Hertie School
Centre for
Digital Governance

# A User-Centered Approach To Content Moderation

22 June 2020

**Oscar Daniel Del Valle Salinas, Sarah Rerbal, Paxia Ksatryo, Evan Yoshimoto, Abhipsha Mahapatro, and Kindye Atnafu Adugna**

#HertieDigitalGovernance

The current EU E-Commerce Directive supports a "notice-and-takedown" approach to content moderation. Even though the approach has been effective for contents that are certainly illegal, divergences have emerged with regards to the "grey zone" of contents that fall between illegal hate speech and the user's right to expression. This is expressed by "overblocking" of contents by social media platforms (SMPs) in fear of being held liable. This policy paper assesses this gap in the content moderation approach in the eyes of the liability framework. The paper provides user-centered policy suggestions that promote chosen principles of freedom of speech, the right to participation, and due process as building blocks for the new legislation. These consist of a positive reward system for SMPs which includes an introduction of an EU-wide good Samaritan clause and accreditation mechanism; proper checks and balances by guaranteeing a user-inclusive appeal mechanism; and establishing an independent regulatory oversight body that comprises all relevant stakeholders.

Key words: Content moderation, Digital Services Act (DSA), E-commerce, EU, Intermediary liability, Social media platforms (SMPs)

Hertie School
Centre for
Digital Governance

# Our vision

**We aim to contribute to a respectful and reliable internet.** Our mission is to share information and knowledge about the development of the Digital Services Act and thus raise awareness of how it relates to addressing online hate speech and disinformation. Under the leadership of the German Presidency, the European Commission will make crucial decisions that shape the development of a European model for digital transformation. This emerging model is not only going to affect the lives of people inside the European Union, it will also serve as a role model for other countries in the World facing similar issues.

**We believe that more public discourse is needed** to better inform the decisions under deliberation by the European Commission. Toward this aim, we provide information about ongoing developments and the different policy approaches that Europe could take to tackle online hate speech and disinformation as part of the Digital Services Act. We will contribute to this debate during the public consultation phase of the European Commission on our website at www.digitalservicesact.eu.

**We are a group of graduate students** at the Hertie School in Berlin, Germany, who have teamed up with Daniela Stockmann, Professor of Digital Governance. The Hertie School's mission is to prepare students for leadership positions in government, business and civil society institutions, to produce knowledge for good governance and policymaking and to encourage responsible stewardship of the common good. We consider ourselves as politically independent and are open to political positions from the entire political spectrum in support of producing positive outcomes for society.

We are grateful for funding by the Hertie School's Center for Digital Governance and the Hertie School's Student and Teaching Activity Fund. The Hertie School is a private university based in Berlin, Germany, accredited by the state and the German Science Council. The school was founded at the end of 2003 as a project of the Hertie Foundation, which remains its major partner. We do not receive funding from Facebook, Google, Twitter or any other tech company.

# 1 Introduction

In January 2018, Twitter temporarily suspended the German satire magazine Titanic's account for posting a reaction to racist tweets composed by an Alternative for Germany (AfD) politician (Thomasson, 2018). Twitter acted on the side of caution to avoid heavy fines imposed under Germany's Network Enforcement Act (NetzDG), a legislation regulating online hate speech pursuant to the liability framework of the 2000 EU E-Commerce Directive (ECD). While NetzDG has been discussed as a test

balloon for the Digital Services Act (DSA), this example illustrates the risk posed onto users' freedom of expression for content in the grey zone of regulation, such as political speech or satire (DG Connect, 2019). The DSA will replace and expand on the ECD, presenting an opportunity to learn from the consequences on free expression posed by a punishment-based intermediary liability framework on social media platforms (SMPs).

## 2 Content Moderation as an Approach towards Platform Regulation

Article 14 of the ECD sets up a conditional liability framework for intermediaries, otherwise known as "safe harbors". While SMPs have not been explicitly mentioned in the ECD, the CJEU has co-opted SMPs into its liability framework (Netlog; C-360-10, 2012). Currently, providers who qualify as "intermediary service providers" are exempt from liability of illegal content shared by its users until two safe harbor provisions have been breached: **(1)** it has been detected or notified, and **(2)** if they "act expeditiously to remove or to disable the access to the information" (ECD Article 14, 2000). SMPs are notified ex ante and ex post the publishing of content through mechanisms such as upload filtering and flagging.

