They found that corporations and business entities were the primary senders of notices, a majority of the notices were sent for competition purposes, one third of the notices were questionable regarding the validity of the copyright infringement claim, and very few individual users responded with a counter-notice.[34] In a follow-up study, Urban, Joe Karaganis, and Brianna Schofield emphasize the role of automation in sending complaints, and compare how the automated notices differ from the manual notices by small right holders.[35] More recently, Daniel Seng compiled a larger dataset and gave more detailed statistics, questioning the validity of many takedown notices, especially those generated by automated systems.[36] Meanwhile, Jonathon Penney surveyed 1,296 panelists with hypothetical scenarios on receiving a takedown notice, and his findings indicate some chilling effects of the policy.[37] Respondents broadly reported being less likely in future not only to share the same content again, but also to share content they themselves had created

30. *Id.* at 163–66.

31. *See generally* Daphne Keller & Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM 220 (Nathaniel Persily & Joshua A. Tucker eds., 2020) (surveying information released by platforms and independent research).

32. Daphne Keller, *Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List*, STAN. L. SCH.: CTR. FOR INTERNET & SOC'Y (Feb. 8, 2021), https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

33. Jennifer M. Urban & Laura Quilter, *Efficient Process or "Chilling Effects"? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA COMPUT. & HIGH TECH. L.J. 621 (2005).

34. *Id.* at 649–80.

35. *See generally* JENNIFER M. URBAN, JOE KARAGANIS & BRIANNA SCHOFIELD, NOTICE AND TAKEDOWN IN EVERYDAY PRACTICE (2nd version 2017).

36. Daniel Seng, *Copyrighting Copywrongs: An Empirical Analysis of Errors with Automated DMCA Takedown Notices*, 37 SANTA CLARA HIGH TECH. L.J. 119 (2020).

37. Jonathon W. Penney, *Privacy and Legal Automation: The DMCA as a Case Study*, 22 STAN. TECH. L. REV. 412 (2019).

(seventy-two percent).[38] Only thirty-four percent said they would send a counter-notice or challenge a takedown they believed was wrong or mistaken.[39]

III. AN ECONOMIC MODEL OF PLATFORM CONTENT MODERATION

There are two distinctive features of platform liability for harmful third-party content. The platform has *imperfect information* about which content is harmful and which is not, and content can have *positive externalities* not captured by the platform itself. These two features, taken together, mean that holding the platform liable for the harmful content it carries can go wrong. Because the platform cannot perfectly distinguish harmful from harmless content, and because it does not internalize the full benefits from the harmless content, the threat of liability can cause the platform to overmoderate, removing too much harmless content along with the harmful content.

A. THE MODEL

The first essential feature that makes intermediary liability distinctive is that a platform has imperfect information about the content that it hosts. Some content is harmful, and other content is not, but they look the same on first glance. A court decides whether a statement is legally defamatory after fact discovery, motion practice, and a trial; a platform does not have the time, the resources, or the power to conduct a full civil lawsuit on every post. A court awards damages in the fullness of time, on relatively complete information; a platform must act now, with radically incomplete information.

Formally, users submit discrete items x_1, x_2, x_3 ... of **content** to a platform. Each of these items is either **harmful** or **harmless**, and the platform can either **host** or **remove** each item. The essential feature of the model is that platform *does not know* whether each item is harmful or not. Instead, it observes the probability $\lambda(x)$ that item x is harmful, so it must make its hosting or removal decision under conditions of uncertainty.

38. *Id.* at 446–47.
39. *Id.* at 451.

Figure 1: Probability of harm

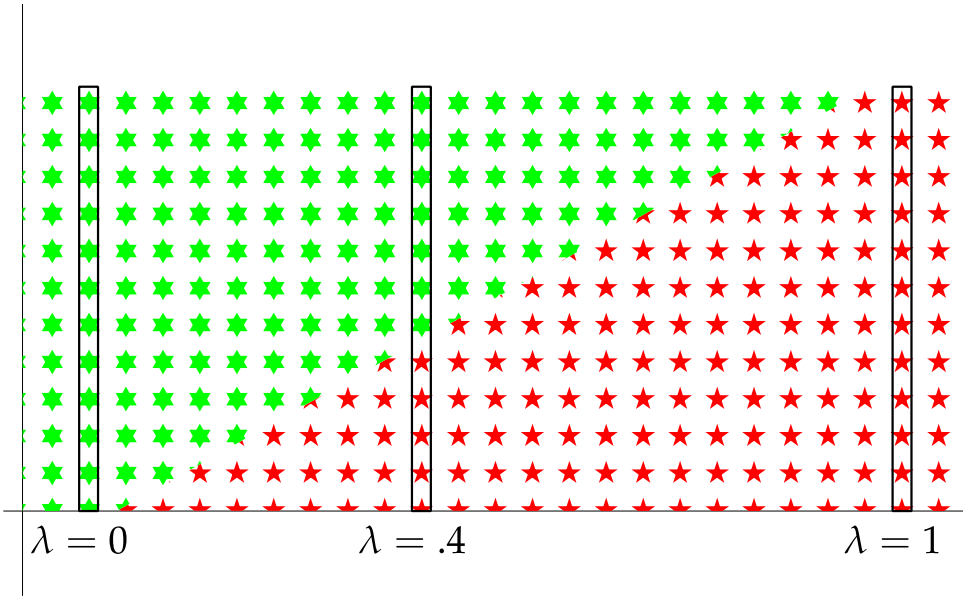


Figure 1 illustrates the platform's imperfect information. Think of the content presented to the platform as being divided into buckets. The platform knows what fraction (the probability $\lambda(x)$) of the content in each bucket is harmful (red five-pointed stars) or harmless (green six-pointed stars). But it does not know which specific items of content (individual stars) are harmful or harmless. If the platform hosts an item x, it has the following consequences:

- The **platform** receives some revenue $p(x)$.
- **Society** realizes some benefits $s(x)$.
- If the item is harmful, a third-party **victim** suffers harm $h(x)$.

If the platform removes the item, then the revenue, social benefits, and third-party harms are all 0.

Note that the social benefits of content $s(x)$ are known with certainty, and so are the harms $h(x)$ *if they happen*. Overall social welfare is therefore $s(x) - h(x)$ for harmful items, and $s(x)$ for harmless ones. Thus, the *expected* social welfare from hosting an item of content is $s(x) - \lambda(x)h(x)$: the known benefits minus the expected harms.

In general, $p(\cdot)$, $s(\cdot)$, $h(\cdot)$, and $\lambda(\cdot)$ could be arbitrarily complicated functions that account for an arbitrarily large number of features of each item of content. So while this expression is almost tautologically simple, it does not say much about how to draw useful lines between different kinds of content.

Therefore, we simplify the model by collapsing all content to a *single axis*. Imagine the content submitted by users to a platform arranged on a spectrum from worthwhile to worthless. At one end, the content is entertaining and informative—cat pictures and civics lessons. At the other end, the content is stomach-churning or worse—gross-out pictures and badly-written spam. A platform sets its moderation policy by deciding where along this spectrum to draw the line.

More formally, we assume that each item content falls within the one-dimensional interval from 0 to $x_{\max}$, where 0 is the "good" end and $x_{\max}$ is the "bad" end. Then as x increases:

- Content is less profitable to the platform: $p(x)$ decreases.
- Content is less beneficial to society: $s(x)$ decreases.
- The harm (if it happens) is fixed: h is a constant.
- Content is more likely to be harmful: $\lambda(x)$ increases.

We assume that $s(x) > p(x)$, i.e., all content has some positive spillover benefits for society that the platform does not capture.[40] We do not assume that $p(x) > 0$ or $s(x) > 0$: it is possible that some content is negative-value even if it is not harmful to third parties. (An example is spam, which is costly for the platform to host and has infinitesimal spillover benefits for anyone else.)

To make the model interesting, and to eliminate some annoying corner cases, we assume that the most innocuous content is profitable for the platform ($p(0) > 0$), beneficial to society ($s(0) > 0$), and known with certainty to be harmless, i.e., $\lambda(0) = 0$. Similarly, we assume that most problematic content is known with certainty to be harmful ($\lambda(x_{\max}) = 1$) and that harmful content is unambiguously bad for society, i.e., $h > s(x)$ for all x. These conditions ensure that some content is definitely good for society and some content is definitely bad for society, so that there is a real interest in treating them differently.

40. Any negative spillovers are separately accounted for by the harm h.

Figure 2: A one-dimensional model of moderation

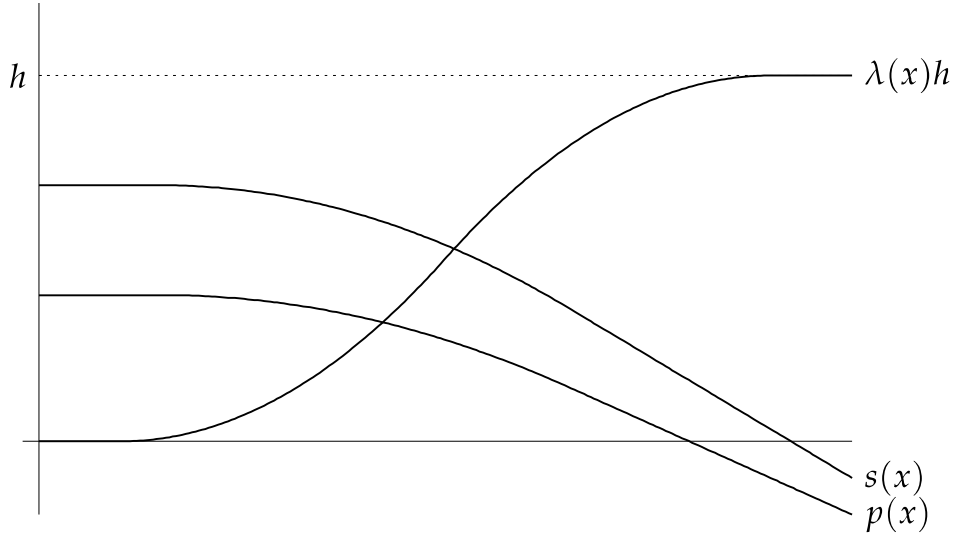


Figure 2 illustrates the essential model. The platform-revenue and social-benefit curves $p(x)$ and $s(x)$ start off positive and drop off. The expected-harm curve $\lambda(x)h$—the probability that content is harmful times the harm if it is—starts at 0 and rises to h.Given these assumptions, it is easy to see that content further to the left is always better ex ante and content further to the right is always worse. If $x_1 < x_2$, then x_1 is more profitable to the platform, better for society, and less likely to be harmful ex ante. Of course, if x_2 turns out to be harmless and x_1 is not, then x_2 might be better ex post, but from behind the veil of probabilistic ignorance, x_1 is the better prospect ex ante.

It follows that a rational moderator who is concerned with maximizing benefits and profits and minimizing harms will set a **moderation threshold** x^*. It will leave up all content x with $x < x^*$, and remove all content x with $x > x^*$. There is no circumstance under which it makes sense for the moderator to take down content x and leave up content y where $x < y$, because it would always be better to leave up x and take down y instead.

Any choice of x^* trades off false positives and false negatives. A low threshold means that more harmless content will be removed; a high threshold means that more harmful content will stay online. We tolerate some harmful content because it is indistinguishable ex ante from harmless content. The choice of x^* incorporates the moderator's judgments about the *acceptable risk of harm*.

This imperfect information is central to our model, and we believe that it is a pervasive fact of content moderation. While the users who upload content and the victims who are harmed by it may be in a better position to know whether content is harmful, platforms and regulators operate from a position of comparative ignorance.

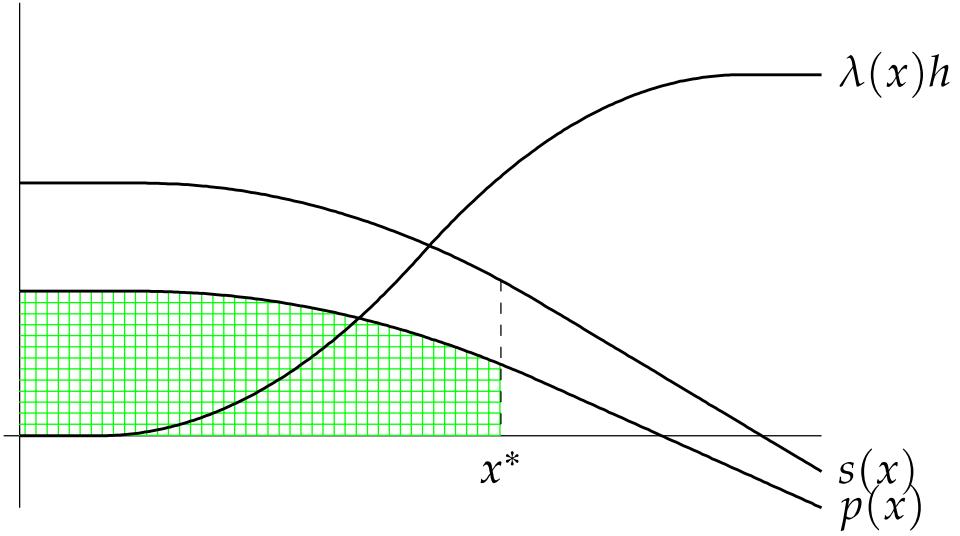
The overall harm here is a *statistical* consequence of a given choice of x^*. If the platform could perfectly distinguish harmful and harmful content, it could choose to host only the harmless content. (Indeed, we will shortly consider an extension of the model under which this distinction is possible, albeit at a cost.) But the point of the current model is that the platform cannot distinguish the two. A choice of x^* is a choice about the acceptable ratio of babies to bathwater.

Figure 2 is an abstract microeconomic diagram. Its purpose is to build qualitative intuition, not to be a scale model of anything specific. The x-axis is measured in abstract "units" of content. Think of each short interval along the axis as being occupied by an indefinitely large number of individual items of content. There are so many items, in fact, that we will treat the interval $[0, x_{\max}]$ as being effectively continuous; while content comes in distinct items, they are individually too small to be visible to the naked eye. Similarly, the y-axis is measured in abstract units of value. They could be dollars, or euros, or utils. Thus the values of the functions $p(x)$ and $s(x)$ and the constant h have the units of "value per unit of content," where, to repeat, a "unit" is made up of many individual items.[41]

We emphasize this point because it is important to remember that this diagram portrays the *marginal* platform revenue, social benefit, and third-party harm per unit of content. The value of $p(x)$ at a point x is the amount of additional revenue the platform will earn by hosting one additional unit of content at x—i.e., from increasing x^* by one unit. The value $p(x)$ is emphatically not the platform's *total* revenue from setting its moderation threshold to x.

41. The value of the function $\lambda(x)$ is a unitless probability, a number between 0 and 1.

Figure 3: Curves represent marginal value; areas represent total value



Rather, total profits, benefits, and harms are illustrated in Figure 1 (and in the numerous diagrams that will follow) by *areas*. For example, Figure 3 illustrates the platform's profits from setting its moderation level at x^*. At any given point, the vertical distance from the x-axis to the revenue curve $p(x)$ is the platform's marginal revenue from hosting the content at x. The platform's total profits are the area of the green checked region.[42]

B. SOCIAL WELFARE, PLATFORM PROFITS, AND BLANKET IMMUNITY

Now we are ready to use the model to draw conclusions about what the platform will do, and what the regulator wants it to do, which are not necessarily the same. We begin by asking what the socially optimal moderation level would be, and then consider whether the platform will set its moderation at that level. (Spoiler alert: no.)

The marginal social welfare from hosting content is the (known) social benefit from that content minus the (expected) harms, i.e.,

$$s(x) - \lambda(x)h. \tag{1}$$

42. In calculus terms, the platform's total profits are the *integral* of its marginal revenues, i.e.,

$$\int_0^{x^*} p(x)\,dx.$$

Another way to look at this expression is that if the platform hosts content at x, a fraction $\lambda(x)$ of that content will be harmful with value $s(x) - h$ per unit: benefits minus harms. Meanwhile, a fraction $1 - \lambda(x)$ will be harmless with value $s(x)$ per unit: all benefits and no harms. In other words, all of the content, harmful and harmless alike, generates benefits of $s(x)$, but only the harmful fraction $\lambda(x)$ also generates harms h.

Geometrically, the marginal social welfare from hosting content is the vertical distance between the benefit curve $s(x)$ and the expected harm curve $\lambda(x)h$. That value is 0 where the two curves cross.[43]

Call this point x_s, i.e., the **socially efficient moderation level**. It is defined by the equation

$$s(x_s) = \lambda(x_s)h.$$

For $x < x_s$, it is net beneficial to society for the platform to host this content. For $x > x_s$, it is net harmful to society. x_s is the point at which content crosses over from being net beneficial to net harmful. The regulator would prefer the platform to set its moderation level to x_s—i.e., to host content just up to x_s and then stop and take down everything else.

Observe how the value of x_s depends crucially on h. Rearranging the defining equation yields $\lambda(x_s) = \frac{s(x_s)}{h}$. The greater the harm h, the lower the probability $\lambda(x)$ of harm worth tolerating, and thus the lower the appropriate threshold of moderation.

43. By the intermediate value theorem, there is some value of x at which $s(x) - \lambda(x)h = 0$, so the curves do cross.

Figure 4: Optimal moderation

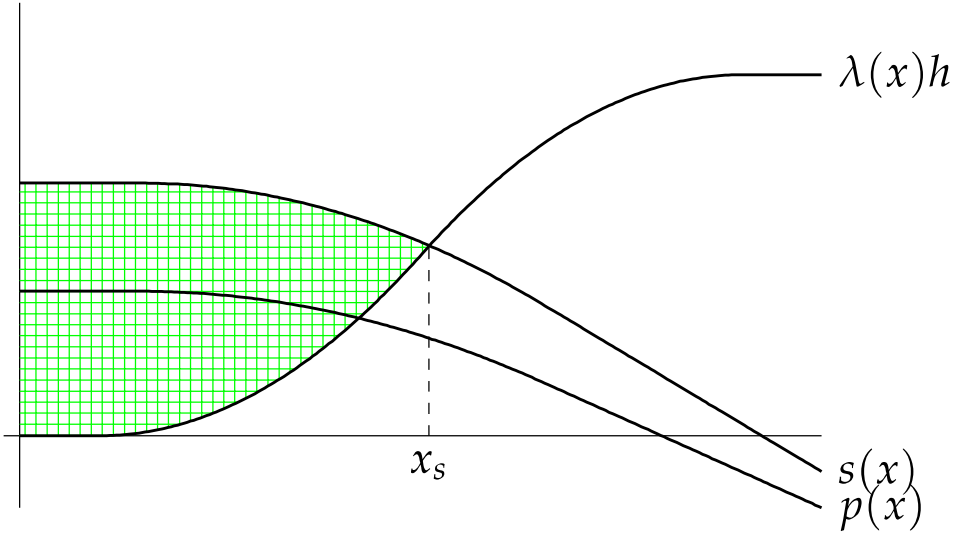


Figure 4 illustrates a socially optimal level of moderation. This is the best moderation that a platform can possibly do without knowing more about which content is harmful and which content is harmless. The green shaded region represents total social welfare under optimal moderation. The platform should host all content to the left of x_s and take down all content to the right of x_s.

Now consider the platform's profits. Since its marginal revenue is $p(x)$, its total profits from setting its moderation level to x^* are the area between $p(x)$ and the x-axis from 0 to x^*. By similar reasoning to the above, the platform maximizes its profits by setting x^* such that

$$p(x^*) = 0.$$

Call this point x_p, the **platform profit-maximizing moderation level**.[44]

44. If there is no such value, which happens when the platform makes positive revenue from all content, the platform should set $x^* = x_{\max}$ and host all content.

