UiO : Faculty of Law
University of Oslo

Shaping a new Strategy in Content Moderation

A comprehensive analysis of legal and technical challenges arising for a new strategy aimed at strengthening platform users' rights and reducing human content moderators' involvement.

Candidate number: 9013

Submission deadline: 01.12.2021

Number of words: 17258

Table of Content

Table of Figures

1 Introduction

1.1 Context and Research Questions

For the last two decades social media became essential part of our life. The globally facilitated internet access as well as an increasing participation in online platforms leads to a raise of users that daily up- and download not exclusively legal content. In fact, social media platforms offer individuals, inclined towards racism, misogyny or homophobia, an opportunity to find niches that can reinforce their views, spread their hate and goad them to violence. In their extreme, those rumours and invectives disseminated online, inspire acts of violence, such as mass attacks, lynching and ethnic cleaning.[1] Not at least the genocide in Myanmar sensitized especially social media platform Facebook to its responsibility for inflammatory posts: The platform was used by Burmese citizen to spread hate speech against the ethnic minority of Rohingyas also residing in Rakhine State, Myanmar. Due to a lack of training data and Burmese-speaking content moderators Facebook took down the harmful posts too slowly which contributed to the death and banishment of thousands of Rohingyas.[2]

Due to incidents like this, platform providers are set under great pressure to proactively take down unlawful content within a limited timeframe. Additionally, regulatory measures, such as the European Union's proposed Digital Service Act or national legislation like the German Network Enforcement Act (NetzDG) incentivize or in the latter case, urge them with high fines to optimize their strategies in the content filtering process mostly consisting of a combination of algorithmic and human review. In response to this, Facebook, as the biggest social media platform with more than 2.3 billion monthly active users[3] increased its proportion of algorithmic detection by six points from 89% to 95% in the second quarter of 2020[4] by employing approximately 15,000 human moderators to review online content manually.[5]

However, what seems a great achievement for content moderation still entails disadvantages for the balancing act of rights: Although, the algorithms' ability to be trained based on big datasets might constitute an analytical advantage over humans, in certain cases they still fail to recognize subjective meanings or the intend behind a text. Limitations arise in terms of both

[1] Laub, 'Hate Speech on Social Media.'

[2] Subedar, 'The country where Facebook posts whipped up hate.'

[3] Feldmann, ,How does Facebook moderate content.'

[4] Rosen, 'Community Standards Enforcement Report.'

[5] Thomas, 'Facebook content moderators paid to work from home.'

technologically insufficient solutions leading to inaccurate outcomes of the detection process and a lack of training data causing biases and misconceptions of non-wide-spread cultures and languages. Especially the resulting inaccuracy bears the risk of overblocking and infringements of users' freedom of expression.

This raises the first research question on how a strategy for algorithmic and human content moderation in the category of 'hate speech' should be shaped to balance platform users' fundamental rights more appropriately? What challenges arise from a technical and legal perspective?

With increasingly facilitated internet access, especially in the global south countries, even more people will join social media platforms and potentially spread hate and violence across the globe with consequences as described above. Solely regarding the quantity of data, we already rely on algorithms in the content filtering process, yet human oversight is still needed to fill the gap between the algorithmic and human ability to parse the nuanced meaning of communication. According to this, the increased need for additional moderators compensating algorithmic deficits by possessing both the language skills and knowledge of local events has recently been demonstrated by internal Facebook documents published by its former product manager and whistleblower Frances Haugen.[6] Confirmed by Facebook's spokesperson Jones 'adding more language expertise has been a key focus for [Facebook]'.[7]
However, what is still disregarded in this debate is the fact that human content moderators are no educated lawyers. Instead, they are contractually obliged by poorly paying third parties[8] to daily watch hateful or violative videos. With an inefficient offer of psychological assistance by their employers they continually run the risk of mental diseases.

Taken as the second object for a new strategy and research question for this study, how can algorithmic inaccuracy be compensated without hiring additional human moderators? Or at least, what technical and legal requirements are needed to protect them from health-damaging working conditions?

A new strategy must be found in which algorithms ensure a high level of fundamental rights, overtake the mentally stressful tasks, and resign human beings to perform control functions in the background.

---

[6] Srivas, 'What Facebook Whistleblower Compliant Touches Upon.'

[7] Culliford, 'Facebook knew about, failed to police, abusive content globally – documents.'

[8] Feldmann, 'How does Facebook moderate content.'