At present, the ECD employs a "notice-and-takedown" approach to content moderation decisions, mostly made by a combination of an automated content classification mechanism and human content moderators (Angelopoulos and Smet, 2017). Legislation limits the response of the intermediary to a 'one-size-fits-all' solution: a 'takedown' procedure that is applied horizontally across all forms of illegal content. While notice-and-takedown has proven to be a straightforward approach to tackle content that is clearly illegal, divergences emerge with the "grey zone" of content, where there lies a tension between illegal hate speech and a user's right to expression. Failure to take down unlawful content upon notification implicates SMP liability and financial consequences. In fear of contracting these consequences, platforms are provoked to engage in "overblocking" — the takedown of allegedly illegal material without proper scrutiny and redress (European Commission, 2019). As a result, SMPs have avoided liability at the expense of free expression of its users. The disincentives for hosting providers to take a proportional approach against such infringements in fear of losing safe harbor protection is described as the "Good Samaritan Paradox" (Eecke, 2011).

Acknowledging the value these platforms provide to users as a public forum, this policy paper aims to reorientate the discourse on intermediary liability with a user-centered approach by addressing the grey zone of content moderation. This user-

centered approach strives to respect freedom of expression, agency for redress, and responsibility for their contested content.

# 3 Framework for Evaluation

The objective of a user-centered approach is to increase the quality of service that SMPs provide by making platforms more effective, safer, and engaging for users (Mikael, 2013). In order to address the Good Samaritan Paradox of content moderation, we consider the following principles in the policy suggestions. These principles help to respect the cornerstone of democratic rights, protect the vulnerable, and safeguard transparency. Our rationales behind selecting these principles are: (1) safeguarding user-interests when facing the "Good Samaritan Paradox" (2) making solutions sustainable, and (3) bolstering the decision-making process in evaluating content.

First, the **right to freedom of expression** is fundamental to the formation of public opinion, without which democracy is not viable (ECHR Article 10, 1950). The principle shares two facets: the individual dimension, which guarantees the right to express opinions freely, and the collective dimension, which guarantees the free circulation of ideas.

Second, the legislation must provide for **unrestricted access to due process** (ECHR Article 13, 1950). Limitation of rights should only be the last resort after a process of legal certainty, ensuring an independent and impartial judgment.

Last, the legislation must uphold **inclusivity and proportionality** to determine the obligations of the intermediaries and to ensure that they have the appropriate tools to make decisions on "grey area" content. Inclusivity is important to understand and entertain different contexts. Proportionality should be considered to make sure the measures are not harsh and unjust. For this reason, precise and unambiguous definitions and **pluralistic participation** in the decision-making processes are essential.

# 4 Policy Suggestions

We recommend a complementary set of policy solutions to shift the attention back to users and encourage a proportional approach to content moderation. The first part focuses on how to provide positive incentives to redress overblocking by introducing a good samaritan clause and EU-wide criterias for content moderation. Corporate accountability is accompanied by checks and balances by the involvement of users and

public stakeholders. We propose a combination of a user-inclusive appeal mechanism, supplemented by a public oversight board with a clearly-specified, principles-based mandate to expand on the responsibility and actionability of content moderation decisions.

### 4.1 Positive Incentives for Social Media Providers

**A. Good Samaritan Clause:**

Switching from a passive to active role by reviewing whether content is legal, platforms lose their safe harbor and contract liability as a consequence (Madiega, 2020). This leads to overblocking, resulting in a decrease of freedom of speech, which will be addressed in the second part of our policy suggestions. SMP's content moderation process must mediate allegedly illegal content without the fear of liability more proactively. To counter the Good Samaritan Paradox, an introduction of an EU-wide 'Good Samaritan Clause' – that will protect SMPs from liability in case self-regulation methods are not efficient in completely restricting availability of harmful content – in the DSA is recommended (Madiega, 2020). This provides reassurance to SMPs from being held liable for hosting illegal content. It also serves as a push for online intermediaries to more diligently detect illegal material.

**B. Accreditation:**

Building from the code of principles laid out by the International Fact Checking Network (IFCN) – fairness, transparency, and a commitment to honest corrections – an EU-wide accreditation for SMPs certifying proportional content moderation practices is recommended (IFCN).