Figure 5: Undermoderation under immunity

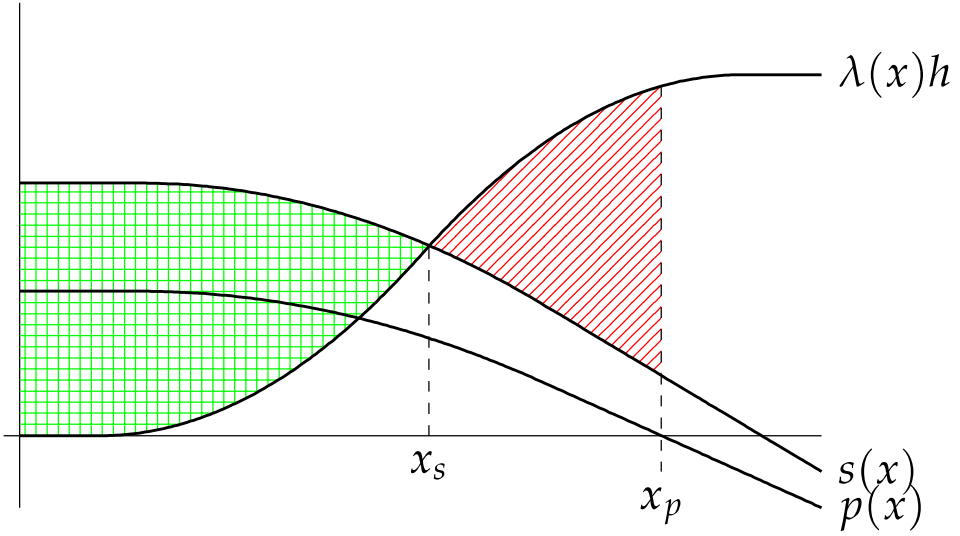
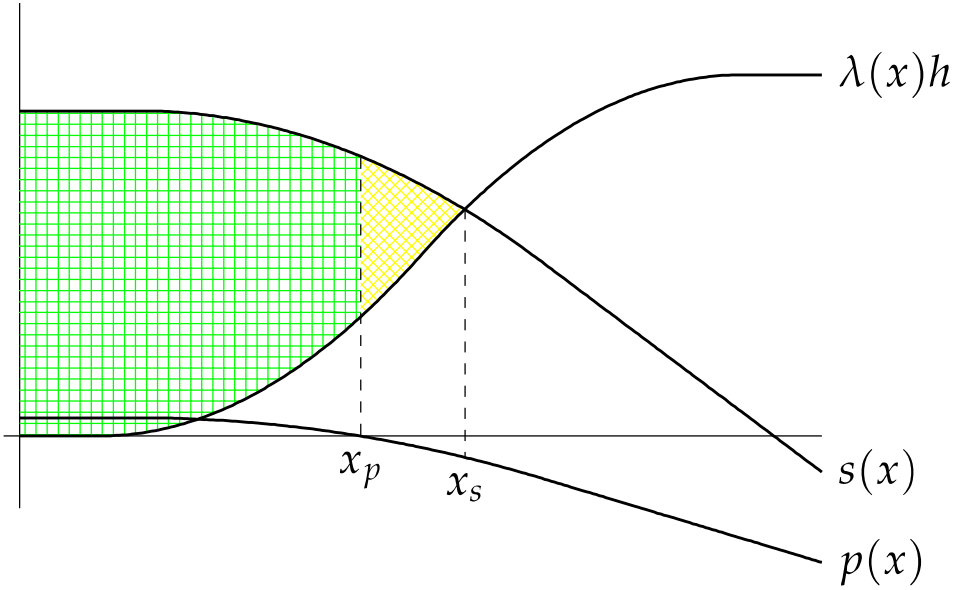


Figure 6: Overmoderation under immunity

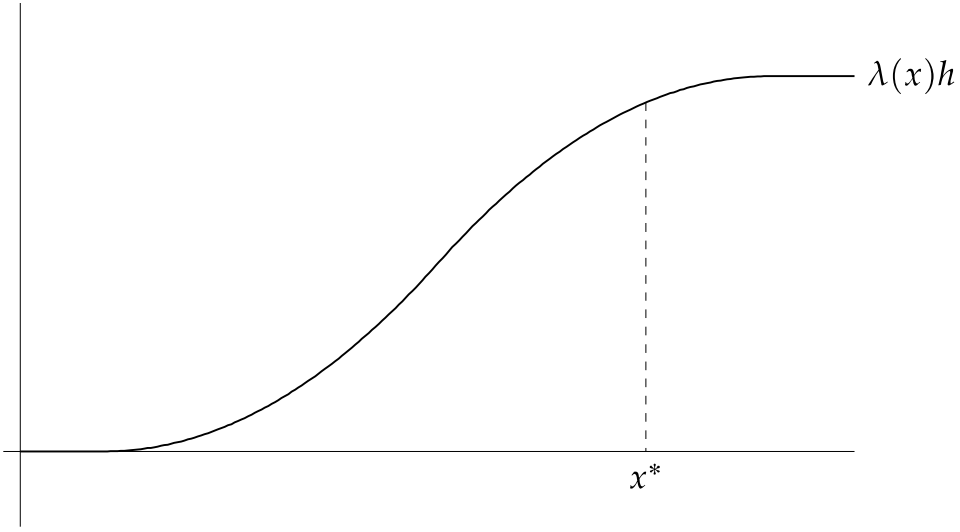


As Figures 5 and 6 show, there is no necessary relationship between x_s and x_p. In Figure 5, the platform undermoderates. It makes money from content that is bad for society, so $x_p > x_s$ and the platform leaves up content that it should ideally take down. The red striped region is the net social loss

from hosting too much content. But in Figure 6, the platform overmoderates. It loses money on content that is good for society, so $x_p < x_s$ and the platform takes down content that it should ideally leave up. The yellow diamond region is the foregone social benefits from content the platform could have hosted but did not.

One way of understanding why both undermoderation and overmoderation are possible is that there are two different effects at work, with opposite signs. On the one hand, the platform fails to internalize the full social benefits of the content that it hosts: $s(x) > p(x)$. On the other hand, when content is harmful the platform does not internalize the harms to third parties: $\lambda(x)h$. On the minimal assumptions we have made, either one of these two effects could dominate. In the real world, both undermoderation and overmoderation are problems that lawmakers have thought serious enough to try to fix. Parts III and IV discuss their various responses in detail.

Figure 7: Blanket immunity



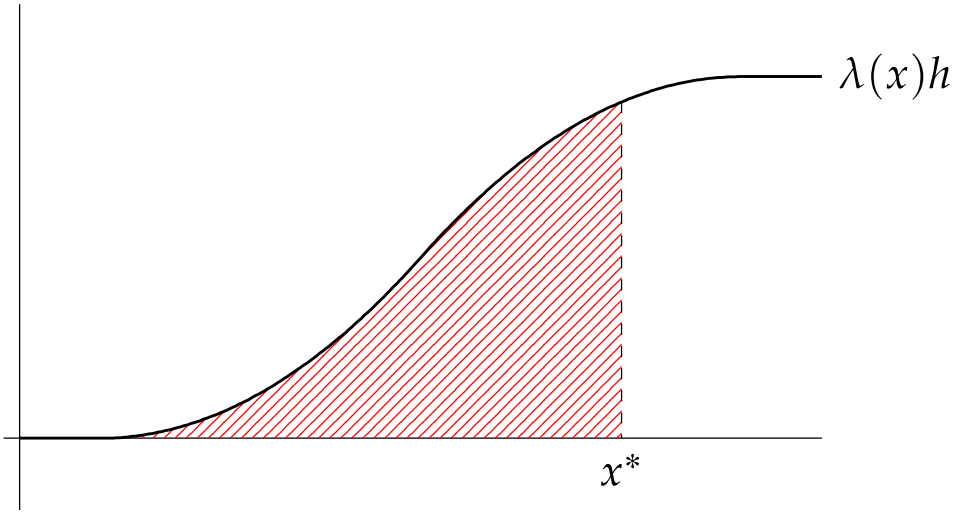
So far, we have been considering a model in which the platform is subject to a legal regime of **blanket immunity**. Figure 7 illustrates this very simple rule. Regardless of where the platform sets x^*, it is not liable for any of the harms that result.

C. STRICT LIABILITY

The essential premise on which any form of liability depends is that some conduct is harmful. The standard law-and-microeconomic response to

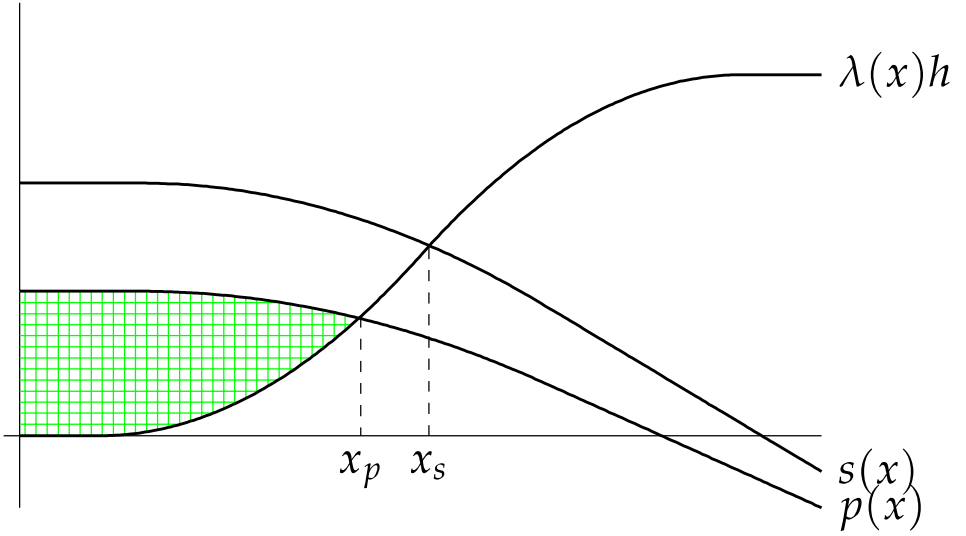
harmful conduct is *strict liability*. If a widget factory is forced to compensate everyone who is injured by defective widgets, the factory will take exactly those manufacturing precautions that are cost-justified. Once the factory internalizes the harms it causes, its incentives are aligned with society's.

Figure 8: Strict liability



For a platform, that conduct is hosting content, and the strict-liability measure of damages is the harm that results from the content that the platform hosts. Figure 8 illustrates that if the platform sets its moderation threshold at x^*, it is liable for all harm caused by the content that it carries (and for none of the harm that would have been caused by content that it could have carried and did not).

Figure 9: Platform's optimal behavior under strict liability



Note that the platform pays no damages for the fraction $1 - \lambda(x)$ of content that is actually harmless. But for the fraction $\lambda(x)$ of content that is harmful, the platform pays total damages of $\lambda(x)h$.

Thus, under strict liability, the platform's marginal profits are $p(x) - \lambda(x)h$. Its profit-maximizing moderation level x^* is defined by

$$p(x^*) = \lambda(x^*)h. \tag{2}$$

Figure 9 illustrates the results. The platform sets its moderation level where its revenue curve $p(x)$ and the damages it must pay $\lambda(x)h$ cross. At that point, its revenues from carrying additional content are exactly cancelled out by the harm that content causes (and hence the damages it must pay).

It follows that *strict liability always results in overmoderation*. Because $p(x) < s(x)$, the platform's profit curve $p(x)$ always intersects the expected-harm curve $\lambda(x)h$ to the left of where the social-benefit curve $s(x)$ intersects $\lambda(x)h$. Thus, $x_p < x_s$.

Figure 10: Strict liability results in overmoderation

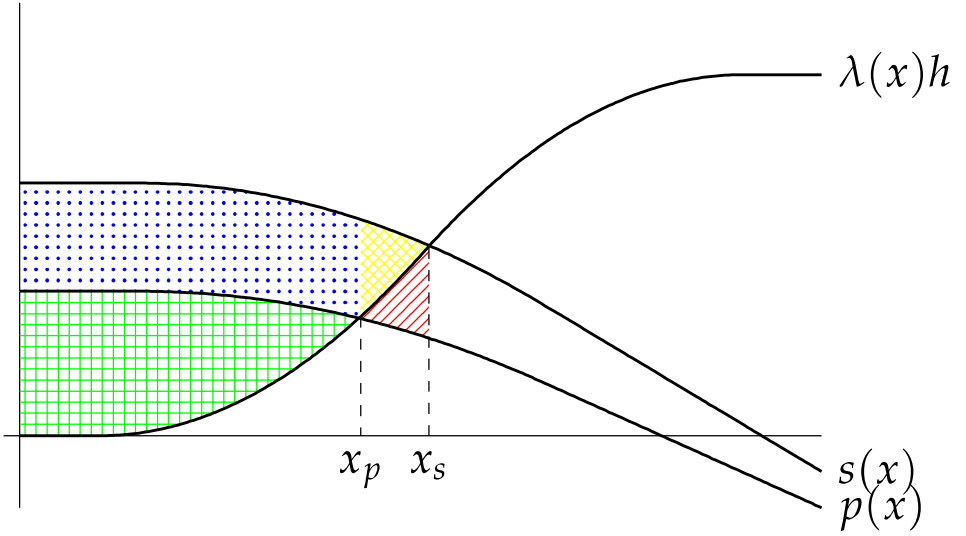


Figure 10 illustrates how strict liability causes overmoderation. The green checked region is the platform's profits, which become 0 exactly at $x_p$where the platform stops hosting content). The blue dotted region is the additional spillover social benefit from the content the platform hosts. Between x_p and x_s, it is unprofitable for the platform to host more content because it would have net losses equal to the area of the red striped region. But content in that range is beneficial overall for society. Society suffers a welfare loss equal to the area of the yellow diamond region from content that platform could have hosted but did not. This content is unprofitable to the platform but beneficial to society, because $p(x) < \lambda(x)h < s(x)$.

D. COSTLESS INVESTIGATIONS

The final moving piece of our model is that a platform can investigate content that it suspects of being harmful. Specifically, we add the option that the platform can pay a cost $c \geq 0$ per unit of content to investigate and determine with certainty whether each item is actually harmful.

To get intuition for how this possibility affects the platform's incentives, we start by presenting extreme cases. When investigation is infeasibly costly, i.e., $c \to \infty$, this model collapses into the previous one because there are no circumstances under which the option to investigate is worth exercising.

On the other hand, when investigation is costless, i.e., $c \to 0$, the platform can perfectly distinguish harmful content and harmless content. That means it

is possible for the platform to take down the harmful content while leaving up the harmless content. From the regulator's perspective, that is exactly what it should do: take down every piece of harmful content and leave up every piece of harmless content.

Naively, it might seem like the effect of costless investigation would be to remove the harm curve $\lambda(x)h$ from the picture, so that the platform earns all the revenue under $p(x)$ and society realizes all the value under $s(x)$. But this is not quite right, because the *harmful content still must be removed*. This means that the platform must forego the revenue, and society the benefits, from the fraction $\lambda(x)$ of content that is removed.

We define $p^*(x) = \big(1 - \lambda(x)\big)p(x)$, i.e., the profits the platform can make by hosting only harmless content. Similarly, we define the corresponding function $s^*(x) = \big(1 - \lambda(x)\big)s(x)$ for social benefits. These new functions represent the maximum revenue and social benefit, respectively, that it is possible to realize with perfect knowledge about which content is harmful.

Figure 11: Platform revenue and social benefits with costless investigations

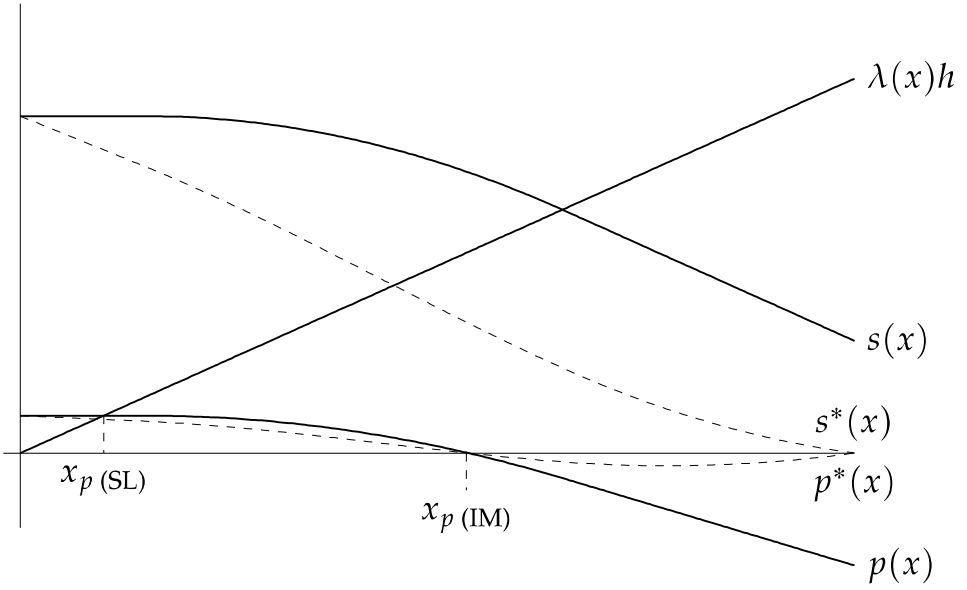


Figure 11 illustrates the platform's profits $p^*(x)$ and social benefits $s^*(x)$ from hosting only harmless content.[45] Their behavior is subtle. $s^*(x)$ starts off equal to $s(x)$ when all content is harmless, and the platform need not remove any content. It immediately dips below $s(x)$ as content must be removed because there is less content available to generate surplus. Eventually, it ends up equal to 0 because all the content is harmful so there is nothing left to host. Similarly, $p^*(x)$ starts off equal to $p(x)$ and immediately dips beneath it. A twist is that $p^*(x)$ becomes 0 exactly when $p(x)$ does because that is the point at which all content, harmful and harmless, is valueless to the platform. From then on $p^*(x) > p(x)$, because the platform saves money by not hosting content that would be unprofitable for it. But like $s(x)$, it eventually ends up equal to 0 because there is nothing left to host.

It is a little difficult to see visually in Figure 11, but costless investigation is always good for society, and society is always better off if the platform removes the harmful content that it knows about. Algebraically, the benefit function with omniscient moderation $s^*(x)$ is always greater than the benefit function with oblivious moderation $s(x) - \lambda(x)h$.[46] Note that society will now prefer to host all harmless content up to the point at which $s(x) = 0$.[47]

From the platform's perspective, omniscient moderation is also never a bad thing. Under immunity, the platform does not care about harmful and harmless content; it still sets its moderation level at $x_{p\,(\mathrm{IM})}$ and it is no worse off. (The "IM" stands for "immunity.") Under strict liability, the platform will still set its moderation level at $x_{p(\mathrm{IM})}$ but it will also remove all of the content $x < x_{p\,(\mathrm{IM})}$ that is actually harmful. This eats into the platform's profits compared with immunity—it makes $p^*(x)$ instead of $p(x)$—but compared with where it would be under strict liability with oblivious moderation it is much better off. Because it does not actually have to pay the harm $\lambda(x)h$, it can move its moderation level from $x_{p\,(\mathrm{SL})}$ to $x_{p\,(\mathrm{IM})}$. (The "SL" stands for "strict liability.") With costless investigation and strict liability, the platform will typically overmoderate for the simple reason that $p(x) < s(x)$, meaning some harmless content may be unprofitable but socially beneficial. The platform may be willing to host more content, but society's preferred moderation level has also shifted to the right.

45. The harm curve $\lambda(x)h$ has been straightened out and the curves $s(x)$ and $p(x)$ separated to make the diagram easier to read.

46. This follows from the definition of $s^*(x)$ and the postulate that $h > s(x)$.

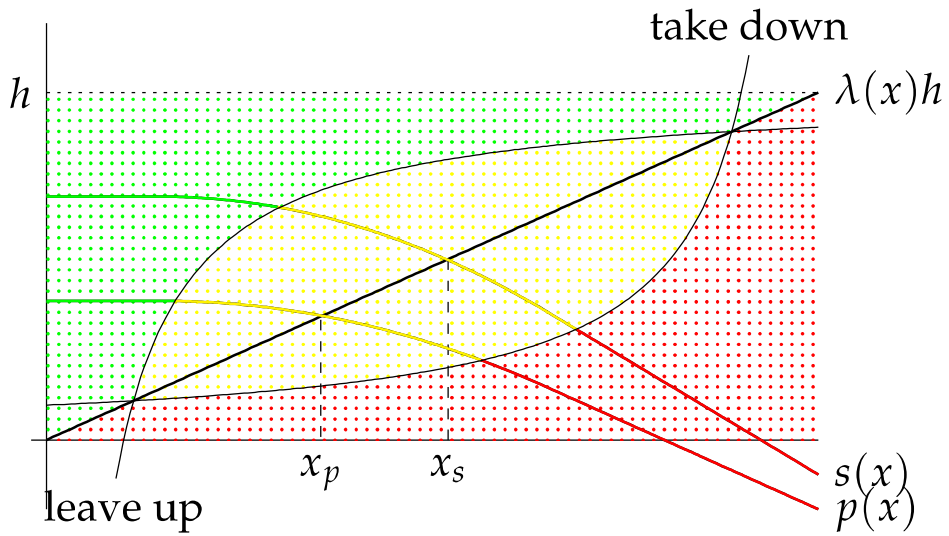
47. Or if there is no such point (as in Figure 11), simply to host all harmless content.