1.2 Purpose and Limitations

The purpose of this study is a comprehensive analysis of the arising challenges and the provision of concrete recommendations for platform providers and the EU legislator facing a new strategy in the hate speech removal process. By analysing the challenges from a legal and technical perspective, the overarching aim of the strategy is a stronger protection of platform users' rights as well as the relief of human content moderators.

To convince the stakeholders mentioned, the study's three main parts are chronologically ordered: Starting with the current shortcomings in content moderation, including the algorithms' limitations and degrading working conditions of human reviewers, Chapter 2 emphasises the urgency for a change. Based on this, Chapter 3 comprehensively analyses the stakeholders' interests and provides them with general recommendations to address the shortcomings mentioned. In the final stage, Chapter 4 presents practical solutions allocated to each stakeholders' interest.

Within the broad range of unlawful content to be detected and taken down from social media platforms, this research is limited to the category of 'hate speech.' Although there might be some national scope for interpretation within the EU, which will not be considered further, 'hate speech' lacks a consistent definition and is therefore critical to identify. Even a recent factsheet of the European Court of Human Rights (ECtHR) admits that '[t]here is no universally accepted definition of…hate speech'[9] and it 'can sometimes appear rational and normal.'[10] Taken for the understanding of this study, the Cambridge Dictionary defines hate speech as 'public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation.'[11] As hate speech consist of many forms of expressions that incite, promote or justify hatred, violence and discrimination against a person or group of persons,[12] the study's understanding of 'hate speech,' in addition to written comments or posts, also covers hate speech related pictures and video material incentivizing violence and discrimination. Due to its blurry definition, the categorization of 'hate speech' is not considered a simple yes-or-no decision in terms of illegality. Compared to other illegal content, such as e.g., 'terroristic content' or mere 'violative content', 'hate speech' opens a broad 'grey zone'

[9] Council of Europe, 'Factsheet - Hate Speech', 1.

[10] Council of Europe, 'Factsheet - Hate Speech' (2008), 2. See also ADF international 'Response to call for submission by the UN Special Rapporteur on the Protection of the Right to Freedom of Opinion and Expression', 3.

[11] Cambridge Dictionary, 'Hate Speech.'

[12] Council of Europe, 'Hate Speech and Violence.'

ranging from 'hate speech related' content to clear 'hate speech,' which makes it more critical to be identified.
Moreover, 'hate speech' contains the interface of two important fundamental rights: The right to non-discrimination protecting individuals from unequal treatment, offence, mental or even physical attack, and the freedom of expression enabling individuals and minorities a far-reaching audience within the digital space. Its critical identification as well as its representation of two important but likewise contradicting fundamental rights, are the reasons why the category of 'hate speech' is chosen for this study.

To be addressed by this research is the EU legislator as well as platform providers worldwide offering their services within the EU. The US legislator as well as Member States national legislators will be mentioned for the sake of completeness but not directly be addressed. Mostly but not exclusively it will be referred to Facebook as the most popular and technically advanced social media platform. Finally, to be determined as a 'human content moderator' is solely the individuum reviewing unlawful content, in distinct from human oversight in general.

1.3 Methodology and Material

At the crossroad between law and technology, this research requires both a comprehensive understanding of the applicable legal framework but also of the technical use of algorithms in the content moderation process. Especially, the latter is important to provide in this study as algorithms became essential to tackle the massive flow of data on social media but likewise raise challenges and concerns by replacing human judgement.[13] Therefore, the studies' first focus is on the technical analysis, preceded by a more general explanation of algorithmic content moderation, but mainly based on a model case made by the author. In order to simplify and illustrate algorithms' inaccuracy, the two-dimensional diagram represents the relation between algorithmic decision boundaries and their consequences for platform users' fundamental rights, which in turn emphasises the strong connection between law and technology.
The second focus of this study contains the legal analysis and will consider the relevant EU legislation that set the current or future regulatory framework for platform providers offering their services within the EU. Not to be considered as comparative, but rather as examples of more extreme regulatory approaches, the analysis briefly introduces parts of the US legislation as well as certain Member States' approaches. Whereas the former represents a more liberal

[13] Gonçalves, 'Common sense or censorshio',2.

approach in terms of platform regulation, some Member States such as Germany (NetzDG) extended the minimum requirements of the E-Commerce Directive[14] and implemented stricter regulatory measures within their national law. To ensure a direct connection to the technical analysis and to increase their comparability, each legislation provided will be allocated to the model case. Based on the analysis' outcome, the last part of this research provides some practical solutions that is built up on other researchers' expertise but will be adapted to the analysis' outcome.