Accreditation allows for the consolidation of trust between users and online platforms, while ensuring healthy practices of content moderation by the latter (Pamment, 2020). It incentivizes SMPs to effectively fight illegal content while safeguarding user experience. The criteria for deciding on accreditation can include multiple parameters – ensuring an annual report from SMPs detailing how they have been proportionally combating illegal content, the timeliness and responsiveness to due process for users, and other proactive measures taken. Adding annual reports under the criteria to achieve accreditation increases effectiveness of the current reporting system and accountability from the SMPs. To ensure that effective due process is achieved, a flexible time frame of intervention to ensure responsible yet timely action should be adopted.

A standardized EU-wide accreditation mechanism directly links to the platform's brand image and consumer loyalty. Studies show that the components and quality of services impact how consumers view a product and interact with it (Seo et al., 2020). Asking SMPs to strive for certification ensures conformity to European standards

while also upholding the values of trustworthiness in a user-centered approach. A similar system for accreditation is already in place for medical devices under the Medical Devices Regulation, where the devices must have a CE (Conformité Européenne) mark that proves the device meets the requirements of the directive. (Deseva, 2014) As the European co-operation for Accreditation notes for Consumers and Citizens, "Through accreditation and harmonized application of standards, consumers can therefore have confidence in the products and services they purchase on the European market." (EA, 2020)

## 4.2 Checks and Balances

### A. A User-inclusive Appeal Mechanism

Addressing the "grey zone" calls for procedural justice in order to better protect and enforce user rights, build trust, and establish credibility in the decision-making process (Tyler and Hollander-Blumoff, 2011). Turning away from the status quo, policymakers should adopt what we call "multi-step notice-and-action".

Social platforms are judge and jury, they draw up their codes of conduct and execute them. Likewise, there is no concept of equality before the law, independence or impartiality (Bingham, 2011). It is essential to guarantee an exhaustive appeal mechanism for users in face of arbitrary decisions that affect them. 'Digital due process' can form the basis of an online infrastructure to restore fundamental freedoms and support the rule of law online.

Digital due process' can entail the following: **(1)** a fair and public review by an independent and impartial competent body within a reasonable time; **(2)** a proper prior notification; **(3)** an opportunity to respond and present evidence; **(4)** the right to legal representation; **(5)** the right to appeal; **(6)** users may seek access to competent courts; **(7)** the right to receive a clearly-articulated decision; and, **(8)** the right to an effective remedy including, for example, stay-up or takedown. Therefore, a key issue to combat the volume of content to moderate and the velocity to moderate it timely, ensuring integrity and freedom of expression, is to focus on uniform digital tools that promote certainty together with a digital due process (Mostert, 2020).

### B. Regulatory Oversight by an Independent Body

Public interest decisions which "should be taken by independent public authorities are now delegated to online platforms, making them *de-facto* regulators without adequate and necessary oversight" (DG Connect, 2019). Currently, regulators for various digital sectors such as data protection already exist, however "no EU-wide specific regulator of platforms are put into place to ensure effective oversight in the area of content moderation (DG Connect, 2019; Hoffmann and Gasparotti, 2020). The DSA aims to improve clarity through effective regulatory features, and effective

enforcement and legitimate oversight are a prerequisite. For this reason, we recommend **a public oversight body to ensure effective oversight** through democratic control, diverse stakeholder involvement, and political independence.

The oversight body shall consist of a set of relevant stakeholders including policy makers, academia, and civil society organizations to signal the objective of a user-centered content moderation approach. Together, these stakeholders shall ensure democratic control in decision-making procedures, bring a diverse set of expertise, and recenter users to the core of the debate.

*Regulatory Structure*

The first condition for an effective oversight body is its regulatory structure within the EU framework. Clarifications concerning the oversight board's capacities must take the shared competencies between EU institutions and its member states into account. The implementation of the oversight body in the area of shared competence in the EU will be difficult to achieve. However, a central authority which is given the power of the member states to contribute to EU wide harmonization in the policy field through overseeing the activities of SMPs would strengthen EU cooperation and trust among member states in an already supranational sphere.