E. COSTLY INVESTIGATIONS

Now consider what happens when an investigation is costly but feasible. Now the platform must make real decisions about whether to investigate because sometimes investigation is worthwhile and sometimes it is not. The platform has three options for any given item of content: leave it up, take it down, or investigate. It is easy to see that the platform will only investigate content where its decision depends on the results of the investigation, i.e., the platform will take the content down if the investigation finds the content harmful and leave it up otherwise. (If the platform intended to take down the content regardless, it could save c by omitting the investigation, and similarly if it intended to leave up the content regardless.)

Thus the expected value to society for content at x is 0 if the platform takes down the content, $s(x) - \lambda(x)h$ per unit if it leaves the content up, and $\big(1 - \lambda(x)\big)s(x) - c$ per unit if it investigates, i.e., the value of a harmless piece of content times the probability that the content is harmless, minus the cost of investigating all content. Intuitively, the platform should prefer takedown for content with $\lambda(x)$ close to 1 and should prefer leaving it up for content with $\lambda(x)$ close to 0, with an interval of investigation somewhere in the middle.

Figure 12: Investigation of intermediate content under strict liability



The regulator is indifferent between takedown and investigation when the value of the content that investigation will allow to remain up minus the costs of investigation ($\big(1-\lambda(x)\big)s(x)-c$) exactly equals the value of taking all content down (which is simply 0). Doing out the math, takedown and investigation are equally efficient when

$$\lambda(x) = 1 - \frac{c}{s(x)}.$$

When c approaches 0, this converges to $\lambda(x) = 1$, i.e., the right end of the investigation interval approaches $x_{\max}$. That is, as the costs of investigation decrease, it is almost always better to investigate than to take down suspected-bad content without first checking.

The regulator is indifferent between investigation and leaving up when the value of the content that investigation will allow to remain up minus the costs of investigation ($\big(1-\lambda(x)\big)s(x)-c$) exactly equals the benefits of all the content minus the costs of the harmful content ($s(x)-\lambda(x)h$). Doing out the math, investigating and leaving up are equally efficient when

$$\lambda(x) = \frac{c}{h - s(x)}.$$

When c approaches 0, this converges to $\lambda(x) = 0$, i.e., the left end of the investigation interval approaches 0. That is, as c decreases, it is almost always better to investigate than to leave up the suspected-good content without first checking.

Put another way, as c decreases, the ideal investigation interval expands to cover more and more content. But as c increases, the investigation interval shrinks and eventually vanishes.[48] When this bound is exceeded, it is never worthwhile from society's perspective for the platform to investigate. It should instead act on the basis of the imperfect information it already has.

These results show that a rational regulator should want platforms to invest resources in investigating only when the cost of investigation is sufficiently low, and then only for a range of intermediate cases where the harmfulness of the content is sufficiently unclear. For content that is highly likely or highly unlikely to be harmful, individual investigation is unnecessary and inefficient. Note that this interval contains x_s—in a sense, affordable

48. It vanishes when:

$$c > \min_{x\in[0,x_{\max}]} s(x)\frac{h-s(x)}{2s(x)-h}.$$

investigations expand the cutoff from a sharp on-off to a range warranting a closer look.

Figure 12 illustrates the intermediate range where investigation is justified.[49] The green dotted region is where no investigation is needed, and the platform should leave up all content; the yellow dotted region is where it should investigate and act accordingly; and the red dotted region is where no investigation is needed and the platform should take down all content. The curve labeled "leave up" is the dividing line between the region where investigation is better than leaving content up and vice versa. The curve labeled "take down" is the dividing line between the region where investigation is better than taking content down, and vice versa. These are two-dimensional regions because whether it is rational to investigate or not depends both on $\lambda(x)$ (the horizontal axis) and on $s(x)$ (the vertical axis). As the probability of content being harmful increases (i.e., as one moves horizontally to the right), one starts in a region where it is optimal to leave content up, passes through a region (possibly zero-width) where investigation is optimal, and then moves into a region where it is optimal to take content down. Similarly, as the value of content increases (i.e., as one moves vertically upwards), the optimal policy changes from takedown to investigation to leaving content up. If the curve $s(x)$ passes through the investigation-justified region at all, then x_s lies within it.

Figure 12 also illustrates the dependence of investigation on c. As c decreases, the upper limit moves upwards and the lower limit moves downwards, increasing the size of the region where investigation is justified. As c increases, these limits converge until eventually the region vanishes entirely. In this case, investigation is never justified, and we are back to the previous model, where $\lambda(x)h$ marks the dividing line between taking down and leaving up.

A nearly identical analysis applies to a platform's incentives under strict liability.[50] Because the platform internalizes all the harm that it causes, the only change is to substitute the platform's private profit $p(x)$ for the overall social value $s(x)$. If there is any range for which investigation is justified, it will contain $x_{p\,(\mathrm{SL})}$. A little algebraic manipulation shows that the platform's preferred interval of investigation is always *shifted left* from the regulator's

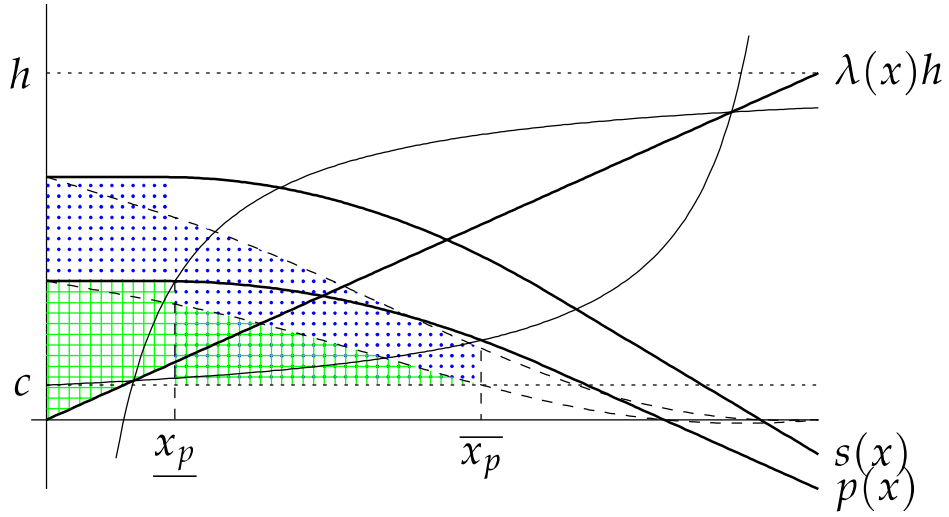
49. Again, for simplicity of illustration, $\lambda(x)$ is shown as a straight line, but the same results hold in the general case where it is any weakly increasing function that goes from 0 to 1 on the interval $[0, x_{\max}]$.

50. Under blanket immunity, a platform will never investigate. Instead, it will always choose to leave all content up.

preferred interval.[51] Intuitively, because the platform has less at stake, it will be more likely to remove content rather than investigating and also more likely to investigate content rather than leaving it up.

Figure 13 shows the platform's profits and social welfare under strict liability when the platform can investigate content. The upper dashed line is $s^*(x)$ and the lower dashed line is $p^*(x)$—respectively, the social benefit and platform-profit effects of omniscient moderation. The difference is that now this moderation must be paid for with investigations, so the net social benefit *if the platform investigates* is $s^*(x) - c$ and the net profit *if the platform investigates* is $p^*(x) - c$.

Figure 13: Platform's profits and social benefits under strict liability with costly investigation



The platform's profits are the green gridded region, and the additional positive externalities for society are the blue dotted region. Beneath the lower limit of investigation, $\underline{x_p}$, matters are as before: the platform's marginal profit is $p(x)$ (income) minus $\lambda(x)h$ (expenses) and marginal social welfare is

51. To be precise, at the lower end

$$\frac{c}{h - s(x)} < \frac{c}{h - p(x)},$$

and at the upper end

$$1 - \frac{c}{p(x)} < 1 - \frac{c}{s(x)}.$$

$s(x) - \lambda(x)h$. When $\lambda(x)$ crosses into the region where investigation is optimal, the platform's marginal revenue is now defined by the difference between $p^*(x)$ (income) and c (expenses). At this point, both income and expenses shift discontinuously downward. The platform is taking in less revenue now that it is removing some content, but that drop is exactly offset by the savings from investigating rather than paying damages. Marginal social welfare discontinuously decreases—intuitively, because society has more to gain from beneficial content and would not have started investigating until later. In this region of investigation, both profits and welfare decrease faster than they did under leave-it-all-up, as more and more content is removed. But this steeper decrease is more than offset by the fact that costs are now constant at c, rather than increasing with $\lambda(x)$. At the upper limit of investigation, $\overline{x_p}$, the platform's marginal profit is zero, so it switches to taking all content down, which zeroes out both marginal profit and marginal welfare going forward. Again, it is visually apparent that the platform is making different tradeoffs than society—it would still be socially beneficial at $\overline{x_p}$ for the platform to continue investigating content.

In short, the ability to investigate increases both social welfare and the platform's profits, but it does not automatically align the platform's incentives with society's incentives.

F. COLLATERAL CENSORSHIP

It is critical to understand why and when strict liability causes overmoderation. Strict liability causes the platform to internalize the harms from the content it carries, but not the offsetting benefits. This asymmetry between harm (for which it faces liability) and benefits (for which it is not compensated) pushes the platform to remove more content than an omniscient regulator would.

This overmoderation fundamentally depends on the platform's imperfect information about content. If the platform could distinguish harmless and harmful content without incurring costs, then strict liability would be efficient. It would be feasible to expect the platform to separate the two and remove only the harmful content. But given imperfect information, the platform *cannot tell with certainty* which content is harmless and creates net positive externalities and which content is harmful and creates net negative externalities. A platform facing strict liability consistently overmoderates and removes more harmless content than it should from society's perspective.

Thus, our model validates Felix Wu's argument for intermediary immunity.[52] The combination of positive externalities and imperfect information causes a platform subject to strict liability to engage in collateral censorship. The platform has less at stake than an original speaker (positive externalities) and responds by removing good content as well as bad (imperfect information). These conditions are jointly necessary and sufficient; if there are no positive externalities (i.e., $s(x) = p(x)$) or the platform has perfect information (i.e., $\lambda(x) = 0$ or $\lambda(x) = 1$ for all content), then strict liability is efficient.

It is worth dwelling for a bit on the nature of these positive externalities. A widget factory might come close to capturing the full social value of the widgets it makes. But a platform does not, for at least two reasons.

First, a platform's "product" is not widgets but speech. Speech consists of information, and information is a public good. Once it has been shared with one listener, then neither the speaker nor the platform can easily prevent them from sharing it with others. A dance video that goes viral on TikTok will be reposted to Twitter and YouTube; the information in a plumbing tutorial will be retained in the minds of viewers and shared with others. All the third-party value is an externality from both the speaker's and the platform's perspectives.[53]

The second source of positive externalities is that platforms do not even capture the full value to speakers of the content they host. As Felix Wu convincingly argues, the value to a user of posting content to a platform is typically much larger than the value to the platform of hosting that content.[54] A platform does not have an original speaker's incentives. This point holds true even for non-speech platforms. For example, Airbnb captures only part of the value that apartment hosts and guests enjoy from rentals made through the platform.[55]

As Wu explains, speech law already provides heightened protections for original speakers—and yet intermediaries have protections that are higher

52. *See* Wu, *supra* note 1.

53. *See* C. Edwin Baker, *Giving the Audience What It Wants*, 58 OHIO ST. L.J. 311 (1997).

54. *See* Wu, *supra* note 1, at 303–8.

55. *See* Chiara Farronato & Andrey Fradkin, *The Welfare Effects of Peer Entry: The Case of Airbnb and the Accommodation Industry*, 112 AM. ECON. REV. 1782, 1783 (2022) (estimating that in 2014 Airbnb generated "$112 million in peer host surplus, or about $26 per room-night"). *See generally* Erik Brynjolfsson, Avinash Collis & Felix Eggers, *Using Massive Online Choice Experiments to Measure Changes in Well-Being*, 116 PROC. NAT'L ACAD. SCI. 7250 (2019) (estimating value to consumers of numerous online platforms).

still. [56] Speakers have private motivations for speaking: financial, self-expression, reputation-building, community-building, or even revenge. Platforms share their speech but not their motivations.

Platforms also differ from speakers in that speakers generally have much better information about the harmfulness of their speech. A speaker knows whether there is a factual basis for allegations of corruption or harassment; a platform does not. A speaker knows whether they wrote a song themselves or copied it from someone else; a platform does not. A speaker is much less likely to be chilled from harmless speech by the threat of liability for harmful speech.

Whether social welfare is higher under strict liability or immunity depends on the parameters of the model: $p(x)$, $s(x)$, h, and $\lambda(x)$. Strict liability always leads to overmoderation; immunity could either undershoot or overshoot the efficient level of moderation. Generally speaking, a blanket immunity regime is most justified when there are large positive externalities (a large difference between $s(x)$ and $p(x)$), highly imperfect information ($\lambda(x)$ has a large intermediate region that is not close to 0 or to 1), and socially harmful content is also unprofitable ($x_p < x_s$). There is a strong argument that these conditions describe many categories of content moderation today.

G. THE MODERATOR'S DILEMMA

Now we are in a position to appreciate the crucial policy arguments at the heart of § 230. Famously, § 230 was enacted against the backdrop of two judicial decisions on the liability of online intermediaries, *Cubby v. CompuServe* and *Stratton Oakmont v. Prodigy*. In *Cubby*, the court held that CompuServe could not be held liable for user-posted content where it "neither knew nor had reason to know" that the content was defamatory.[57] But in *Stratton Oakmont*, the court held that Prodigy could be held liable for user-posted content, even where it lacked such knowledge.[58] Both courts treated the cases as involving imperfect information—the issue was how a platform *without* specific knowledge should be treated.

Notoriously, the *Stratton Oakmont* court distinguished *Cubby* on the grounds that Prodigy's "conscious choice, to gain the benefits of editorial control, has opened it up to a greater liability than CompuServe and other computer networks that make no such choice."[59] On this reasoning, moderated services

56. Wu, *supra* note 1, at 304.

57. Cubby, Inc. v. CompuServe, 776 F. Supp. 135, 141 (S.D.N.Y. 1991).

58. Stratton Oakmont, Inc. v. Prodigy Servs. Corp., No. 031063/94, 1995 WL 323710, at *3 (N.Y. Sup. Ct. May 24, 1995).

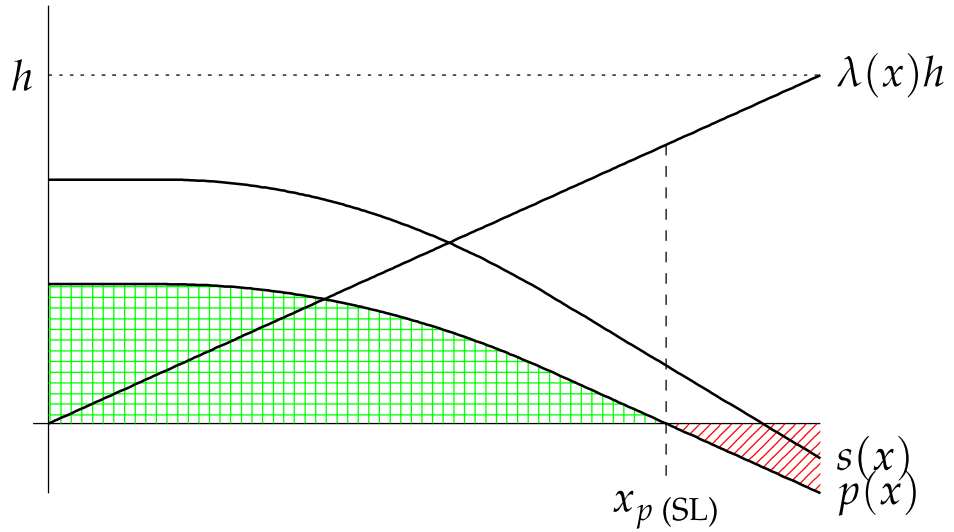
59. *Id.* at *5.

like Prodigy that exercise "editorial control" face strict liability, whereas unmoderated services like CompuServe that exercise no editorial control are immune.

In terms of our model, the rule in *Stratton Oakmont* forces platforms to make a choice. If they host *all* content, they face no liability. But if they remove *any* content, they are strictly liable for the harms caused by any content they do not remove.

Section II.E of this Article analyzed the platform's behavior if it chooses to moderate and thus commits to optimal investigations to maximize its profits in the presence of strict liability. Figure 11 shows the range of content for which the platform will investigate, and Figure 12 shows the platform's profits (green gridded) and additional social welfare (blue dotted) that result.

Figure 14: Platform's profits if it carries all content



Compare that situation with Figure 14, which shows the platform's profits (green gridded) and losses (red striped) if it chooses simply to carry all content. While the platform ends up taking some losses on the spammy, negative-revenue content at the right, it also makes substantial profits on the positive-revenue content at the left—and since the platform no longer has to pay damages, it can pocket all of that revenue without concern for the resulting harms.

Comparing Figure 12 with Figure 14, it is visually clear that the platform is better off not moderating at all. This is contingent on the precise values of the

parameters in the model, especially its profit function $p(x)$. For a different and lower $p(x)$, the platform might lose so much money hosting the worst of the worst content that it would be better off moderating and accepting liability.

These diagrams also reinforce an important point about moderation: *almost all platforms have their own strong incentives to engage in at least some moderation*. The platform here would moderate at $x_{p\,(\mathrm{SL})}$ even in the absence of liability because the worst content is genuinely bad for the platform and its users. Liability is not the only incentive to moderation, and by putting the platform to the choice between voluntary moderation and immunity, the regulator runs the risk that the platform will choose to give up its voluntary moderation efforts.

Figure 15: Social welfare if the platform carries all content

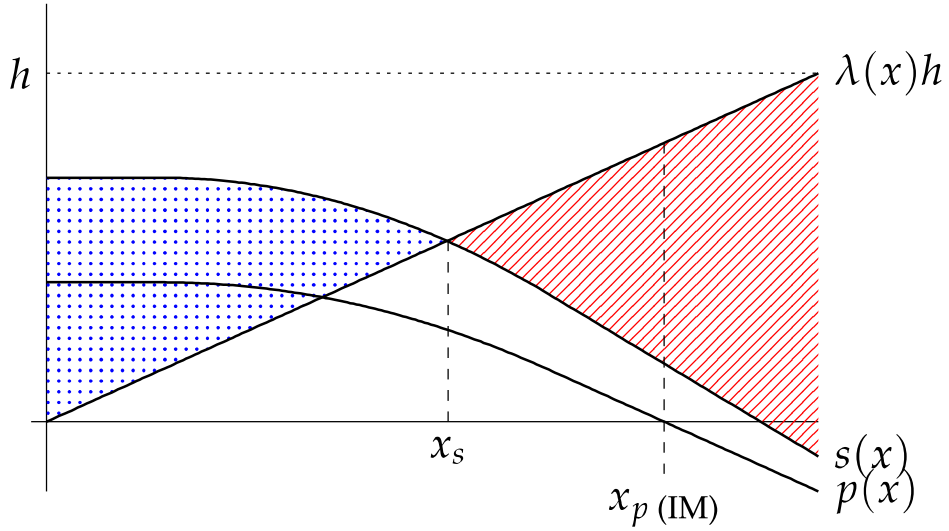


Figure 15 shows the resulting social welfare if the platform chooses to carry all content. The blue dotted region is social benefit and the red striped region is social harm. The red striped area is large. The additional social welfare loss from $x_{p\,(\mathrm{IM})}$ to $x_{\max}$ is particularly substantial. Intuitively, the content that the platform is most selfishly interested in removing is also the content that the regulator most wants it to remove.

Take a moment to let the implications sink in. The platform here is much better off carrying all content; society is much worse off as a result. Strict liability induces the platform to increase its moderation effort from $x_{p\,(\mathrm{IM})}$, where it would moderate in the absence of liability at all. Society gains as a result. But when the platform has the option of not moderating at all—or put

another way, when strict liability is the price it must pay for engaging in moderation—the platform is better off turning off its moderation. It no longer has to pay damages for the content it carries, it no longer has to pay investigatory costs, and it can carry content that would have been unprofitably risky before. These gains are more than enough to outweigh the harms to its platform from the negative-value content on the right. From society's perspective, this is a disaster. In trying to encourage the platform to moderate, the regulator has perversely discouraged it from doing so.