1.4 Outline

The study is divided into three main sections. Prior to that, the introduction provides the context and main issues of current algorithmic content moderation leading to the two research questions on shaping a new strategy with stronger user protection and less human moderators' involvement. This part frames the study's purpose by determining its limits and key definitions.
Chapter 2 provides a general overview of the hate speech removal process including a technical explanation of algorithmic functioning, its limitations in detecting the nuanced meaning of a text, but also describes the involvement and working conditions of the human content moderators employed to compensate the algorithmic deficits.
Chapter 3 analyses the challenges arising when shaping a new strategy from a technical, but mainly from a legal perspective. By means of a model case two technical approaches that are applicable for platform providers will be introduced in the technical part, whereas in the legal part current and future legislation within the EU will be analysed and allocated to the model case. Additionally, recommendations for a new legislation regarding both a stronger protection of users' rights and moderators' working conditions will be provided.
Chapter 4 faces three practical solutions for platform providers and legislators to implement in the algorithmic content moderation as well as in the legal framework based on the recommendations acquired in the previous part. Finally, some practical thoughts and outlooks including moral concerns will close this chapter.
Lastly, Chapter 5 will summarize the critical points, remarks and recommendations arising from the analysis and amplified by the practical approaches.

[14] Directive 2000/31/EC.

2 What is Content Moderation?

As defined by Grimmelmann, content moderation is the 'governance mechanism that structures participation in a community to facilitate cooperation and prevent abuse.'[15] Especially for online communities, participation is already permitted through simple internet access, meaning that users can simply join from all over the world not limited by national borders.

However, what entails an advantage for its participation, concurrently provides a disadvantage for its regulation: As the services of platform providers are applicable across national boundaries, user-generated content is rather regulated based on the private companies' own policies than under governmental control. The fact, that these self-regulatory 'platform policies', 'terms of services' or 'terms and conditions' often lack any user-friendliness, transparency, and platforms' insight information[16] must be considered as an issue beyond the scope of this study.

Anyways, the dominance of platforms' self-regulatory regimes has its roots in the liberal U.S. jurisdiction where prominent platforms, such as Facebook, Twitter or YouTube have benefitted from Section 230 of the US Communications Decency Act, that grants them immunity from liabilities related to third party hosted content. However, according to hate speech related incidents increasing over the last decade, reforming legislation has been proposed and implemented at both EU level (Digital Services Act) and national level (e.g., German NetzDG). While (national) legislators increasingly demand a faster and stricter removement of hateful and offensive material, experts in contrast fear the risk of overblocking and an infringement of the freedom of expression.[17] Leaving room for discussion, the balancing act between the right to non-discrimination and the freedom of expression, can be considered as one of the most difficult decisions to make in the hate speech removal process.

2.1 Platforms' Hate Speech Removal Process – A Balancing Act of Rights

Facebook, taken as representative example for other big social media platforms, started with content moderation more than a decade ago. What at that time solely relied on users' reports double-checked by human reviewers, today is supported by sophisticated technology. Promoted by Facebook itself, those AI tools include '*proactive detecting*' of the company's Community

[15] Grimmelmann, 'The Virtues of Moderation', 47.

[16] See e.g., Brunk, 'Effect of Transparency and Trust on Acceptance of Automatic Online Content Moderation Systems.'

[17] See e.g., Schulz, 'Regulating Intermediaries to Protect Privacy Online.'

Standards independent from user reports, '*automation*' of decisions which are highly likely to be violating, and '*prioritization*' of content that is most harmful to users based on multiple factors such as virality, severity of harm and likelihood of violation.[18]
However, once content has been identified by either AI tools or users as potentially harmful, it is either being flagged or deleted. According to Gorwa et.al., in the former case content is placed in either in a regular queue, indistinguishable from a user-flagged content, or in a priority queue where it will be seen faster, or by human content moderators. In the latter case, content is removed outright or prevented from being uploaded in the first place. [19]

Although, it is the respective platform's policy setting the benchmark for the final decision in the filtering process, algorithms' as well as human moderators' major challenge remains the differentiation of ambiguous content that is either considered as discriminative or as covered by the freedom of speech. The high significance of both rights becomes apparent in the fact, that they are included in multiple national constitutions and international human rights treaties, such as the European Convention on Human Rights (ECHR) or the Universal Declaration of Human Rights (UDHR).