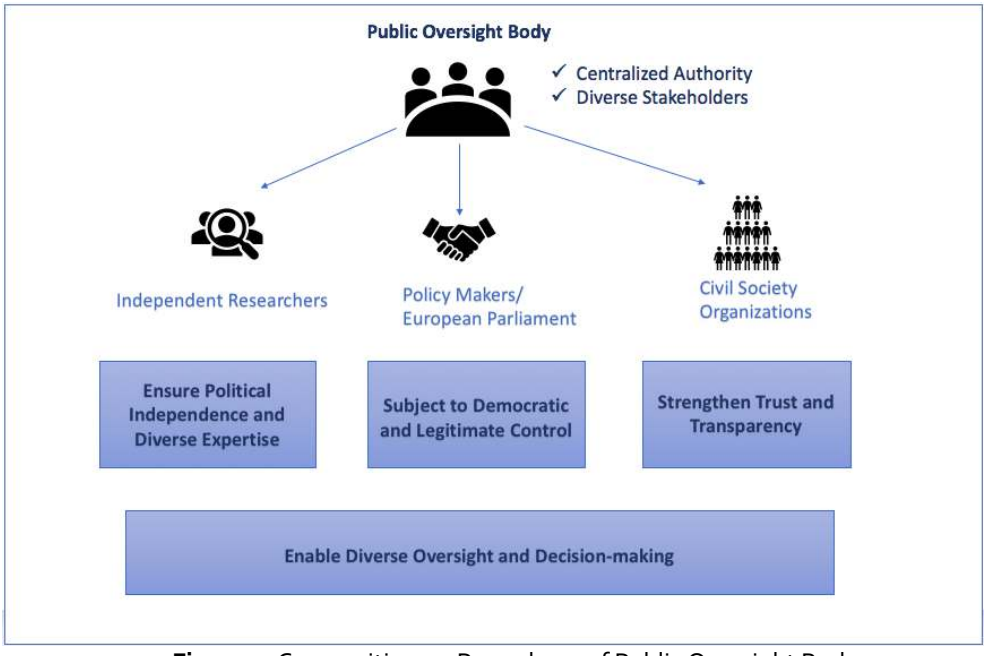


**Figure 1:** Composition on Procedures of Public Oversight Body

*Stakeholder Engagement*

(1) **Civil Society Organizations representing users:** Civil society organizations function as a representation of user perspectives and ensure the maintenance of equal participation and freedom of expression. Additionally, civil society engagement in the formal body will strengthen trust in formal decision-making and transparency by diverse channels of information on the body's activities.

**(2) Independent Researchers:** The involvement of independent researchers aims to ensure political independence and to guarantee diverse expertise in the complex decision-making process.

**(3) Policy Makers/European Parliament:** In order to verify that the oversight body is subject to legitimate democratic control, elected policy makers in the European Parliament are involved in the oversight body.

In order to ensure the effectiveness as well as the consistency with the rule of law, the public oversight body must abide by a **clearly specified mandate prescribing** (1) the substance of the oversight body and (2) procedural aspects of its work. **The mandate should specify** the involvement of stakeholders, objective, and tasks. Procedural aspects ensure regular meetings of the oversight body in a transparent manner through annual reports on the work of the body up for review by the European Parliament that will be available to the public.

## BIBLIOGRAPHY

Angelopoulos, C. and Smet, S., 2016. Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability (SSRN Scholarly Paper No. ID 2944917). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.2944917 (accessed 5.24.20).

Bingham, T., 2011. The Rule of Law. Penguin UK. UK. URL https://books.google.de/books/about/The_Rule_of_Law.html?id=6UsjcX-IUJ4C&redir_esc=y (accessed 4.23.20).

Devesa, F.R.S., 2014. Medical software certification processes in Europe, USA and Brazil. URL https://estudogeral.sib.uc.pt/handle/10316/26543 (accessed 5.31.20).

DG Connect, 2019. Digital Services Act Note. Leaked Document. Page 2. Online available at: https://cdn.netzpolitik.org/wp-upload/2019/07/Digital-Services-Act-note-DG-Connect-June-2019.pdf. (accessed 5.24.20).

ECD, 2000. 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'). 2000. eur-lex.europa.eu, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031.(accessed 5.24.20).