Eric Goldman refers the platform's choice as the "moderator's dilemma."[60] The platform wants to moderate in order to improve its offerings for its users. But when moderation also becomes the legal trigger for liability, the platform must consider whether moderation is still worthwhile.

This is why § 230(c) is titled "Protection for 'Good Samaritan' blocking and screening of offensive material."[61] It was enacted to remove the perverse disincentive to moderation created by the rule of *Cubby*. A platform protected by § 230 is now free to move its moderation off of $x_{\max}$ without fear that it will now open itself to liability and be forced to move much further to the left.

IV. POLICY RESPONSES TO UNDERMODERATION

The fundamental challenge of platform liability law is that content has both harms and benefits to society that the platform does not internalize. A profit-maximizing platform makes its decisions based on how much it can make from hosting content, paying no attention to either positive or negative spillovers. We have seen that under blanket immunity, either of these effects can dominate, so both overmoderation and undermoderation are possible.[62] It is technically possible for these effects to cancel out, so that the platform arrives at an appropriate level of moderation on its own. But there is no particular reason to expect that this would be the case. Instead, a particular platform, hosting a particular type of content, with particular harms and benefits, will typically fall on one side or the other.

This Part gives a comparative analysis of the ways that a regulator could respond to undermoderation; the next Part similarly considers responses to overmoderation. We have already discussed strict liability in detail; this Part considers liability on notice, negligence, and conditional immunity. The point

60. Eric Goldman, *Internet Immunity and the Freedom to Code*, 62 COMMC'N OF THE ACM 22, 22 (2019), https://ssrn.com/abstract=3443976.

61. 47 U.S.C. § 230(c).

62. *See supra* Section III.B.

is not to settle on one or another as optimal, but instead to bring out the intuitions behind each and to get a sense of the conditions they depend on.

A. ACTUAL KNOWLEDGE

At common law, a "distributor" of defamatory speech published by a third party (e.g., a bookstore) was liable "if, but only if, [it] knows or has reason to know of its defamatory character."[63] Section 512(c)(1)(A) removes a platform's immunity as to specific material if it has "actual knowledge that the material . . . is infringing"[64] and the platform does not "act[] expeditiously to remove, or disable access to, the material."[65]

These are examples of *actual knowledge*: a platform is liable for harmful content that it hosts, but only when it has specific knowledge that a particular item is harmful. The intuition behind an actual-knowledge regime is that while it might not be feasible to require a platform to *acquire* the knowledge to show that an item of content is harmful on its own, once the platform *has* such knowledge (from whatever source derived), it is reasonable to expect the platform to take action on it.

In our model, actual knowledge corresponds to cases where the cost of investigation c is 0 as to a particular item of content. As we saw in Section II.D, imposing liability for harmful content when $c = 0$ does not distort the platform's incentives. The platform takes down harmful items where $c = 0$ and it is socially optimal for it to do so. This is a strict improvement over immunity. (The platform leaves up harmless items when $p(x) > 0$, which is not socially optimal, but adding an actual knowledge test to a baseline of immunity does not change matters.)

It is crucial, however, that "actual knowledge" actually means actual knowledge. When investigation is costly because $c > 0$, imposing not-actually "actual knowledge" liability does distort the platform's incentives. In Section III.C below, we analyze the platform's responses to a rule that holds it liable when content has a *high probability* of being harmful, and we observe some of the same potential distorting effects as strict liability.

B. LIABILITY ON NOTICE

If someone else is willing to bear the expense of investigating content, then from the platform's perspective, it receives investigation for free. Put another way, the value of a takedown notice is that it reduces the investigation costs as

63. RESTATEMENT (SECOND) OF TORTS § 581(1) (AM. L. INST. 1977).
64. 17 U.S.C. § 512(c)(1)(A)(i).
65. *Id.* § 512(c)(1)(A)(iii).

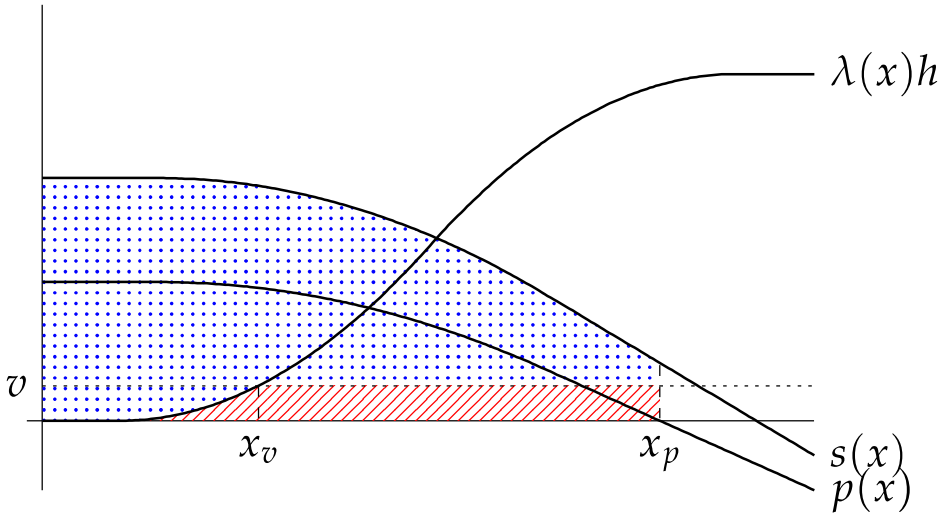
to specific content by narrowing the issues that the platform must investigate. When investigation is expensive, as we have seen, a rational platform will not bother searching for the needle. Instead, it will overmoderate and throw out the entire haystack. But when someone points to an alleged needle, it is far easier for the platform to decide whether it is actually a needle.

The most straightforward way to model liability on notice in our framework is to introduce additional agents: the *victims* of harm, who can investigate content and provide notice to the platform. In this modification, each individual item of content is indexed to a distinct victim, and that victim is the specific person one who suffers the harm if the item is harmful and the platform carries it. The victims, like the platform, can investigate content. Their cost to investigate need not be the same, so we write c_p for the platform's cost of investigation and c_v for the victim's cost. The victims are also able to send notices to the platform for any content they choose, and the platform is liable to the relevant victim for any harm that victim suffers from content about which the platform has received a notice.[66]

Intuitively, it seems that liability on notice should induce the state of affairs depicted in Figure 16. In this figure, the red striped region shows victims' uncompensated harms and investigation costs. The blue dotted region above it shows the social surplus. For content at x, the relevant victim has the option of doing nothing and suffering harm $\lambda(x)h$ or of investigating at cost c_v and giving the platform notice if the content is harmful. For low $\lambda(x)$ they prefer to suffer the harm; for high $\lambda(x)$ they prefer to investigate, with crossover at the point x_v for which $\lambda(x_v)h = c_v$. The platform will always remove any harmful content for which it receives a notice, because a costless removal is better than paying to compensate a harm h that outweighs its profits $p(x)$. Thus, the platform never actually has to pay compensation. (The platform cuts off hosting content entirely all at x^*, where its revenues go negative.)

66. All parties can observe the functions $p(x)$, $s(x)$, and $\lambda(x)$, and the parameters h, c_p, and c_v.

Figure 16: Naive model of liability on notice



But this reasoning is incomplete. The problem is that victims *are not restricted to sending notices for content that is actually harmful*. The victims have a third option for content besides ignoring it and investigating it—they can also send a notice without investigation. In economic terms, liability on notice creates a signaling game between victims and platform. For each item of content, the relevant victim chooses whether to ignore it, investigate and give notice if the content is harmful, or give notice without investigation. The platform either does or does not receive a notice and then chooses whether to take the content down, investigate it, or leave it up. The above reasoning applies only if the signals are all truthful.

When the signals are not truthful, notice on takedown might collapse into strict liability. The victim never investigates but always sends a notice. Because the platform receives a notice regardless of whether the content is harmful or not, the notices are of no use to the platform in distinguishing harmful from harmless content. At the same time, the platform is now legally on notice of all harmful content, and thus subject to strict liability for failure to remove it. The platform faces exactly the same incentives, with exactly the same knowledge, and exactly the same options as in the strict liability case. The notices do no useful work.

This analysis bears out academic criticism of the § 512(c) notice-and-takedown regime. Copyright claimants frequently send notices based on no or minimal investigation, including on content that involves no reuse of copyrightable expression or is obviously a fair use.

One way to make the signal provided by a notice more credible is to make it more expensive to send notices against harmless material. Section 512(f) tries to do this by imposing liability on anyone who "knowingly materially misrepresents" that material is infringing in a takedown notice.[67] Unfortunately, judicial interpretations have almost completely defanged this remedy. Courts have held that a subjective belief of infringement, however unreasonable, is a sufficient defense to a § 512(f) suit.[68] They have also held that even the most cursory investigative process is sufficient.[69] These holdings undermine the effectiveness of notices as signals.

Another way to make a notice more useful is to require it to contain specific evidence of harmfulness, thereby making the platform's own investigation cheaper. To rephrase the standard test for copyright infringement slightly, a claim of copyright infringement requires proof that (1) particular material (2) uses a copyrighted work (3) in a way that infringes.[70] In the abstract, investigation is expensive because a platform must investigate all of its content, compare that content to all copyrighted works, and consider all possible justifications (such as licenses, fair use, etc.). The statutory template for a takedown notice addresses these elements by requiring, respectively, "[i]dentification of the material that is claimed to be infringing . . . and information reasonably sufficient to permit the service provider to locate the material,"[71] "[i]dentification of the copyrighted work claimed to have been infringed,"[72] and "[a] statement that the complaining party has a good faith belief that use of the material . . . is not authorized by the copyright owner, its agent, or the law."[73]

Experience has shown that these three requirements stand on somewhat different footings. Courts have generally been unwilling to relax the requirement of identification of specific material, recognizing that without that specific identification the platform must investigate a vast array of content.[74] And plaintiffs have also been held to the requirement that they identify the

67. *See* 17 U.S.C. § 512(f).

68. *See* Rossi v. Motion Picture Ass'n of Am., 391 F.3d 1000, 1004–05 (9th Cir. 2004).

69. *See* Lenz v. Universal Music Corp., 815 F.3d 1145, 1154 (9th Cir. 2015).

70. Feist Publ'n, Inc. v. Rural Tel. Serv. Co., Inc., 499 U.S. 340, 361 (1991) ("To establish infringement, two elements must be proven: (1) ownership of a valid copyright, and (2) copying of constituent elements of the work that are original.").

71. 17 U.S.C. § 512(c)(3)(A)(iii).

72. *Id.* § 512(c)(3)(A)(ii).

73. *Id.* § 512(c)(3)(A)(v).

74. *See* Perfect 10, Inc. v. CCBill LLC, 340 F. Supp. 2d 1077, 1099–101 (N.D. Cal. 2004), *aff'd in relevant part*, 488 F.3d 1102 (9th Cir. 2007).

relevant copyrighted works.[75] (Indeed, in a world where copyright subsists on fixation, almost every upload will contain material that is copyrighted by someone, so that all of the important questions about the copyright itself go to whether the uploader had the right to do so.) But, as noted above, courts have held that the "good faith belief" required by § 512(c)(3)(A)(v) can be satisfied by a subjective belief, regardless of whether that belief is reasonable or not. Even if the notice-sender acts in bad faith, the damages against them are likely to be nominal at best.[76]

This analysis also shows why commentators have generally regarded liability on notice as producing similar chilling effects to strict liability, even outside the copyright space.[77] It is simply too easy to send a notice against content that is not actually harmful. Proposals to instate some kind of liability on notice need to affirmatively demonstrate that the notices they allow will be credible signals.

C. NEGLIGENCE

Strict liability is not the only form of liability. Another version, which is modeled on the negligence tort, sets an objective standard of care. If the actor complies with the standard of care, it is not liable, even if harm results. But if the actor's conduct falls beneath the standard of care, it is liable for any resulting harm. Although scholars dispute the extent to which the standard of care in negligence is defined mathematically, it is sometimes described in terms of the "Hand formula," $B = P \times L$. Under this formula, an actor is liable for failure to invest in a precaution that would have prevented a harm if the cost of the precaution B is less than the ex ante probability of harm P times the magnitude of the harm L.

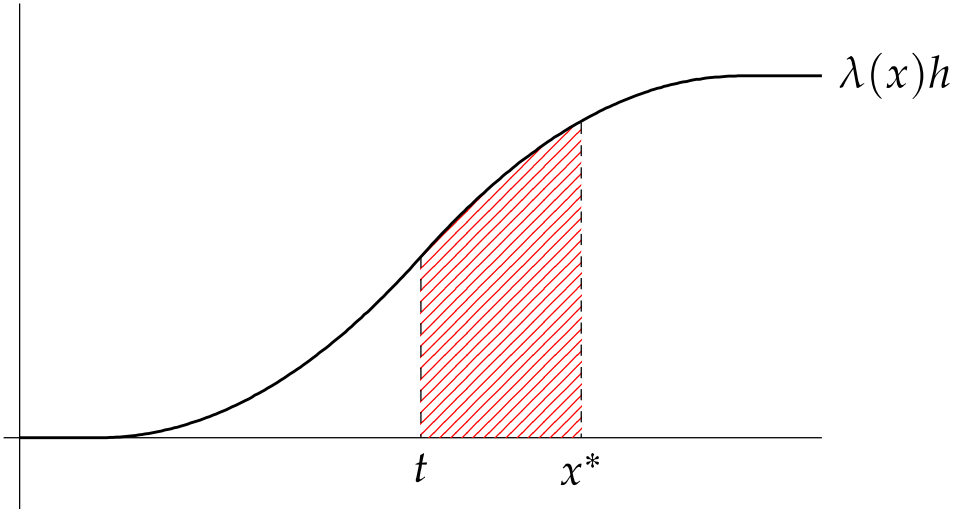
In our model, the regulator imposes negligence liability on the platform by setting a threshold t. The platform is liable for the full harm resulting from hosting any content with $x > t$ but it is not liable for any harm from content with $x \leq t$. Figure 17 illustrates this concept. Note the sharp discontinuity at t.

75. *See id.*

76. *See* Rossi v. Motion Picture Ass'n of Am., 398 F.3d 1000, 1004–05 (9th Cir. 2004); Lenz v. Universal Music Corp., 815 F.3d 1145, 1154, 1156 (9th Cir. 2015).

77. *E.g.*, Wu, *supra* note 1; Schruers, *supra* note 7.

Figure 17: Negligence



The platform's behavior under a negligence regime is identical to its behavior under strict liability, except that it always chooses to leave content up for $x < t$. Thus, the regulator should set t equal to the value of x for which the social benefit of leaving content up is equal to the social benefit of investigation. But this is just the socially optimal lower limit of investigation $\underline{x_s}$. Setting t higher means that the platform will leave up content the regulator would prefer it to investigate (or even take down); setting t lower means that the platform will investigate content the regulator would prefer it to leave up without investigation.

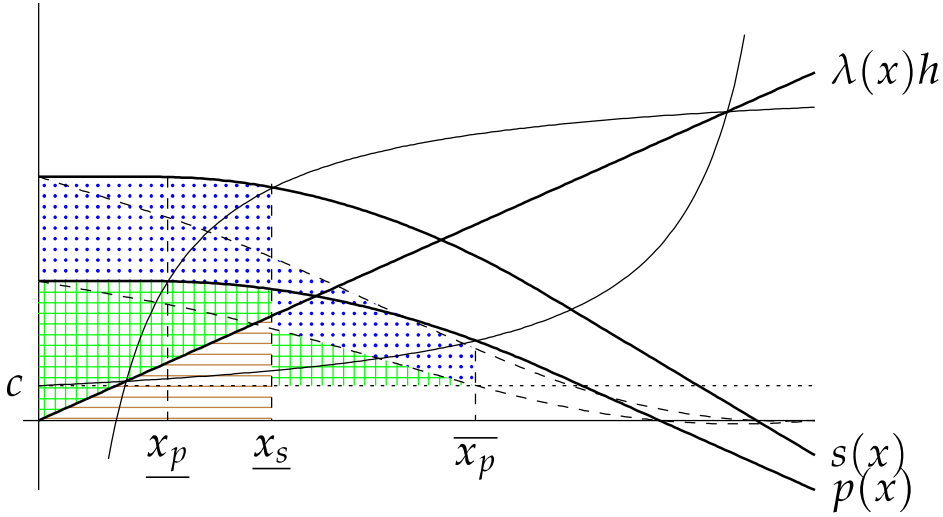
Figure 18 shows the consequences of a negligence rule. Most importantly, it pushes the platform's lower limit of investigation up from $\underline{x_p}$ (where $p(x)$ intersects the lower-limit curve) to $\underline{x_s}$ (where $s(x)$ intersects the lower-limit curve). This is welfare-improving because throughout this range, the value of leaving up ($s(x) - \lambda(x)h$) exceeds the value of investigation ($\big(1 - \lambda(x)\big)s(x) - c$). (Compare this figure to Figure 14, which shows what happens under strict liability.)

There are also distributional consequences. The brown striped region represents uncompensated harm to victims—this region is not part of the social surplus from content on the platform. But it is part of the platform's profits. Note that this is harm that is socially optimal not to attempt to prevent because imposing liability on the platform causes it to inefficiently spend resources investigating. This is not a case where the platform takes down harmless content in ignorance of its harmless nature (that occurs in at higher

values of x, at the right of the diagram, beyond the upper-limit curve). It is a case where the platform *spends too much* on investigation under strict liability, and society is better off overall moving the threshold of liability upwards from 0 (strict liability) to $\underline{x_s}$ (optimal standard of care under negligence).

This system is still not efficient. It gets the platform's incentives right at the boundary between leaving up and investigating, but not at the boundary between investigating and taking down. The platform will still spend too little on investigation at that boundary (from the regulator's perspective) and take down harmless content because it is too similar to possibly harmful content. The optimal negligence rule still results in overmoderation.

Figure 18: Social welfare under negligence



There is an additional challenge. A negligence regime improves on strict liability if the regulator can calculate $s(x)$, h, $\lambda(x)$, and c to set the appropriate threshold. This is not necessarily an easy task, because it involves weighing the full benefits and harms of content, the ex ante likelihood that given content is harmful, and the cost of investigation to make sure. If the regulator sets t too low, it blends into strict liability. If the regulator sets t too high, it blends into immunity. Negligence is always at least as good as one of these two, but it is not necessarily any better.

An example of a negligence rule in platform law is § 512(c)(1)(A)(ii), which removes the platform's immunity as to specific content if it is "aware of facts or circumstances from which infringing activity is apparent" and fails to

remove the content.[78] This exception, known in the caselaw and scholarship as the "red flag" provision, is best understood as a judgment that in certain cases, the probability of infringement is high enough to justify removal.[79] In other words, the red flag provision is a negligence-style rule—beyond some threshold t of high likelihood that content is infringing, the platform will be liable for all such infringing content. Caselaw confirms that t is high.[80] It is not enough that the platform is aware in general that some content is infringing; it must be awareness of "facts that would have made the specific infringement 'objectively' obvious to a reasonable person."[81]

D. CONDITIONAL IMMUNITY

A hybrid of strict liability and immunity is *conditional immunity*. Informally, the platform is immune provided that it keeps total harm small enough. Formally, the regulator sets a harm threshold T. If the total harm caused by the content that the platform hosts is less than or equal to T, the platform's liability is zero (Figure 19). In Figure 19, the yellow dotted region shows harm caused to victims for which the platform is not liable. But if the total harm exceeds this threshold, the factory loses its immunity and is liable for *all* the harm it caused, even those beneath the threshold (Figure 20). In Figure 20, the red striped region shows harm caused to victims for which the platform is now liable. The region that was yellow in Figure 19 is now red in Figure 20. The platform's liability to victims of harm depends on how much other harmful content it allows.

78. 17 U.S.C. § 512(c)(3)(A)(ii).

79. *See, e.g.*, Edward Lee, *Decoding the DMCA Safe Harbors*, 32 COLUM. J.L. & ARTS 233, 251–59 (2008); Viacom Int'l, Inc. v. YouTube, Inc., 676 F.3d 19, 31–32 (2d Cir. 2012).

80. *Viacom Int'l, Inc.*, 676 F.3d at 31–32.

81. *Id.* at 31.