Article 14 of the ECHR claims that any discrimination shall be prohibited 'on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.'[20] Even applicable in the digital space, discrimination occurs either directly or indirectly when a person is treated disfavouring or when a person's dignity is violated. As potentially contradicting, Article 10 of the ECHR protects people's right to hold their own opinion and to receive and impart information and ideas without interference by public authority and regardless any frontiers.[21] Nevertheless, the freedom of expression might explicitly be restricted to protect other people's rights or reputation. Required is a certain level of 'proportionateness', meaning that it must be appropriate and no more necessary to address the issue concerned. In how far this should be considered in the content moderation process, that is constantly updated by platforms' new measures, will be part of the study's analysis.

[18] King, 'How we review Content.'
[19] Gorwa, 'Algorithmic Content Moderation', 6. See also Caplan, 'Content or Context Moderation?', 14.
[20] ECHR, Art. 14.
[21] ECHR, Art.10.

2.2 The Algorithm in the Hate Speech Removal Process

By applying the definition of 'content moderation' to online communities, the component of *algorithmic* is inevitably: As a system that classifies user-generated content based either on matching or prediction, leading to a decision and governance outcome (e.g. removal, geo-blocking, account takedown)'[22] *algorithmic content moderation* has been developed by platform providers as a machine learning system to cope with the mass of data up- and downloaded every day.

In a nutshell, machine learning as a brand of artificial intelligence (AI) is based on data and computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment.[23] Its sub-categories, the *neural networks* and *deep learning* approaches, focus on one single technique, the miming of human brain functionality and vary mainly in their depth of analysis and degree of automation.[24] This technique enables the machine in e.g., content moderation processes to ingest unstructured, unlabelled data in its row form (e.g. texts, images) and automatically determines the set of features which distinguish different categories of data from one another (e.g. hate speech from copyright infringements).[25] Without human intervention, which is still needed for conventional computers to succeed, deep learning approaches are able to learn based on examples and cannot be programmed to perform a specific task.[26]

Apart from the depth of analysis a further categorization can be made in the way the algorithm is trained: depending on the data provided and the result expected, the learning styles of algorithms are either unsupervised, supervised, self-supervised or reinforced. In the following the former approaches of unsupervised and supervised learning will be explained in more detail, exemplified by their most common algorithms used in content moderation, namely the *matching system*, the *classification system*, and the *neural language processing*

[22] Gorwa, 'Algorithmic Content Moderation', 3.
[23] El Naqa, 'What is machine learning?', 3.
[24] Mueller, 'Deep Learning for Dummies',9-24.
[25] IBM Cloud Education, 'Machine Learning.'
[26] Maind, 'Research Paper on Basic of Artificial Neural Networks', 96.

2.2.1 Unsupervised Learning

The Approach of Unsupervised Learning

In the unsupervised learning method, the algorithm analyses and clusters unlabelled data sets by discovering hidden patterns in the data without the need for human intervention (hence, they are 'unsupervised'). Provided with inputs, but not with the desired outputs, the system itself must decide what features it will use to group the input data. Although, those models provide insights from large volumes of new data, they still require human adjustments to validate the output variables and to counteract their widely inaccuracy.

The Matching System

A representative example of unsupervised learning is a *matching system* that makes knew data categorizable by underlying a uniquely identifiable string of data, the so-called 'hash'. [27] Although these 'digital fingerprints' are easy to compute and compare, their uniqueness makes them resistant to collisions (when two different pieces of content share the same hash) and prevents the data from unauthorised modification.[28] Nevertheless, the matching system reaches its limits by identifying other than previous known keywords, e.g., when parsing nuanced meaning of context[29] or detecting minor modifications (e.g. changing the colour of one pixel in an image).[30] To make these techniques more robust for changes, alternative techniques, such as 'fuzzy hashing' or 'perceptual hashing' aim to blur the similarities between two inputs, by rather matching 'homologies' than exact matches.[31] Exemplarily, 'perceptual hashing' focuses on the identification of prominent characteristics, such as corners of images, to become more robust to changes that are irrelevant to how humans perceive the content.[32]

2.2.2 Supervised Learning

The Approach of Supervised Learning

As the name supervised learning already indicates, the algorithm is trained under human supervision and based on labelled data sets. With the goal to predict outcomes for new data, the

[27] Gorwa, 'Algorithmic content moderation', 4.

[28] Ibid.