Eecke, P., 2011. Online service providers and liability: A plea for a balanced approach 48, 1455–1502. URL https://www.researchgate.net/publication/298435815_Online_service_providers_and_liability_A_plea_for_a_balanced_approach (accessed 5.24.20).

European Accreditation (EA), 2018. For Consumers & Citizens. European Accreditation. URL https://european-accreditation.org/accreditation/for-consumers-citizens/ (accessed 5.27.20).

European Commission, 2019. Code of Practice on Disinformation [WWW Document]. Shaping Europe's digital future - European Commission. URL https://ec.europa.eu/digital-single-market/en/news/code-practice- disinformation (accessed 4.23.20).

European Commission, 2019. Hosting intermediary services and illegal content online : an analysis of the scope of article 14 ECD in light of developments in the online service landscape : final report. [WWW Document]. URL http://op.europa.eu/en/publication-detail/-/publication/7779caca-2537-11e9-8d04-01aa75ed71a1/language-en (accessed 5.24.20).

European Commission, 2016. The EU Code of conduct on countering illegal hate speech online [WWW Document]. European Commission - European Commission. URL https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct- countering-illegal-hate-speech-online_en (accessed 4.23.20).

European Commission, 2000. e-Commerce Directive [WWW Document]. Shaping Europe's digital future - European Commission. URL https://ec.europa.eu/digital-single-market/en/e-commerce-directive (accessed 4.23.20).

European Convention on Human Rights (ECHR), 1950. Article 13, Article 10. URL: https://www.echr.coe.int/Pages/home.aspx?p=basictexts (accessed 5.24.20).

Hoffmann A. and Gasparotti A., 2020, Liability for illegal online content- Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a "Digital Services Act", published by cepStudy, page 35. URL https://www.cep.eu/fileadmin/user_upload/cep.eu/Studien/cepStudie_Haftung_fuer_illegale_Online-Inhalte/cepStudy_Liability_for_illegal_content_online.pdf (accessed 5.24.20).

IFCN, n.d. IFCN Code of Principles [WWW Document]. URL https://ifcncodeofprinciples.poynter.org/ (accessed 5.24.20). (accessed 5.24.20).

Madiega, T., 2020. Reform of the EU liability regime for online intermediaries: Background on the forthcoming digital services act - Think Tank [WWW Document]. URL https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_IDA(2020)649404 (accessed 5.24.20).

Mikael, J., 2013. How Social Media Changes User-Centered Design. URL: https://www.researchgate.net/publication/255708713_How_Social_Media_Changes_User-Centred_Design_Cumulative_and_Strategic_User_Involvement_with_Respect_to_Developer-User_Social_Distance. (accessed 04.24.20)

Mostert, F., 2020. "Digital Due Process": A Need for Online Justice [WWW Document]. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3537058 (accessed 5.24.20).

Netlog, C-360-10, 2012. JUDGEMENT OF THE COURT (Third Chamber) [WWW Document]. URL http://curia.europa.eu/juris/document/document.jsf?docid=119512&doclang=EN (accessed 5.24.20).

Pamment, J., 2020. EU Code of Practice on Disinformation: Briefing Note for the New European Commission [WWW Document]. Carnegie Endow. Int. Peace. URL https://carnegieendowment.org/2020/03/03/eu-code-of-practice-on-disinformation-briefing-note-for-new-european-commission-pub-81187 (accessed 5.24.20).

Seo, E., Park, J.-W., Choi, Y., 2020. The Effect of Social Media Usage Characteristics on e-WOM, Trust, and Brand Equity: Focusing on Users of Airline Social Media. Sustainability 12, 1691. https://doi.org/10.3390/su12041691 (accessed 5.24.20).

Thomasson, E., 2018. German hate speech law tested as Twitter blocks satire account. Reuters. URL https://www.reuters.com/article/us-germany-hatecrime/german-hate-speech-law-tested-as-twitter-blocks-satire-account-idUSKBN1ES1AT (accessed 4.23.20)

Tyler, T. and Hollander-Blumoff, R., 2011. Procedural Justice and the Rule of Law: Fostering Legitimacy in Alternative Dispute Resolution. Faculty Scholarship Series. URL https://digitalcommons.law.yale.edu/fss_papers/4992/ (accessed 5.24.20).



Friedrichstraße 194
D – 10117 Berlin
Tel.: +49 (0)30 467 26 09 – 01