Figure 19: Conditional immunity (below threshold)

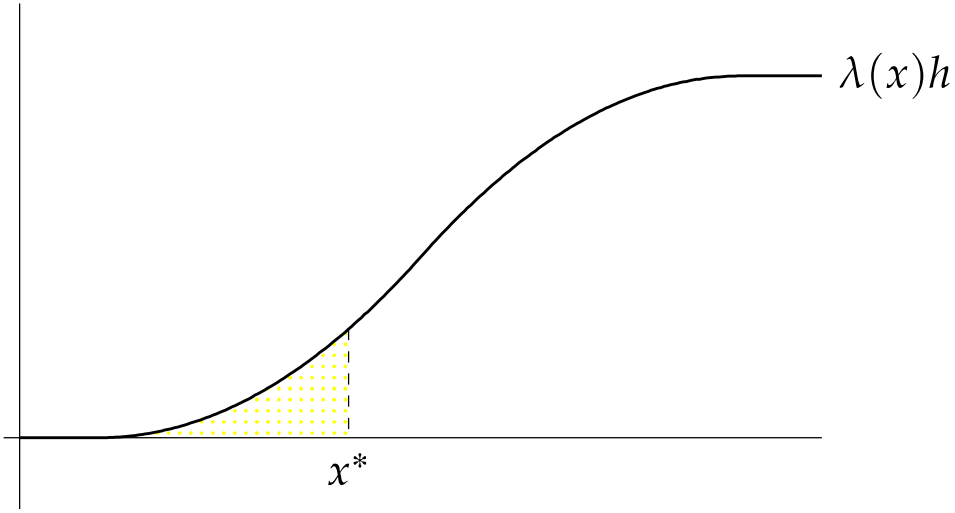
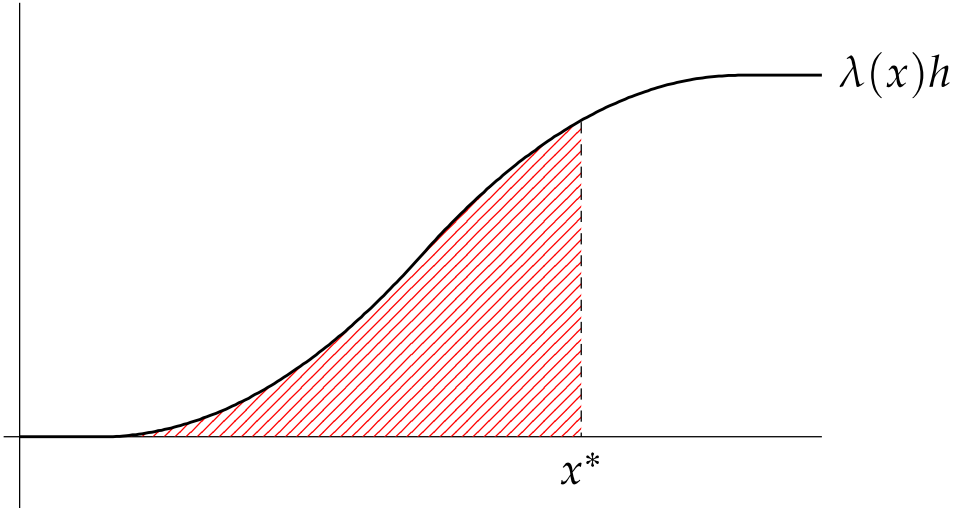


Figure 20: Conditional immunity (above threshold)



An example of conditional immunity is the repeat-infringer provision (RIP) of § 512. To be eligible for the safe harbor at all, a platform must "adopt[] and reasonably implement[] . . . a policy that provides for the termination in appropriate circumstances of . . . repeat infringers."[82] If a platform doesn't do a good enough job at removing content posted by repeat

82. 17 U.S.C. § 512(i)(1)(A).

infringers, it is not eligible for the safe harbor at all, even for material posted by others. Another example of conditional immunity is Danielle Citron and Benjamin Wittes's proposal to condition § 230 immunity on the platform making reasonable efforts to prevent the posting of illegal content.[83]

Both negligence and conditional immunity use a threshold to shape the platform's liability. But they do so in different ways. Negligence imposes liability for *specific content* that exceeds the threshold. Conditional immunity imposes liability for *all content* if total harm exceeds the threshold.

Despite this difference, conditional immunity and negligence have similar incentive effects. Under conditional immunity, the platform in effect has a "budget" of harm it can cause without incurring liability. If the platform has the choice of two items of content at x_1 and x_2 to leave up rather than investigate, where $x_1 < x_2$, it is always better off picking x_1 because $\lambda(x_1)h \leq \lambda(x_2)h$ (i.e., x_1 uses less of the harm budget) and $p(x_1) \geq p(x_2)$ (i.e., x_1 makes more profit for the platform). A similar argument shows that if the platform is choosing between two items to investigate rather than take down, it is always better off choosing the one to the left to investigate. And finally, the platform is best off spending all its budget—it should leave up content until the total harm equals T and then use investigation and takedown to ensure that no further harm ensues.

It follows, therefore, that the optimal level threshold level is

$$T = \int_0^{\underline{x_s}} \lambda(x)h\,dx.$$

If the regulator does so, the platform's behavior and social welfare are exactly the same as under negligence.

There are, however, meta-level concerns about conditional immunity. The first is that the calculation problem is more difficult. The regulator must be able evaluate $\lambda(x)$ and $p(x)$ at every point in $[0, \underline{x}]$, not just at $\underline{x_s}$. A second is that conditional immunity is much more sensitive to errors in this calculation process. Small errors in setting the negligence threshold lead to small changes in the platform's liability. But small errors in setting the conditional immunity threshold can lead to large changes in liability if a platform that thought it qualified for the immunity discovers it did not. The ISP Cox, for example, faced a $1 billion damage award after the court held that its repeat-infringer

83. Citron & Wittes, *supra* note 18, at 455–56.

policy was insufficient to qualify for the § 512(a) safe harbor.[84] Where a negligence regime acts as a price, a conditional immunity has characteristics of a sanction. Prices are more appropriate when the harm can be quantified but the appropriate level of activity is uncertain.[85]

V. POLICY RESPONSES TO OVERMODERATION

A. SUBSIDIES

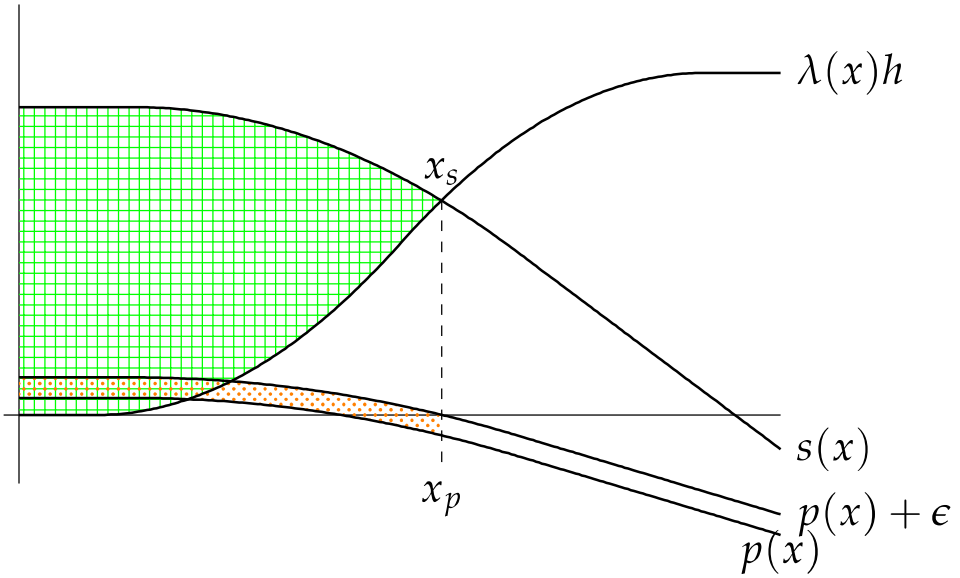
Many responses to overmoderation are familiar from telecommunications and intellectual-property law. One of the most common is *subsidies*, in which the government pays the platform to carry content. Figure 21 shows a case in which the government gives the platform a subsidy of ϵ for any content that it carries. Here, ϵ has been chosen so that it pushes the platform's profits up to the point that $x_p = x_s$ and it carries the socially optimal level of content.

There are at least three challenges in providing subsidies. First, the regulator must accurately estimate x_s, which requires an understanding both the value of content $s(x)$ and the harm of the content $\lambda(x)h$. Second, the regulator must choose an appropriate subsidy ϵ, which requires an understanding of the platform's revenues $p(x)$. And third, the subsidy must be one that the regulator is willing to pay. The orange dotted region in Figure 21 is money that must come from somewhere. It is not a welfare loss to society, just a wealth transfer (ignoring administrative costs and the distortionary effects of taxation, that is). Below-cost mail service is an example of this type of subsidy.

84. Sony Music Ent. v. Cox Commc'ns, Inc., 464 F. Supp. 3d 795, 837–39 (E.D. Va. 2020) (damage award); BMG Rights Mgmt. v. Cox Commc'ns, Inc. 881 F.3d 293, 301–05 (4th Cir. 2018) (safe harbor).

85. *See generally* Robert Cooter, *Prices and Sanctions*, 84 COLUM. L. REV. 1523 (1984).

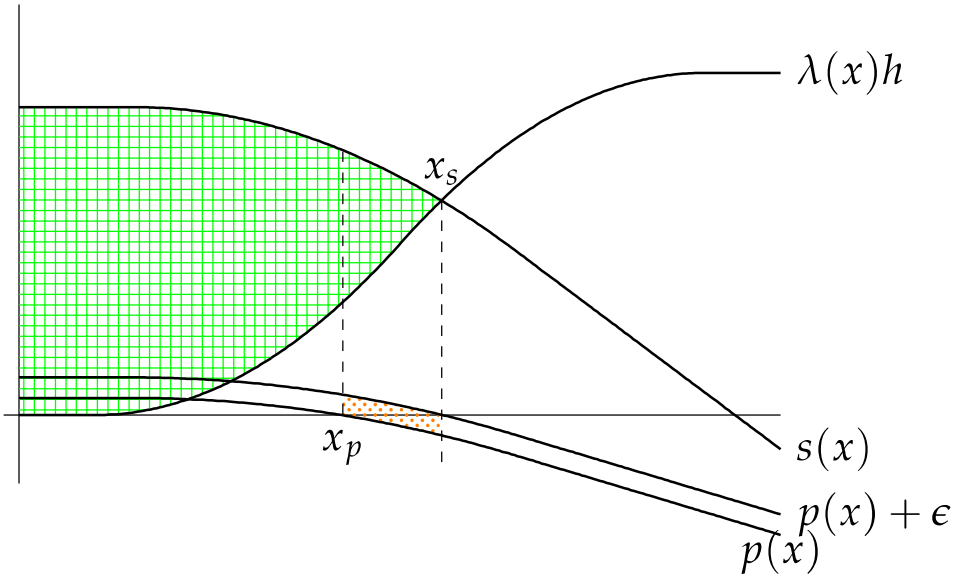
Figure 21: Flat subsidies



A partial solution to the third problem is *targeted subsidies*. Here, the government subsidizes content only in the range where subsidies make a difference in the platform's decision of whether to carry it (between x_p and x_s). This reduces the size of the subsidies required, but it increases the difficulty of the regulatory problem because now the regulator must be able to accurately estimate x^* and not just know the behavior of $p(x)$ in the neighborhood of x_s. For example, the FCC's Universal Service Fund is a targeted subsidy. It helps make broadband internet access more widely available by supporting its availability to people and communities for whom it would not otherwise be profitable for telecom companies to provide it.[86]

86. 47 C.F.R. pt. 54 (2022).

Figure 22: Targeted subsidies



Subsidies can also be provided indirectly, by subsidizing the users who create content and distribute it through platforms and the consumers who receive it. The idea here is that if distribution is more valuable to creators and consumers, they will be willing to pay more to distributors, thus shifting the $p(x)$ curve upwards. There is an argument that the copyright system has some of these features, although it is not typically described in these terms.

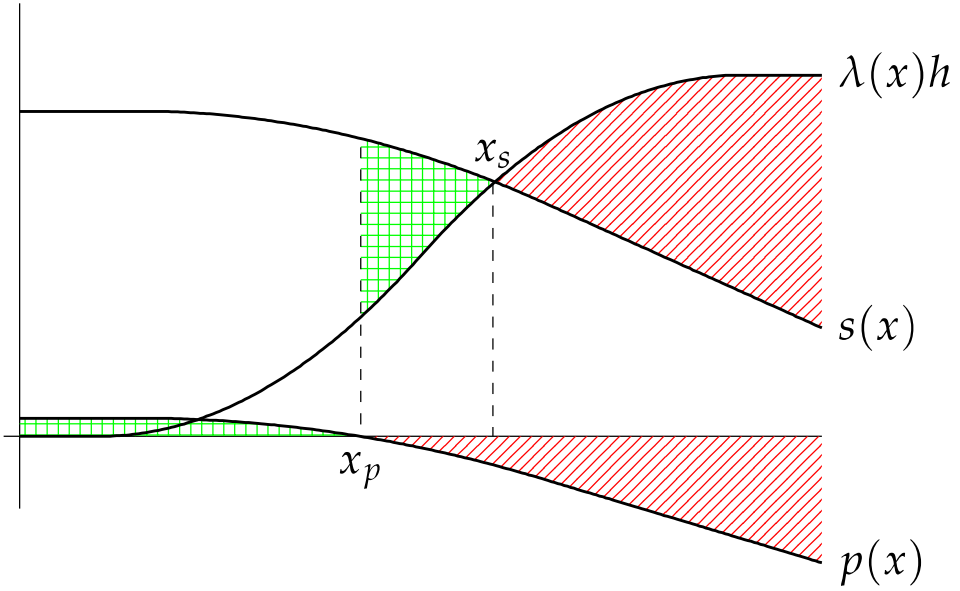
B. MUST-CARRY

Another response to overmoderation is to impose a *must-carry* rule, in which the platform must host all content submitted to it. Formally, the regulator forces the platform to set $x^* = x_{\max}$, i.e., the far right of the diagram. Compared to subsidies, a must-carry system is simpler to design and almost by definition requires less outlay. It also removes discretion from the platform, which may be a concern if the platform has a conflict of interest due to other business lines or does not agree with the regulator's understanding of which content is valuable. Something like this, for example, is a commonly advanced argument for network neutrality.

A must-carry rule, however, must satisfy two conditions to be justified compared with the baseline. First, it must actually result in hosting more worthwhile than worthless content. In Figure 23, the upper green gridded region is the positive-value content that must-carry causes to be hosted, and the upper red striped region is the negative-value content it also causes to be

hosted. If the red region is larger than the green one, must-carry is counter-productive because the bad additional content outweighs the good.[87]

Figure 23: Must-carry



A little more subtly, must-carry can also counter-productively drive a platform out of the market. In Figure 23, the lower green gridded region is the platform's profits from hosting the content it wants to, and the lower red striped region it is losses from hosting the content it is forced to. If the red region is larger than the green one, it is unprofitable for the platform to operate at all, and the platform will rationally shut down rather than comply with a must-carry mandate.

C. LAWFUL MUST-CARRY

An issue with a pure must-carry regime is that it compels platforms to carry content that society itself considers harmful, even illegal. So it is common to see must-carry mandates limited to "lawful" content. The FCC's Obama-era network neutrality regulations had such a carveout,[88] as do the Texas and

87. This analysis omits the investigation option because it is never rational for a platform to investigate content that it is just going to leave up anyway.

88. Safeguarding and Securing the Open Internet, 88 Fed. Reg. 76048, 76096 (Nov. 3, 2023) (to be codified at 47 C.F.R. § 8.2(b)) (prohibiting broadband providers from "block[ing] *lawful* content, applications, services, or non-harmful devices" (emphasis added)).

Florida social media must-carry bills whose constitutionality is currently being litigated.[89]

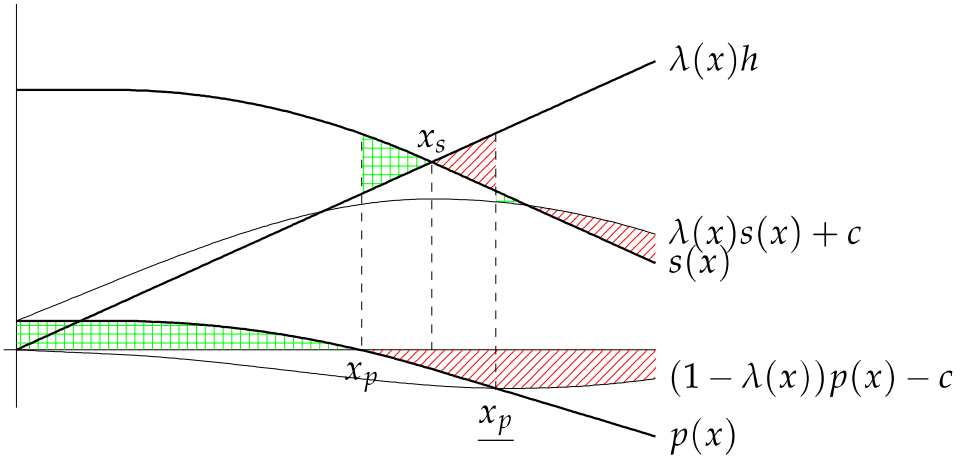
We can model a *lawful must-carry* rule by stating that the platform *must* host all harmless content, but it has discretion whether or not to host harmful content. Of course, to know with certainty whether content is harmless, the platform must investigate it. Thus, under lawful must-carry, the platform has two choices: it can either leave the content up without investigation, or it can investigate it and take it down if harmful.

As Figure 24 illustrates, the platform's marginal revenue from leaving up is $p(x)$, and its marginal revenue from investigation is $\big(1-\lambda(x)\big)p(x)-c$. Thus, the platform finds the two equal when $p(x)=-c/\lambda(x)$, which can only occur when the platform's profit $p(x)$ has gone negative. If these two curves meet at all, we call this intersection $\underline{x_p}$ because this is the point at which the platform starts investigating in the hopes of being able to find and remove harmful unprofitable content. To the left of $\underline{x_p}$, the platform leaves up content, so its profits and social welfare are as above in Section B. But to the right of $\underline{x_p}$, the platform investigates all content and takes down all harmful content. Compared with a flat must-carry requirement, the platform can reduce its losses from the content it is compelled to carry and thus may be better able to keep operating in the face of a lawful must-carry requirement.

Lawful must-carry can also be better for social welfare because the platform will filter out some content that is both harmful and unprofitable. Figure 24 shows that the welfare effects can be subtle and complex. $\underline{x_p}$ creates a discontinuity. To the left, social welfare is the benefits of all content $s(x)$ minus the harm of all content $\lambda(x)h$. To the right, investigation eliminates the harm $\lambda(x)h$ but introduces two new costs: the cost of foregone benefits from removed harmful content $\lambda(x)s(x)$ and the costs of investigation c.

89. *See* NetChoice, LLC v. Att'y Gen., 34 F.4th 1196 (11th Cir. 2022); NetChoice, LLC v. Paxton, 49 F.4th 439 (5th Cir. 2022).

Figure 24: Lawful must-carry



VI. EXISTING AND PROPOSED LAWS

A. SECTION 230

Section 230, in our model, is a blanket immunity regime. The platform is not liable for any harmful content, regardless of its knowledge and regardless of whether it has made any effort to investigate.[90] As discussed in Section II.G, such a regime is a reasonable response to the perverse incentives of *Stratton Oakmont, Inc. v. Prodigy Services Co.*[91] Platforms have their own commercial reasons to moderate content, so it is important not to create a system in which they are disincentivized from moderating at all.

Our model also illustrates the wide range of proposed reforms to § 230. These reforms have profoundly different economic consequences.

To begin, the Citron-Wittes proposal is a straightforward conditional immunity.[92] Courts would be asked to assess a platform's overall moderation efforts and to deny platforms the § 230 safe harbor if they fell beneath that threshold. It therefore functions like the RIP limitation on § 230 and can be expected to have some of the same consequences, including the occasional massive verdict against a platform that miscalculates the required level of effort and a corresponding *in terrorem* effect against other platforms that will cause

90. 47 U.S.C. § 230.

91. *See* Stratton Oakmont, Inc. v. Prodigy Servs. Corp., No. 031063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995) (holding a platform liable for user-posted content even where it lacked knowledge of the content).

92. *See* Citron & Wittes, *supra* note 18.

them to engage in overmoderation due to the uncertainty they face about their legal exposure.

Other scholars have proposed that *Zeran v. America Online, Inc.* should be overturned[93] and platforms be subject to common-law distributor liability.[94] This would in effect create a liability on notice regime, much like the notice-and-takedown system of § 512. Similarly, the Platform Accountability and Consumer Transparency (PACT) Act would have created a system for material that a court had determined to be unlawful and would have defined a platform to have knowledge only when it was provided with a copy of the court order and information reasonably sufficient to locate the material.[95] This is a substantial improvement on the deficiencies of notice-and-takedown under § 512 because it sets a meaningful threshold for sending an effective notice. On the other hand, the process of obtaining a court order will be slow and expensive, so this would be a solution only for egregiously harmful material.

B. SECTION 512

Our model sheds light on the notice-and-takedown regime of § 512 of the Copyright Act.[96] The basic rule of § 512 is that a hosting platform "shall not be liable for monetary relief . . . for infringement of copyright by reason of the storage at the direction of a user of [infringing] material."[97] This is a blanket immunity, but it is qualified by five (!) exceptions.

First, § 512(c)(1)(A) removes the platform's immunity as to specific material if it has "actual knowledge that the material . . . is infringing"[98] and the platform does not "act[] expeditiously to remove, or disable access to, the material."[99] This exception reflects the intuition that where the platform *has* performed an investigation into specific content, it can remove harmful items without affecting non-harmful items. It does not matter where along the $\lambda(x)$ curve the item falls; once the platform has knowledge, it must act.

93. Zeran v. Am. Online, Inc., 129 F.3d 327 (4th Cir. 1997) (holding that Section 230 provides a blanket immunity from defamation liability for platforms carrying third-party content).

94. *See, e.g.*, Shlomo Klapper, *Reading Section 230*, 70 BUFF. L. REV. 1237 (2022).

95. *See* S. 4066, 116th Cong. (2020); *see also* Daphne Keller, *CDA 230 Reform Grows Up: The PACT Act Has Problems, but It's Talking About the Right Things*, STAN. L. SCH.: CTR. FOR INTERNET & SOC'Y (July 16, 2020), https://cyberlaw.stanford.edu/blog/2020/07/cda-230-reform-grows-pact-act-has-problems-it%E2%80%99s-talking-about-right-things.

96. *See* 17 U.S.C. § 512.

97. 17 U.S.C. § 512(c)(1).

98. *Id.* § 512(c)(1)(A)(i).

99. *Id.* § 512(c)(1)(A)(iii).

Second, as discussed above, § 512(c)(1)(A)(ii) removes the platform's immunity as to specific content if it is "aware of facts or circumstances from which infringing activity is apparent" and fails to remove the content.[100] This is a negligence rule.

Third, § 512(c)(1)(B) removes the platform's immunity if it "receive[s] a financial benefit directly attributable to the infringing activity, in a case in which the service provider has the right and ability to control such activity."[101] This standard, which resembles but is not identical in application to the common-law vicarious-infringement standard,[102] is not in theory tied to the platform's knowledge at all. Instead, it is designed to smoke out situations in which a platform that could block infringement has especially bad incentives to turn a blind eye to it. In terms of our model, we think these are situations in which c is small (so that the platform has the "ability to control" infringement) and P is large (so that the platform has strong private incentives to allow as much infringement as it can). These are circumstances under which in the absence of liability, the platform might under-invest in investigating likely-to-be-infringing content.

Fourth, § 512(c)(1)(C) removes the platform's immunity if it receives a "notification of claimed infringement" and fails to remove it.[103] As discussed above, this creates a notice-and-takedown regime, which is effective only to the extent that sending a notice is a signal that conveys information.

And fifth, again as discussed above, the repeat-infringer provision of § 512(i) creates a conditional immunity.[104]

To summarize, the five limitations on the § 512(c) safe harbor all function in different ways. The actual-knowledge provision deals with cases where $c = 0$ and no investigation is required; the red-flag provision deals with cases where $\lambda(x)$ is high and the content is likely to infringe; the financial-benefit provision deals with cases where c is low and P is high so the platform has bad incentives not to investigate; the notice-and-takedown provision deals with cases where the copyright owner has taken on the investigative costs c itself; and the RIP provision requires the platform to keep overall infringement beneath a total threshold. Notably, four out of these five limitations have to do with investigation costs.

100. *Id.* § 512(c)(1)(A)(ii).

101. *Id.* § 512(c)(1)(B).

102. *See* R. Anthony Reese, *The Relationship Between the ISP Safe Harbors and the Ordinary Rules of Copyright Liability*, 32 COLUM. J.L. & ARTS 427 (2009).

103. 17 U.S.C. § 512(c)(1)(C).

104. *See supra* Section IV.D.

Section 512 also has important text on investigation. The safe harbor does not depend on "a service provider monitoring its service or affirmatively seeking facts indicating infringing activity."[105] A useful way to understand this statement is as creating a rule that a platform's choice of whether to investigate content is not a basis for liability. Only the fact that the platform hosts infringing content is a liability trigger, and the safe harbor is removed only when the platform's conduct falls into one of the five limitations above. These are all performance standards based on the platform's knowledge or activity *with respect to the infringing content*. The platform is free to arrange its activities as it chooses, investigating only the content it chooses to, as long as it acts when it has knowledge.

C. THE DIGITAL SERVICES ACT

The Digital Services Act (DSA) makes a number of interesting choices.[106] The first is that it sharply distinguishes between platforms that serve as a "mere conduit" and those that store information at the request of a user. A mere conduit is not liable for user-provided content and has no content-moderation obligations.[107] But a hosting service is liable only when it has knowledge (actual or red-flag) and fails to act.[108] Thus, mere conduits have a blanket immunity, while hosting services have a notice-and-takedown regime.[109]

Above, we criticized the two-track regime under pre-§ 230 law for creating a disincentive for platforms to engage in content moderation. The DSA's distinction is more sensible because it is tied to the nature of a platform's services rather than the nature of its moderation. A platform can qualify for the mere-conduit safe harbor (or the similar safe harbor for caching services[110]) only when it is completely passive with respect to the information, selecting neither the material nor its destination and playing no role in modifying the material.

The DSA's hosting safe harbor is in some respects broader than the safe harbor under § 512. It has actual-knowledge and red-flag exceptions, but it does not have anything that looks like the vicarious-liability provision of § 512

105. 17 U.S.C. § 512(m)(1).

106. *See* 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), 2022 O.J. (L. 277).

107. *See id.* art. 4.

108. *See id.* art. 6.

109. Section 512 draws a similar distinction, but only applies to copyright infringement, whereas the DSA applies to all "illegal activity or illegal content." *See id.* art. 6(1).

110. *See id.* art. 5.

or the conditional immunity of the RIP. While there is a requirement that platforms suspend service to users "that frequently provide manifestly illegal content," this is simply an independent requirement of law, not a condition on the safe harbor.[111] It therefore avoids some of the error costs and overdeterrence associated with the RIP. The DSA has a takedown procedure that is based not on private notices but on orders from appropriate authorities, which functions as a high-threshold notice-and-takedown procedure.[112]

The DSA also has a separate procedure for "notice and action mechanisms" that allow private parties to send notices that

> shall be considered to give rise to actual knowledge or awareness for the purposes of Article 6 in respect of the specific item of information concerned where they allow a diligent provider of hosting services to identify the illegality of the relevant activity or information without a detailed legal examination.[113]

This is a U.S.-style notice-and-takedown regime. And, most interestingly, it has a "trusted flagger" provision that allows member states to designate certain entities as trusted flaggers, whose notices of illegal material "are given priority and are processed and decided upon without undue delay."[114] A trusted flagger is required to act "diligently, accurately and objectively" and should therefore not send notices without suitable investigation.[115] This is a clever response to the signaling problem we discuss above with respect to notice-and-takedown under § 512.

The DSA emphasizes that platforms remain eligible for the safe harbors even if they "in good faith and in a diligent manner, carry out voluntary own-initiative investigations into, or take other measures aimed at detecting, identifying and removing, or disabling access to, illegal content."[116] This is an important limitation to prevent the *Stratton Oakmont* trap discussed in Section II.G above. Contrariwise, it adds that platforms have "[n]o general obligation to monitor the information" they carry "nor actively to seek facts or circumstances indicating illegal activity."[117] This is (like the corresponding provision in § 512) a way of emphasizing that the platform can choose not to investigate content and that those choices by themselves do not create liability.

111. *Id.* art. 23(1).
112. *See id.* art. 9.
113. *Id.* art. 16.
114. *Id.* art. 22.
115. *Id.*
116. *Id.* art. 7.
117. *Id.* art. 8.

VII. CONCLUSION AND FUTURE EXTENSIONS

Our model is deliberately simple. Nonetheless, it yields vivid, straightforward intuitions about a wide range of intermediary-liability problems. We hope that it can provide a clean foundation for modular extensions to model a wider range of fact patterns and legal responses.

Indeed, our treatment of liability on notice is intended as an example of how to extend our basic model. We introduced a parsimonious extension: a new type of actor (victims of harm), who can take two types of actions (investigate and give notice), and whose features are captured by a single parameter (their cost of investigation c_v). A more sophisticated treatment of liability on notice might add costs of sending notices (or of sending false notices) and allow victims and platforms to negotiate deals.

Other extensions of our model might introduce other actors, such as the users who post content in the first place. One could posit, for example, that these users know whether the content they are posting is harmful or not and have private gains from posting that are distinct from the platform's revenues but do not exhaust the social benefits their posting creates. Add in a feature to model the comparative difficulty of seeking enforcement against these users, and again, one is in a position to draw interesting conclusions. Perhaps these users might negotiate the price they pay for posting on the platform, or perhaps the platform competes with other platforms, and so on.

Another way in which the model presented in this Article might be limited is the assumption that harmful content is less profitable and has fewer positive spillovers. We made this assumption because it simplifies the analysis in our initial presentation. Our results would be robust if, for example, $p(x)$ is increasing but $\lambda(x)h$ increases faster than $p(x)$ (as measured by the slope). In other cases, however, it becomes possible for $s(x)$ and $\lambda(x)h$ to intersect multiple times—even infinitely often—and it is no longer rational for a moderator to set a single threshold x_s. Instead, as $s(x)$ and $\lambda(x)h$ take turns surging ahead, the moderator might choose to turn moderation on and off repeatedly. A similar point applies to the platform's revenues $p(x)$, and one might also consider whether the harms h and costs of investigation c should vary.

Although the analysis will be more mathematically difficult, there are important classes of content for which $s(x)$ and $p(x)$ plausibly increase even as the content becomes more likely to be harmful. The most scandalous accusations against public figures are both more likely to be false and more important to air publicly if true. Indeed, our analysis of § 512 vicarious liability suggests that it makes the most sense in a world where $p(x)$ increases with

$\lambda(x)$. The fact that space limitations prevent us from addressing this scenario in the depth a proper analysis would require should not be taken as a statement that the scenario does not occur or is not worth understanding when it does.

Another important set of extensions relates to error costs. We have considered errors by *platforms* about whether content is harmful. But our model assumes that courts eventually reach the truth. This assumption may not be warranted because courts themselves have an error rate and may classify harmful content as harmless or vice versa. In our discussion of negligence and conditional immunity, we noted that courts and regulators may mismeasure factors that go into setting and applying liability thresholds. If a court misunderstands the threshold or a platform's efforts, the consequences can be significant. Platforms must make their moderation decisions in the shadow of the possibility that courts could err.

WHEN THE DIGITAL SERVICES ACT GOES GLOBAL

Anupam Chander†

ABSTRACT

The European Union's Digital Services Act ("DSA") establishes a "meta law"—public regulation of the private regulation conducted by internet platforms. The DSA offers an attempt to balance private technological power with democratic oversight. The DSA will likely prove an attractive model for other governments to assert control over massive global internet platforms. What happens when other countries borrow its approach, in an instantiation of the vaunted Brussels Effect? This Article evaluates the DSA using the "Putin Test"—asking what if an authoritarian leader were given the powers granted by the DSA? The Article argues that authoritarians might well exploit various mechanisms in the DSA to enlarge their control over the dissemination of information, and, in particular, to target the speech of critics.

TABLE OF CONTENTS

DOI: https://doi.org/10.15779/Z38RX93F48

† Scott K. Ginsburg Professor of Law and Technology, Georgetown University; Visiting Fellow, Berkman Klein Center's Institute for Rebooting Social Media, Harvard University. I'm grateful to Pam Samuelson and Martin Senftleben for inviting me to participate in the stellar Berkeley symposium on the Digital Services Act, and to Eric Goldman, Florence G'sell, Martin Husovec, Daphne Keller, and my student Chris Haines for insightful comments, Christine Vlasic for helpful research assistance, and Marley Macarewich for excellent editing. All views expressed herein, and all errors, are my own.

I. INTRODUCTION

The European Union's Digital Services Act (DSA)[1] represents the most comprehensive effort by liberal democratic states to regulate content moderation by internet platforms. The DSA requires internet intermediaries—such as social media, hosting services, online marketplaces, app stores, and search engines—to adopt a wide array of measures designed to ensure transparency and reduce harmful material. Because these companies regulate the actions of their users, providing or denying service, amplifying or dampening speech, the DSA seeks to regulate how these companies engage in this private decision-making. It is thus a meta-content moderation law—public regulation of the private regulation conducted by internet platforms. This means it is a literal Meta law, the law of Meta Platforms, Inc. In April 2023, the European Commissioner for Internal Market Thierry Breton adapted a line from Spiderman to describe the DSA's regulatory philosophy: "With great scale comes great responsibility."[2] With this caution, Commissioner Breton announced the initial list of services subject to the law's strictest rules—namely, certain services provided by Alibaba, Amazon, Apple, Booking.com, Facebook, Google, Microsoft, Pinterest, Snapchat, TikTok, Twitter, Wikipedia, and Zalando.[3] To make the announcement, he chose Twitter (now "X"), one of the companies designated a "Very Large Online Platform" (VLOP) under the Act.[4]

The DSA is part of the European Union's efforts to bring private power under democratic control. As European scholar Florence G'sell notes, the DSA responds to the "trend towards the privatization and automation of

1. Regulation 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022, on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), O.J. (L 277) 1 (EU). The Act could perhaps more precisely be described the "Digital Intermediary Services Act," as it does not govern all digital services, but rather only internet platforms that serve as intermediaries for other goods and services. Recital 6 of the Act provides, "This Regulation should apply only to intermediary services and not affect requirements set out in Union or national law relating to products or services intermediated through intermediary services."

2. @ThierryBreton, TWITTER (Apr. 25, 2023, 6:30 AM), https://twitter.com/ThierryBreton/status/1650854765126107136.

3. *See* European Commission Press Release, Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines (Apr. 25, 2023), https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413. Zalando, the only designated platform that hails from an E.U. member state, has sued to void the designation. Molly Killeen, *Zalando Files Suit Against Commission Over Very Large Platform Designation*, EURACTIV (June 27, 2023), https://www.euractiv.com/section/platforms/news/zalando-files-suit-against-commission-over-very-large-platform-designation.

4. *Id.*

online speech control."[5] Margrethe Vestager, European Commissioner for Competition and Executive Vice President of the European Commission for A Europe Fit for the Digital Age, summarizes its goals: "It aims to protect online consumers from unsafe and illegal products, and it protects our right to speak freely online."[6] One prominent example of online speech control was the decision by many of the largest internet platforms to ban then-President Donald Trump in the wake of the January 6, 2021 attack on the Capitol by his supporters. European leaders, including then-German Chancellor Angela Merkel, criticized the ban.[7] French Finance Minister Bruno Le Maire declared: "What shocks me is that Twitter is the one to close his account. The regulation of the digital world cannot be done by the digital oligarchy."[8]

The Digital Services Act will likely follow its groundbreaking European predecessor, the Data Protection Directive, in achieving a global impact—what the scholar Anu Bradford has famously labeled a "Brussels Effect."[9] Americans, Bradford points out, may be surprised to find that "EU regulations determine . . . the privacy settings they adjust on their Facebook page."[10] Legal scholar Dawn Nunziato has written persuasively that the DSA, too, will likely carry a Brussels Effect, "influenc[ing] how social media platforms globally moderate content."[11] Large platforms are likely to apply various aspects of the DSA, such as transparency rules and perhaps dispute resolution systems, across their worldwide operations. I want to consider here another mechanism associated with the Brussels Effect—the explicit adoption of laws modeled on European rules by countries elsewhere.[12] That is, rather than relying on corporations to globalize European legal norms, governments can choose to do so themselves.

---

5. Florence G'sell, *The Digital Services Act: a General Assessment*, 1 CONTENT REGUL. IN THE EUROPEAN UNION: THE DIGITAL SERVICES ACT 85–86 (Antje von Ungern-Sternberg ed., 2023).

6. Margrethe Vestager, *Shared Objectives for Framing the Tech Economy*, 41 BERKELEY J. INT'L L. 137, 141 (2023).

7. Pierre-Raul Bermingham, *Merkel Among EU Leaders Questioning Twitter's Trump Ban*, POLITICO (Jan. 11, 2021), https://www.politico.eu/article/angela-merkel-european-leaders-question-twitter-donald-trump-ban.

8. *Id.*

9. Anu Bradford, *The Brussels Effect*, 107 NW. U. L. REV. 1, 22–26 (2012).

10. *Id.* at 3.

11. Dawn Carla Nunziato, *The Digital Services Act and the Brussels Effect on Platform Content Moderation*, 24 CHI. J. INT'L L. 115, 117 (2023).

12. Anu Bradford, who named the "Brussels Effect," observes that the de facto Brussels Effect can be "reinforced with a de jure Brussels Effect . . . when other countries' legislators affirmatively adopt the EU's strict standards." *The Brussels Effect*, *supra* note 9, at 8.

Introducing the Digital Services Act at the Berkeley symposium, Irene Roche Laguna, one of the European Commission architects of the Act, asked: "What would Putin do with this instrument?"[13] She answered, "I want to think that the DSA passes the Putin Test."[14] I think the "Putin Test" would go something like this: If Russia cut and pasted the European Union's Digital Services Act into Russian law, would that raise human rights concerns? Why single out Russia? Vladimir Putin has served continuously as either the President or Prime Minister of Russia since 1999. Putin's government uses legal tools to stifle dissent and target political opponents with politically-motivated charges.[15] In this Article, I propose to put the DSA through the Putin Test. We can also imagine versions of that test for other imperfect democracies such as Brazil, India, or Nigeria. These hypotheticals allow us to imagine a future Brussels Effect of the DSA.

Imagining a Russian Digital Services Act allows us to see how ruthless actors might deploy its provisions to target political enemies by utilizing its regulatory infrastructure—from digital services coordinators to trusted flaggers to local representatives—to ensure a favorable information environment in the country.

Some will consider this endeavor unfair or unreasonable. After all, the DSA was written by and for the European Union. It exists within a constitutional framework, the Charter of Fundamental Rights of the European Union, and national constitutional constraints.[16] The DSA also exists within the human rights framework covering the members of the Council of Europe through the European Convention on Human Rights—though, by that measure, so did Russia, which was a member of the Council of Europe until it was expelled in the wake of its invasion of Ukraine in 2022.[17]

There is much to praise in the DSA, as it increases transparency and risk assessment and mitigation. My argument is not that the DSA grants national or regional authorities clearly excessive power. The DSA is nowhere close to a

13. Irene Roche Laguna, Keynote Address at the University of California, Berkeley School of Law 27th Annual BTLJ-BCLT Symposium: From the DMCA to the DSA—A Transatlantic Dialogue on Online Platform Liability and Copyright Law (Apr. 6-7, 2023). Irene Roche Laguna is the Deputy Head of Unit dealing with the Implementation of the Telecom Regulatory Framework in the European Commission, at the Directorate-General for Communications Networks, Content and Technology (DG CONNECT).

14. *Id.*

15. U.S. DEP'T OF STATE, 2022 COUNTRY REPORTS ON HUMAN RIGHTS PRACTICES: RUSSIA (2022).

16. *See* Digital Services Act recital 3, 2022 O.J. (L 277).

17. *Russia Ceases to be a Party to the European Convention on Human Rights*, COUNCIL OF EUR. (Mar. 23, 2022), https://www.coe.int/en/web/portal/-/russia-ceases-to-be-a-party-to-the-european-convention-of-human-rights-on-16-september-2022.

grant of plenary control over the internet. Rather, it generally cabins itself within the rules of existing national laws that determine legality and illegality. Furthermore, the DSA does not provide government authorities the direct power to order the removal of material (though some provisions might be deployed in ways that approximate such power, as we shall see); rather it leaves questions of such removal powers to the national laws of the member states. The goal of this Article is not to criticize the DSA, but to begin to anticipate the ways that it, or laws modeled after it, can be abused by determined actors.

This Article proceeds as follows. Part II argues that the Digital Services Act will likely carry a Brussels Effect. Part III then applies the Putin Test to the DSA—what would happen were such a law adopted in Putin's Russia—and finds that there is cause for concern about the globalization of the DSA.

II. THE DIGITAL SERVICE ACT'S LIKELY BRUSSELS EFFECT

Before we evaluate the Digital Services Act as a part of law outside Europe, it is useful to understand that such a result is both likely and, to some extent, intended. This Part argues that the Digital Services Act will likely carry a Brussels Effect, both de facto through changes in the practices of multinational corporations, and de jure through changes in foreign law. This is not to suggest that the DSA will be adopted in whole either by corporations or governments worldwide, but rather to suggest that it will be substantially influential on digital content regulation well beyond Europe. Other states are likely to cite it to support their own efforts to regulate the content moderation practices of internet platforms.

As Dawn Nunziato argues, the DSA will likely have global impact.[18] The DSA will, she explains, "likely incentivize platforms to skew their content moderation policies toward the [European Union]'s approach. This is because the DSA levies huge financial penalties for violating its provisions, including maximum fines of six percent of a platform's annual worldwide turnover."[19] Given these potential fines, platforms will ensure that European hate speech norms are reflected in their community guidelines, at minimum for E.U. states. Many internet platforms may find it convenient to globalize their content policies written for E.U. member states, which would allow moderators around the world to be trained on a uniform set of policies to be applied globally with exceptions for specific categories of speech that have different

18. Nunziato, *supra* note 11, at 120.
19. *Id.*

rules, such as nudity.[20] Of course, even within the European Union, the definition of hate speech has varied among the member states.[21]

Furthermore, some of the DSA's obligations carry global implications. Parts of the DSA—such as those requiring platforms to provide explanations of why adverse actions were taken against users or requiring platforms to offer dispute settlement systems[22]—might be rolled out by at least some VLOPs globally. All covered platforms must publish annual reports on their content moderation practices, including their use of automated tools, training measures, and complaints received.[23] While these reports may limit their information to their European operations, they are likely to provide some insight into their global operations as well. Because some firms may adopt largely unified practices across the world, these transparency reports may prove useful to users across the world, not just in the European Union. Internet intermediaries must also publish their "terms and conditions," which include their community guidelines.[24] VLOPs and VLOSEs (very large online search engines) must provide a summary of their terms and conditions in machine-readable form.[25]

A number of mechanisms might globalize the DSA.[26] First, companies could adopt DSA-compliant practices worldwide. This is a common form of

20. *See* Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power over Online Speech*, HOOVER WORKING GROUP ON NATIONAL SECURITY, TECHNOLOGY, AND LAW, AEGIS SERIES PAPER 1902 (Jan. 29, 2019), https://s3.documentcloud.org/documents/5699593/Who-Do-You-Sue-State-and-Platform-Hybrid-Power.pdf ("[P]latforms' operational preference [is] for a single set of rules. Teams that review massive volumes of user content struggle with logistics and enforcement consistency in the best of circumstances. Enforcing dozens of different rules around the world would, as Facebook's Monika Bickert has pointed out, be 'incalculably more difficult' than applying a single, consistent set of Community Guidelines. For social networks and other communications platforms, inconsistent rules also create bad user experiences, interfering with communication between people in different countries. Maintaining a single set of standards—and perhaps expanding them to accommodate national legal pressure as needed—is much easier."); *cf.* ANU BRADFORD, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD 161 (2020) (arguing that U.S. internet platforms "made a strategic choice to switch to a more restrictive European style of hate speech regulation").

21. Natalie Alkiviadou, *Regulating Hate Speech in the EU*, *in* ONLINE HATE SPEECH IN THE EUROPEAN UNION: A DISCOURSE-ANALYTIC PERSPECTIVE 6–7 (2017) (observing that "there is little coherence amongst EU member states on the definition of hate speech").

22. Digital Services Act art. 20, 2022 O.J. (L 277) (internal complaint-handling system); *id.* at art. 21 (out-of-court dispute settlement).

23. *Id.* art. 15.

24. *Id.* art. 14.

25. *Id.* art. 14(1).

26. Charlotte Siegmann and Markus Anderljung offer a set of possible mechanisms to globalize the EU's Artificial Intelligence Act, which is currently being finalized:

the Brussels Effect in Anu Bradford's account—when companies align their global practices with Brussels' rules largely out of possible efficiency of adopting those same standards worldwide.[27] This is also the main mechanism in Nunziato's account of the global effects of the DSA.[28]

Second, governments might find much to envy in the Digital Services Act—which validates burgeoning efforts to bring the internet under government control, provides special tools for speeding up the removal of illegal content under local law, includes procedural rules that might limit the power of platforms to label or suppress other content, conveys power to evaluate risk mitigation measures, and sets out "break glass" crisis control mechanisms—complete with the possibility of getting six percent of the company's global revenue for violations.

A third mechanism is possible as well. The European Union could itself promote the DSA as a global model, perhaps incorporating parts of it into its model free trade agreements. Here, the DSA does not offer a mechanism like the adequacy determination used in the European Union's General Data Protection Regime (GDPR), where foreign governments might seek to align their laws to the European standard in order to win easier access to digital trade with the European Union.[29] Because European countries are generally seen as well-governed, democratic, and compliant with human rights norms,

---

- Foreign jurisdictions may expect EU-like regulation to be high quality and consistent with their own regulatory goals.
- The EU may promote its blueprint through participation in international institutions and negotiations.
- A de facto Brussels Effect with regard to a jurisdiction increases its incentive to adopt EU-like regulations, for instance by reducing the additional burden that would be placed on companies that serve both markets.
- The EU may actively incentivise the adoption of EU-like regulations, for instance through trade rules.

Charlotte Siegmann & Markus Anderljung, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, CTR. FOR GOVERNANCE of AI 5 (Aug. 2022), https://arxiv.org/pdf/2208.12645.pdf.

27. *Cf.* Anu Bradford, *The Brussels Effect: How the European Union Rules the World*, OXFORD UNIV. PRESS 232 (2000) (describing the de facto Brussels Effect as "common").

28. Nunziato, *supra* note 11, at 115 (arguing that the DSA "will incentivize platforms to skew their global content moderation policies toward the EU's instead of the U.S.'s balance of speech harms versus benefits").

29. *See* Regulation 2016/679 of the European Parliament and the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119/1) 1 (EU), art. 45.

they may be especially attractive generators of legal norms. This last possibility is the focus of this Article—governments across the world borrowing the European approach for their own laws.

There would be good reason for this adoption. The problems of disinformation, hate speech, communal violence, and election interference are particularly acute in many countries in the Global South.[30] The concerns motivating the DSA in Europe are shared across the world.[31]

At the same time, institutional capacity and resources can be more limited, and civil society institutions and an independent press more fragile. Furthermore, states are often at grave risk of democratic backsliding, or already have authoritarian tendencies, with serious concerns about freedom of expression. This makes the adoption of similar rules in those states more likely to lead to abuse.

A Brussels Effect is likely to occur as a byproduct of the European Union's efforts to regulate the internet enterprises that serve its own population. A Brussels Effect, however, is not an official ambition of the DSA. At times, however, one can see some European regulators express the hope that the DSA might offer the world what they believe are more enlightened digital regulatory standards.

One E.U. leader has hinted at a hope that the DSA might create a Brussels Effect. European Union President Charles Michel has embraced the Brussels Effect and included the DSA in the list of E.U. rules that might have such an effect. At the Munich Security Conference on February 20, 2022, he stated:

> There is something else that is very important: our regulatory power, often called the 'Brussels effect.' Our standards, inspired by our European values, tend to become global standards. And this is true in many sectors . . . In the digital field, the General Data Protection Regulation (GDPR) had a similar effect, and we are working on our Digital Services Act and our Digital Markets Act.[32]

30. Vittoria Elliott, Nilesh Christopher, Andrew Deck & Leo Schwartz, *The Facebook Papers Reveal Staggering Failures in the Global South*, REST OF WORLD (Oct. 26, 2021), https://restofworld.org/2021/facebook-papers-reveal-staggering-failures-in-global-south (providing examples of disinformation in the Global South and arguing that Facebook underinvests in content moderation there).

31. António Guterres, *Guardrails Urgently Needed to Contain "Clear and Present Global Threat" of Online Mis- and Disinformation and Hate Speech, Says UN Secretary-General*, AFRICA RENEWAL (June 11, 2023), https://www.un.org/africarenewal/magazine/june-2023/guardrails-urgently-needed-contain-%E2%80%9Cclear-and-present-global-threat%E2%80%9D-online-mis.

32. European Council Press Release 130/22, Remarks by President Charles Michel at the Munich Security Conference (Feb. 20, 2022).

Despite President Michel's warning elsewhere in this speech of "massive sanctions" upon Russian aggression in the Ukraine, Putin ordered Russian troops to invade a few days later.[33] President Michel had linked the DSA to the Brussels Effect earlier as well as part of the European Union's "regulatory power."[34] As President Michel explained in 2021, the Brussels Effect stems from the European Union's position as one of the world's largest markets:

> We have unique and undeniable strengths. Our market of 450 million people. And with it, comes our regulatory power. The famous "Brussels effect"—that enables us to set the highest standards for our citizens, while projecting these standards across the world. This is especially true in the digital domain. Take our General Data Protection Regulation (GDPR) in 2016. And today's Digital Services Act and Digital Markets Act, proposed by the Commission.[35]

Here, President Michel mentions the possibility that the DSA might be one of the standards that the European Union "project[s] . . . across the world."[36]

It is important to recognize that the hope to bring Europe's local rules to a global stage is not merely the assertion of power for its own sake. As President Michel explained in 2021:

> And as our Union has taken shape, it has also become a project of influence. Our large single market has made us the biggest trader in the world and, in turn, the largest exporter of standards—known as the 'Brussels effect.' Yet, the standard we propagate most successfully in our neighbourhood is democracy, fundamental values, and the rule of law.[37]

President Michel's hope, shared with many European regulators, is that Brussels can lead the way towards a democratic and liberal world order founded on fundamental rights.

At the initial level, the Brussels Effect can be understood simply as a descriptive account of regulatory globalization with Brussels driving global regulatory standards across a variety of domains. But it can sometimes also

---

33. Austin Ramzy, *The Invasion of Ukraine: How Russia Attacked and What Happens Next*, N.Y. TIMES (Feb. 24, 2022), https://www.nytimes.com/2022/02/24/world/europe/why-russia-attacked-ukraine.html.

34. European Council Press Release, *supra* note 32.

35. European Council Press Release 63/21, Digital sovereignty is central to European strategic autonomy - Speech by President Charles Michel at "Masters of digital 2021" online event (Feb. 3, 2021).

36. *Id.*

37. European Council Press Release 666/21, Speech by President Charles Michel at the opening session of the Bled Strategic Forum (Sept. 1, 2021).

become a normative goal—that the rest of the world should move towards E.U. rules, perhaps especially for global digital technologies. The Brussels Effect consists not just in a descriptive account of the globalization of a legal norm, but rather an occasional hope that the legal rules and standards of the European Union will become global rules and standards.

Anu Bradford has noted that the civil law style of continental Europe might have special appeal for developing countries:

> [T]he civil law tradition of the [European Union] typically leads to precise and detailed rules, drafted in multiple languages, which are easier to emulate in developing countries that may have less-skilled administrative agencies and judiciaries. The Brussels Effect presents these countries with an opportunity to outsource their regulatory pursuits to a more resourceful and experienced agency.[38]

III. PUTIN'S DSA

We now turn to the Putin Test for the DSA. In order to imagine the Russian Digital Services Act—or the Nigerian Digital Services Act or the Indian Digital Services Act—we need to understand both the institutions and the norms embedded within the DSA.

The normative focus of the DSA is to promote more vigorous takedown of illegal speech by internet platforms and, at the same time, provide due process-style rights for speakers who are disciplined by internet platforms.[39] As Florence G'sell observes, the DSA's "goal is to encourage [online platforms] to fight objectionable content while respecting users' fundamental rights."[40] Both of these are, of course, noble goals and are likely shared by many governments across the world. Like the governments in the European Union, many governments are concerned about disinformation or hate speech circulating online. At the same time, they are rightly concerned that internet platforms exercise extraordinary power over global speech—deciding who speaks, who doesn't, who can monetize their speech via advertising, and whose speech is promoted or demoted on their platforms.

38. Anu Bradford, *The European Union in a Globalized World: The "Brussels Effect,"* 2 REVUE EUROPÉENNE DU DROIT 75, 75 (Mar. 2021).

39. *Questions and Answers: Digital Services Act*, EUROPEAN COMM'N (Apr. 25, 2023), https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 (stating that the Digital Services Act provides "[e]ffective safeguards for users, including the possibility to challenge platforms' content moderation decisions based on a new obligatory information to users when their content gets removed or restricted").

40. G'sell, *supra* note 5, at 86 ("[The Digital Service Act's] goal is to encourage [online platforms] to fight objectionable content while respecting users' fundamental rights").

But there are also more self-interested reasons for many governments to embrace these goals. After all, many governments complain about what they believe to be disinformation on internet platforms, especially disinformation about their own activities. And governments are annoyed when their own speech is labeled as misleading or false, disabled from amplification, or removed altogether without appropriate process from the governments' perspective. This makes the various mechanisms that the DSA offers to achieve its twin goals attractive for governments with authoritarian tendencies. We turn to those mechanisms now.

A. DSA: THE DIGITAL SERVICES COORDINATOR

The Digital Services Act relies on a variety of enforcers, some newly created and some currently existing. Some of these enforcers operate at the national level, and some at the E.U. level.[41] E.U.-level enforcement is a response to concerns of insufficiently strict enforcement by the national or sub-national data protection authority charged with regulating any particular data gathering or processing enterprise.[42] Separate national regulators increase the variation in enforcement priorities, and even quality among the regulators, but also increase the local knowledge they have.

The DSA establishes a new regulator—the Digital Services Coordinator.[43] This is a position at the national level, not at the Europe-wide level, and thus requires the creation of at least twenty-seven such Coordinators across the Union.[44] A single supranational regulator would, in theory, increase uniformity in the application of the law (potentially at the price of some local knowledge), but it would not make sense here because criminal laws related to speech have not been harmonized across the European Union. The national Digital Services Coordinators will themselves coordinate through a newly created European Board for Digital Services, with each member state having one representative on that Board.[45]

---

41. Supervision and enforcement powers are divided between the Member State in which the main establishment of the intermediary service is located and the Commission itself, which shall work "in close cooperation." Digital Services Act art. 56, 2022 O.J. (L 277).

42. G'sell, *supra* note 5, at 22 ("The DSA framework grants the EU Commission significant supervisory and enforcement powers, a departure from the usual jurisdiction of Member States. This may be attributed to criticisms of the country-of-origin principle, which gives exclusive jurisdiction to Irish regulators since many large technology companies are based in Ireland").

43. Digital Services Act art. 49, 2022 O.J. (L 277).

44. Member States are required to designate one or more Digital Services Coordinators. *Id.*

45. *Id.* arts. 61–62.

The title suggests a relatively ministerial, administrative role, but the powers vested in the Coordinator are substantial.[46] The Coordinator has a significant role in the enforcement of the DSA. Like the Act itself, the title of the role, Digital Services Coordinator, is somewhat of a misnomer because the role does not give jurisdiction over all digital services. Rather, it gives jurisdiction only over internet platforms that serve as intermediaries for other goods and services.[47]

The Coordinator can request data from VLOPs or VLOSEs.[48] Importantly, there is a civil liberty constraint: these requests must

> take due account of the rights and interests of the providers of very large online platforms or of very large online search engines and the recipients of the service concerned, including the protection of personal data, the protection of confidential information, in particular trade secrets, and maintaining the security of their service.[49]

The Coordinator receives user complaints and assesses them.[50] The Coordinator also has the power to investigate the complaints.[51] The Coordinator can demand information from platforms related to a suspected infringement of the DSA.[52] This includes the powers to conduct on-site investigations and to seize information, regardless of storage medium.[53]

---

46. The European Union did not make the mistake of the U.S. State Department when it stood up a new body it called "The Disinformation Governance Board." Amanda Seitz, *Disinformation Board to Tackle Russia, Migrant Smugglers*, ASSOCIATED PRESS (Apr. 28, 2022), https://apnews.com/article/russia-ukraine-immigration-media-europe-misinformation-4e873389889bb1d9e2ad8659d9975e9d. Senator Josh Hawley denounced the board, arguing that "Homeland Security has decided to make policing Americans' speech its top priority." @HawleyMO, TWITTER (Apr. 27, 2022, 4:01 PM), https://twitter.com/HawleyMO/status/1519406288756785152?s=20. While such a role was not the intention of the State Department, the name invited this confusion about its purpose.

47. Digital Services Act recital 6, 2022 O.J. (L 277) ("This Regulation should apply only to intermediary services and not affect requirements set out in Union or national law relating to products or services intermediated through intermediary services").

48. *Id.* art. 40(1).

49. *Id.* art. 40(2).

50. *Id.* art. 53.

51. *Id.* art. 51(1).

52. *Id.* art. 51(1)(b) (granting the Digital Services Coordinator "the power to carry out, or to request a judicial authority in their Member State to order, inspections of any premises that those providers or those persons use for purposes related to their trade, business, craft or profession, or to request other public authorities to do so, in order to examine, seize, take or obtain copies of information relating to a suspected infringement in any form, irrespective of the storage medium").

53. The Coordinator can enforce its orders by seeking a competent judicial authority to order the platform to temporarily cease services. *Id.* art. 51(2)(b).

The Coordinator chooses those who will be the "trusted flaggers" for platforms, whose requests for takedowns are to be prioritized.[54] The Coordinator can grant trusted flagger status only to those entities meeting the following conditions:

> (a) it has particular expertise and competence for the purposes of detecting, identifying and notifying illegal content;
>
> (b) it is independent from any provider of online platforms;
>
> (c) it carries out its activities for the purposes of submitting notices diligently, accurately and objectively.[55]

Once designated, trusted flaggers are supposed to notify platforms of illegal material on their services, and the platforms are to act upon those notices "without undue delay."[56] Because of the special power of trusted flaggers to cause rapid suppression of speech online, the selection of which entities are entrusted with that power is of great significance. What if the approved "trusted flaggers" are not in fact to be trusted?

The Coordinator also has a critical role in the access that researchers will have under the DSA to information held by internet platforms: the Coordinator determines who is a "vetted researcher."[57]

The Coordinator also has broad enforcement powers, which include "the power to order the cessation of infringements and, where appropriate, to impose remedies proportionate to the infringement and necessary to bring the infringement effectively to an end, or to request a judicial authority in their Member State to do so . . . ."[58] In significant part, the Digital Services Coordinator exercises power over digital services companies that have their main establishment (or their legal representative) in the jurisdiction of that Coordinator.[59]

The Coordinator can issue extraordinary fines that could reach amounts that are historic for all penalties: up to "6% of the annual worldwide turnover of the provider of intermediary services concerned in the preceding financial year."[60] This is a fifty percent higher fine than available under the GDPR.[61] The Coordinator can also seek a judicial order to temporarily suspend a

54. *Id.* art. 22.
55. *Id.* art. 22(2).
56. *Id.* art. 22(1).
57. *Id.* art. 40(8).
58. *Id.* art. 51(2)(b).
59. *Id.* art. 3(n) (providing definition).
60. *Id.* art. 52(3).
61. General Data Protection Regulation art. 83(4), 2016 O.J. (L 119/1).

platform under appropriate circumstances.[62] The Coordinator also certifies dispute settlement bodies.[63]

Each of these roles permits a personally, politically, or ideologically-motivated Coordinator to exercise those powers not on behalf of all of the people of the country, but rather a particular interest. Political or personal preferences might favor flaggers aligned with those preferences. Accredited dispute settlement bodies might favor a particular viewpoint. Researchers who may have better relations with the Coordinator may be more likely to be approved, which may help produce studies that favor the government's viewpoint. Indeed, a provision designed to allow the public to scrutinize platform actions through outside research could be weaponized to strengthen government control.[64]

So, what might Vladimir Putin do with a Russian Digital Services Coordinator? Such a Coordinator might exercise his or her statutory powers in the interests of Putin himself. For example: the Russian Digital Services Coordinator might compel information from wayward platforms about critics of the Ukraine invasion, select trusted flaggers that mark opposition speech as illegal, approve researchers that were sympathetic to the Russian strongman, select dispute settlement bodies favorable to the government, and seek a judicial order to temporarily suspend a recalcitrant platform. A Russian Digital Services Coordinator might well become a Digital Czar.

B. DSA: CRISIS PROTOCOLS AND EMERGENCY POWERS

Finalized after the Russian invasion of Ukraine, the Digital Services Act includes special emergency-type powers.[65] First, it empowers the European Commission to establish crisis protocols for VLOPs and VLOSEs.[66] Second, the DSA also grants the Commission the power to issue guidelines for the risk mitigation measures that the platforms undertake.[67] Specifically, the Commission has the power to order "interim measures" against VLOPs and VLOSEs "where there is an urgency due to the risk of serious damage for the recipients of the service."[68]

It makes sense to prepare for inevitable crises with protocols in place to respond. And certain crises will demand immediate response. But emergency powers raise risks of abuse. Governments could use them to target what the

62. Digital Services Act art. 51(3)(b), 2022 O.J. (L 277).
63. *Id.* art. 21(3).
64. I thank Daphne Keller for this important insight.
65. G'sell, *supra* note 5, at 103.
66. Digital Services Act art. 36(1), 2022 O.J. (L 277).
67. *Id.* art. 35.
68. *Id.* art. 70.

governments believe to be election disinformation produced by the opposition, or communal discontent that might undermine those in power.

Russia again provides a warning. Prior to elections in 2021, Russian authorities ordered Apple and Google to remove an app created by supporters of opposition leader Alexei Navalny. Russian authorities claimed that the app supported a political movement that had been outlawed as extremist.[69] Russia's Roskomnadzor—the Federal Service for Supervision of Communications, Information Technology, and Mass Media—banned Facebook in the wake of the Ukraine invasion, arguing ironically that Facebook was restricting "the free flow of information" because of its constraints on Russian state media.[70] Washington Post technology writer Will Oremus aptly described this Russian claim as "Orwellian."[71]

A Russian DSA would offer Putin or his designated Digital Services Coordinator powers that would allow such repressive measures. A Russian DSA would allow the Roskomnadzor to designate trusted flaggers that would label any criticism of Putin as defamatory and thus illegal. A Russian DSA would permit Putin to order the removal of the Navalny supporters' political app on the grounds that it was carrying illegal content—in this case, election-related information provided by an opposition candidate. And it would permit Putin to ban Facebook itself on the ground that it was violating Russian law by interfering with Russian state propaganda.

C. DSA: RISK MITIGATION

One of the DSA's principal regulatory mechanisms is a requirement that VLOPs and VLOSEs undertake risk assessments and then put in place risk mitigation measures.[72] The companies must perform an initial risk assessment and then perform annual assessments and additional risk assessments when introducing new functionalities that might raise risks. The bulk of the risk assessment and mitigation work is thus assigned to the companies themselves.

However, there remains a role for the government regulators. The risk assessments must be provided, upon request, to the relevant Digital Services Coordinator, as well as to the European Commission itself.[73] The Board, in

69. Anton Troianovski & Adam Satariano, *Google and Apple, Under Pressure from Russia, Remove Voting App*, N.Y. TIMES (Sept. 23, 2021), https://www.nytimes.com/2021/09/17/world/europe/russia-navalny-app-election.html.

70. Will Oremus, *The Real Reason Russia is Blocking Facebook*, WASH. POST (Mar. 5, 2022), https://www.washingtonpost.com/technology/2022/03/05/russia-facebook-block-putin-ban-roskomnadzor.

71. *Id.*

72. Digital Services Act arts. 34–35, 2022 O.J. (L 277).

73. *Id.* art. 34(3).

cooperation with the Commission, shall publish best practices for risk mitigation. The Commission, in cooperation with the Digital Services Coordinators (which presumably would occur through the Board), may issue guidelines that present best practices and recommend possible measures.[74]

These mechanisms seem reasonably measured. However, an authoritarian state could use the recommendation and guidelines powers to pressure platforms to implement rules to control speech.

D. DSA: LOCAL REPRESENTATIVE

The Digital Services Act requires foreign-based digital intermediaries that serve the European Union to designate a local legal representative.[75] This representative could be held liable for non-compliance. Jason Pielemeier of the Open Network Initiative has worried that similar laws that require a human representative can amount to "hostage-taking" laws.[76] The European Union's version explicitly permits a corporate entity to play the role of legal representative.[77] Thus, the European Union's version of this requirement does not offer the opportunity for a government to threaten local employees with jail if they do not comply—only the financial consequences of having their legal representative fined (extensively, as the case may be). Yet, other governments might implement the local representative requirement to require physical employees. As we will see, Twitter apparently agreed to establish a local office in Nigeria in order to be permitted to return to Nigeria after it censored tweets by the Nigerian President for promoting violence.[78]

Russia passed a so-called "Landing Law" in 2021 to require physical presence in Russia of certain internet platforms, including through a branch,

74. *Id.* art. 35(2).

75. *Id.* art. 13.

76. Vittoria Elliott, *New Laws Requiring Social Media Platforms to Hire Local Staff Could Endanger Employees*, REST OF WORLD (May 14, 2021), https://restofworld.org/2021/social-media-laws-twitter-facebook; Jason Pielemeier, *Mind the Gap: The UK is About to Set Problematic Precedents on Content Regulation*, JUST SEC. (Mar. 6, 2023), https://www.justsecurity.org/85358/mind-the-gap-the-uk-is-about-to-set-problematic-precedents-on-content-regulation; Asha Allen & Ophélie Stockhem, *A Series on the EU Digital Services Act: Tackling Illegal Content Online*, CTR. FOR DEMOCRACY & TECH. (Aug. 22, 2022), https://perma.cc/HG99-U28Q ("The threat of prosecution or imprisonment of employees or representatives if platforms do not comply with government demands even if unjustified, as seen during the recent Russian Federal elections and cases in Brazil, poses a significant risk to human rights and freedom of access to information.").

77. Digital Services Act art. 13(1), 2022 O.J. (L 277) (specifying that "a legal or natural person" can serve as a legal representative for purposes of the Act).

78. *See infra* notes 89–90 and accompanying text.

representative office, or other entity within the country.[79] It is unclear whether a legal representative is enough, or whether actual employees are necessary to comply with the physical presence requirement.[80] A Russian news website claims that of the large foreign internet service providers, "only Apple and Spotify" have fully 'landed' in Russia.[81] The human rights group Article 19 worries that this Landing Law "could be used to suppress freedom of expression and access to information by making internet companies vulnerable to online content removal requests or demands to disclose users' personal data from the authorities."[82] Apple and Google removed the app by supporters of Alexei Navalny only after Russian authorities threatened to arrest their local employees.[83]

E. THE DSA ELSEWHERE

It is not only Putin that may relish the powers of a Digital Services Act. Many governments across the world may embrace the ability to rapidly take down content they believe to be illegal and to punish platforms severely for lack of compliance with government views of what content is permissible or harmful.

Indeed, we see similar moves in laws across the world. Brazil's new "fake news" law also adopts crisis protocols and state-supervised risk mitigation, all in the service of slowing the spread of misinformation online.[84] However, "critics decry it as draconian, rushed and open to abuse by special interests."[85]

---

79. [Federal Law of the Russian Federation on Activities of Foreign Persons on the Information and Telecom Network "Internet" in the Territory of the Russian Federation], Sobranie Zakonodatel'stva Rossiĭskoĭ Federatsii [SZ RF] [Russian Federation Collection of Legislation] 2011, No. 236.

80. *Physical Presence Requirements for Foreign Tech Companies in Russia: How to Respond*, BAKER MCKENZIE, https://www.bakermckenzie.com/-/media/files/insight/publications/resources/webinar-re-landing-law--faq.pdf.

81. *Law On Landing in Russia (Digital Residency)*, TADVISER (July 15, 2022), https://tadviser.com/index.php/Article:Law_on_Landing_in_Russia_(Digital_Residency)#Authorities_will_tighten_the_law_.22on_landing.

82. *Russia: Internet Companies Must Challenge 'Landing Law' Censorship*, ARTICLE 19 (Jan. 21, 2022), https://www.article19.org/resources/russia-internet-companies-must-challenge-censorship-under-new-law.

83. Troianovski & Satariano, *supra* note 69.

84. Joan Barata, *Regulating Online Platforms Beyond the Marco Civil in Brazil: The Controversial "Fake News Bill,"* TECH POL'Y PRESS (May 23, 2023), https://techpolicy.press/regulating-online-platforms-beyond-the-marco-civil-in-brazil-the-controversial-fake-news-bill.

85. Brian Harris & Hannah Murphy, *Brazil's Lawmakers to Vote On 'Fake News' Bill Opposed by Tech Groups*, FIN. TIMES (Apr. 30, 2023), https://www.ft.com/content/827326e9-7433-4fb4-9fb5-36a76658d106; *Brazil: Reject "Fake News" Bill*, HUMAN RIGHTS WATCH (June 24, 2020), https://www.hrw.org/news/2020/06/24/brazil-reject-fake-news-bill.

Many have also criticized similar moves in Indian intermediary liability rules.[86] A proposed Digital India Act will "regulate Big Tech," but might raise similar concerns.[87] The Indian government labeled a BBC documentary about the Prime Minister defamation, and the Indian information technology ministry ordered Twitter to take down tweets promoting that documentary.[88]

In 2021, after Twitter deleted tweets by the Nigerian President because it found that they might promote violence, the government banned Twitter from the country. Twitter negotiated a return half-a-year later, with the government proclaiming that Twitter agreed to its terms.[89] Twitter reportedly agreed to open a local office, pay taxes, and be sensitive to national security.[90] Would a Nigerian DSA give the government power over all of the speech platforms, committing them to not censor government speech, for example?

Even in the European Union, uses of the authorities provided by the DSA might raise questions. In July 2023, French Prime Minister Emmanuel Macron floated the possibility of a shutdown of TikTok, Snapchat, and Telegram after accusing them of contributing to the riots that followed the police shooting of

86. Tejasi Panjiar & Prateek Waghre, *Many Mysteries of 'Digital India Bill,'* INTERNET FREEDOM FOUND. (Feb. 20, 2023), https://internetfreedom.in/many-mysteries-of-the-digital-india-bill; Aarathi Ganesan, *Why does the Delhi HC Think Search Engines are Responsible For Taking Down Non-Consensual Intimate Images?*, MEDIANAMA (May 11, 2023), https://www.medianama.com/2023/05/223-search-engines-non-consensual-intimate-images-delhi-hc.

87. Umang Poddar, *Digital India Bill May Change the Internet as We Know It*, SCROLL.IN (Mar. 24, 2023) https://scroll.in/article/1045731/the-proposed-digital-india-bill-may-change-the-internet-as-we-know-it#:~:text=In%20the%20recent%20past%2C%20too,power%20to%20take%20down%20content.

88. Prasanna S, *Why Did Twitter Agree to Take Down Tweets Linking to BBC Documentary?*, WIRE (Feb. 18, 2023), https://thewire.in/law/why-did-twitter-agree-to-take-down-tweets-linking-to-bbc-documentary.

89. James Vincent, *Nigeria Lifts Twitter Ban, Says the Company Has Agreed to Government Demands*, VERGE (Jan. 13, 2022), https://www.theverge.com/2022/1/13/22881580/nigeria-twitter-ban-agreed-government-demands-local-office-tax.

90. Abubakar Idris & Peter Guest, *How Twitter Rolled Over to Get Unblocked in Nigeria*, REST OF WORLD (Jan. 13, 2022), https://restofworld.org/2022/how-twitter-rolled-over-to-get-unblocked-in-nigeria. Almost a year after the Twitter ban was lifted in Nigeria, however, Twitter told a news service that it had not yet opened an office in that country; Francis Onyemachi, *8 Months After Lifting Ban, Twitter Office Not in Nigeria*, BUS. DAY (Sept. 17, 2022), https://businessday.ng/technology/article/8-months-after-lifting-ban-twitter-office-not-in-nigeria/. In 2022, the Court for the Economic Community of West African States (ECOWAS) ruled that the Nigerian Twitter ban was an unlawful infringement on freedom of expression and access to media. Jason Kelley, *Nigerian Twitter Ban Declared Unlawful by Court*, ELEC. FRONTIER FOUND. (July 20, 2022), https://www.eff.org/deeplinks/2022/07/nigerian-twitter-ban-declared-unlawful-court-victory-eff-and-partners.

Nahel M.[91] Thierry Breton, the European Union's Commissioner for Internal Market, explained that the DSA would provide this power in the future. "When there is hateful content, content that calls—for example—for revolt, that also calls for killing and burning of cars, they will be required to delete [the content] immediately," Breton stated, citing the Digital Services Act obligations which were not yet enforceable in July 2023.[92] "If they don't act immediately, then yes, at that point we'll be able not only to impose a fine but also to ban the operation [of the platforms] on our territory."[93] Any such ban would only be temporary, however, as Florence G'sell points out: "Under Article 82 of the DSA, the European Commission can request a regulator to ask a judicial authority to temporarily restrict user access if there is a serious and persistent breach causing significant harm and involving a criminal offense threatening people's safety or lives."[94]

IV. CONCLUSION: AVOIDING DIGITAL CZARS

All of the powers that the Digital Services Act grants are worthy and well-intentioned, designed to respond to the critical role of internet platforms in our daily lives, from politics to business to culture. But by pointing out the risks entailed in the globalization of the Digital Services Act, this Article hopes to ensure we anticipate the ways that the powers granted by the law can be abused.

Many of these rules can be weaponized by activist politicians, who have an incentive to promote speech favorable to them, and to punish speech that criticizes them or regales an opponent. As we have seen, there are a variety of mechanisms in the Digital Services Act that are open to such exploitation. That may be true of all laws: they depend on the prosecutor, the regulator, the policeman, and the judge to administer them impartially. But when it comes to

91. Lyric Li, *Macron Says Social Media Could Be Blocked During Riots, Sparking Furor*, WASH. POST (July 6, 2023), https://www.washingtonpost.com/world/2023/07/06/france-macron-social-media-block-riots; Théophane Hartmann, *Macron Mulls Social Media Shutdowns to Contain Civil Disorder*, EURACTIV (July 5, 2023), https://www.euractiv.com/section/digital/news/macron-mulls-social-media-shutdowns-to-contain-civil-disorder; Laura Kayali & Elisa Bertholomey, *Macron Floats Social Media Cuts During Riots*, POLITICO (July 5, 2023), https://www.politico.eu/article/macron-mulls-cutting-access-social-media-during-riots/#:~:text=PARIS%20%E2%80%94%20French%20President%20Emmanuel%20Macron,his%20speech%20seen%20by%20POLITICO.

92. Clothilde Goujard & Nicolas Camut, *Social Media Riot Shutdowns Possible Under EU Content Law, Top Official Says*, POLITICO (July 10, 2023), https://www.politico.eu/article/social-media-riot-shutdowns-possible-under-eu-content-law-breton-says.

93. *Id.*

94. G'sell, *supra* note 5, at 20–21.

speech regulations, we should be especially concerned—because here the motivations to bend the law in one's favor are at their apex.

The DSA exists within a European legal framework that is interpreted by an independent judiciary. Evaluating its rules on its own does not recognize the constraints that may arise from other sources of law within the European Union. Any adoption of the DSA in foreign jurisdictions may not find a similar protective framework of laws and institutions. Legal transplants do not necessarily carry the old soil of institutions, practices, and people.

Governments across the world seeking to bring the internet under control have often turned to the European Union as a model. For example, one of the sponsors of the Brazilian "fake news" law, Federal Deputy Orlando Silva, explained that the German NetzDG law had served as an inspiration for the debates of the working group that created the Brazilian fake news law.[95] It is likely that the Digital Services Act will prove an attractive model as well for countries across the world.

At the same time, it is important to recognize that it is not only the countries of the former Soviet Union or the Global South that pose the risk of abuse. Even a President of the United States might install partisans in key spots to deploy agencies for political ends.[96] Even within the European Union, there are potential areas of risk. Many of the twenty-seven member states of the Union share a recent history of authoritarian leaders. And every leader, regardless of political party, has an incentive to control information flow—to designate the opposition's critique as defamation and lies.

This reflects what I have elsewhere described (with Haochen Sun) as the double-edged nature of digital sovereignty—digital sovereignty is both necessary and dangerous.[97] Legal regimes must anticipate the exploitation of

95. Tales Tomaz, *Brazilian Fake News Bill: Strong Content Moderation Accountability but Limited Hold on Platform Market Power*, 30 JAVNOST - THE PUBLIC, J. EUR. INST. FOR COMMC'N & CULTURE 253, 259 (2023).

96. Adam Goldman, *Justice Dept. Investigated Clinton Foundation Until Trump's Final Days*, N.Y. TIMES (May 22, 2023), https://www.nytimes.com/2023/05/22/us/politics/fbi-clinton-foundation.html; Maggie Jo Buchanan, *Trump's Politicization of the Justice System*, CTR. FOR AM. PROGRESS (Feb. 20, 2020), https://www.americanprogress.org/article/trumps-politicization-justice-system; Priscilla Alvarez & Geneva Sands, *Trump Uses Homeland Security Agency To Fight His Political Battle Against Democratic Cities*, CNN (July 20, 2020), https://www.cnn.com/2020/07/20/politics/trump-homeland-security-portland-chicago/index.html; Lisa Rein & Juliet Eilperin, *The White House Installs Political Aides At Cabinet Agencies To Be Trump's Eyes and Ears*, WASH. POST (Mar. 19, 2017), https://www.washingtonpost.com/powerpost/white-house-installs-political-aides-at-cabinet-agencies-to-be-trumps-eyes-and-ears/2017/03/19/68419f0e-08da-11e7-93dc-00f9bdd74ed1_story.html.

97. Anupam Chander & Haochen Sun, *Sovereignty 2.0*, 55 VAND. J. TRANSNAT'L L. 283 (2022).

powers over our digital activities not only by corporations but also by governments. We need law to protect us not only from profit-maximizing industrialists but also self-interested politicians.

Indeed, the DSA does seek to build in what Americans might call checks and balances in certain cases. The DSA includes, at various points, limits to the powers it grants as well as procedural protections. For example, the Digital Services Coordinator is supposed to be an independent body. Its investigatory power is to be exercised in conformity with the E.U. Charter of Rights and subject to safeguards in national law.[98] In this way, the DSA relies on national legislation to provide critical safeguards. The crisis protocol, too, must include safeguards to protect Charter rights.[99] Trusted flaggers are to meet various conditions.[100] The guidelines issued by the Commission for risk mitigation measures occur only after public consultations and must take "due regard to the possible consequences of the measures on fundamental rights enshrined in the Charter of all parties involved."[101] This is hardly an Act that is indifferent to the possibility that government officials might deploy those powers in abusive ways.

Even domestic checks and balances may prove inadequate. Many have advocated the use of international human rights law to guide and constrain the actions of private internet platforms.[102] But at present there is no international law mechanism to enforce speech and political rights within offending states.

---

98. Digital Services Act art. 51(6), 2022 O.J. (L 277) ("Member States shall lay down specific rules and procedures for the exercise of the powers pursuant to paragraphs 1, 2 and 3 and shall ensure that any exercise of those powers is subject to adequate safeguards laid down in the applicable national law in conformity with the Charter and with the general principles of Union law. In particular, those measures shall only be taken in accordance with the right to respect for private life and the rights of defense, including the rights to be heard and of access to the file, and subject to the right to an effective judicial remedy of all affected parties.").

99. *Id.* art. 48(4) ("The Commission shall aim to ensure that crisis protocols set out . . . safeguards to address any negative effects on the exercise of the fundamental rights enshrined in the Charter, in particular the freedom of expression and information and the right to non-discrimination.").

100. *Id.* art. 22. The recitals to the DSA recognize the possibility that trusted flaggers might themselves engage in abuse: "In order to avoid abuses of the trusted flagger status, it should be possible to suspend such status when a Digital Services Coordinator of establishment opened an investigation based on legitimate reasons." *Id.* recital 62. The solution to this concern seems to be to rely on the Digital Services Coordinator to investigate that abuse, but this solution only works if the Coordinator and flagger are not politically aligned.

101. *Id.* art. 35(3).

102. *See, e.g.*, David, *U.N. Doc. A/HCR/38/35: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UNITED NATIONS HUMAN RIGHTS OFFICE OF THE HIGH COMMISSIONER (Apr. 6, 2018). *But cf.* Evelyn Douek, *The Limits of International Law in Content Moderation*, 6 U.C. IRVINE J. INT'L, TRANSNAT'L & COMPAR. L. 37

When we think about fundamental rights, we must always keep in mind that we should protect fundamental rights both against private corporations and the state. Opportunities for abuse lie in both sources—indeed, the state may pose a special threat in some cases—targeting dissidents or political minorities. Alongside our understandable desire to bring internet companies under democratic control, we should remain mindful of the dangers of excessive government control as well.

(2021) (observing the limits of international human rights law to resolve difficult questions of content moderation).