[29] Duarte, 'Mixed Messages?', 3.

[30] Gorwa, 'Algorithmic content moderation',4.

[31] Datar, 'Locality-sensitive hashing scheme based on p-stable distributions', 255.

[32] See Gorwa, 'Algorithmic content moderation',4, and Niu, 'An overview of perceptual hashing', 426-427.

type of result can be expected from its classification and tends to be more accurate than unsupervised learning.

More precisely, the algorithm is provided with certain input and already determined output data used to create a model that it fits to. By comparing the known examples and the model estimate, its discrepancies are reported back through the system and cause its adjustments. This progress is repeated until the updated weights autonomously reach the desired thresh-old of accuracy.

Concerning the targets to reach, supervised learning can be further subdivided into regression and classification. While the former's aims to provide a numeric value, such as an average price for a specific good, the latter's target is a qualitative variable, such as a class or tag. Especially, the classification system is predominately used for content moderation, when e.g., categorizing images or text passages as hate speech or discriminating and is therefore the representative algorithm described in more detail:

The Classification System

In contrast to the matching system that categorizes pieces of content against an existing data base, the classification system assesses newly uploaded content that has no previous version.[33] Based on examples labelled by humans as either belonging or not to a targeted category of content (e.g., hate speech or not hate speech), classifying algorithms identify patterns and learn rules of sorting new, unlabelled examples of the targeted content.[34] As classification systems became more sophisticated, a discipline of computer science emerged, called natural language processing (NLP), using neural networks[35] to parse texts and features to classify them.[36] This technique enables to cover the position of a word in relation to all other words that usually appear around it (word embeddings).[37] Those technique already today contribute to the success of hate speech detecting tools in the content moderation process.

2.3 What are today's Limitations of Algorithms?

Machine learning approaches significantly relieve human beings in the content moderation process by filtering a quantity of data that is daily up- and downloaded on social media platforms. However, an algorithm is trained on big data sets and makes decisions on a yes-or-

[33] Gorwa, 'Algorithmic content moderation',4.

[34] Duarte, 'Mixed Messages?', 10.

[35] More information on neural networks in Tanz, 'Neural networks made easy'.

[36] Duarte, 'Mixed Messages?', 10.

[37] Gorwa, 'Algorithmic content moderation',4.

no basis. What seems an analytical advantage over humans, still fails to recognize the nuanced meaning or intent behind texts. Limitations arise in terms of both technologically insufficient solutions leading to inaccurate outcomes of the detection process and a lack of training data causing biases and misconceptions of non-wide-spread cultures and languages.

2.3.1 Accuracy Rate and False Positive Rate

To scale the efficiency of content filtering algorithms, either as classification or as matching systems, the benchmarks of accuracy rates as well as of false positive rates are considered as most expressive: Regarding the former, an accuracy rate constitutes the percentage of correct predictions for a given dataset. This means, the closer the algorithm comes to coincide with the human coder's result, the higher is its accuracy rate. For example, when an algorithm's accuracy rate is at 80%, the machine filtered 80 out of 100 cases the same way as human reviewers would do. Consequently, a high rate only reflects the majority of human decisions, so potentially may bias the training data towards the majoritarian view of what is 'hateful' and might ignore a wholly legitimate expression of minority voices. In addition to that, one should be aware that even with an accuracy rate of 80%, one person out of five is treated incorrectly, what potentially could affect any individual's civil liberties and human rights.[38]
Moreover, domain-specific hate speech detection tools still drop in accuracy when applied into the diverse, dynamic speech environment of a social media platform. Confirmed by Abassi et al., tools that achieve a high accuracy in one context may suffer when exposed to one other context or way of speaking.[39]

Beside the demand for high accuracy rates, another important benchmark constitutes the false positive rate, defined as the incorrect identification of anomalous data, e.g., classifying as 'unlawful data' which is in fact lawful. This might have significant consequences in practice, when algorithms too often filter benign speech like jokes, sarcasm and literary devices[40] and risk an infringement of the user's freedom of expression.
Fact is, that automated decision-making tools still lack accuracy and remain far behind the reality. Therefore, human validation is still required to avoid inacceptable outcomes for platform users' rights.

[38] Duarte, 'Mixed Messages?',17.
[39] Abbasi, 'Benchmarking Twitter Sentiment', 6.
[40] Duarte, 'Mixed Messages?', 18.

2.3.2 Language and Cultural Biases

According to Hirschberger et.al, '*a major limitation of NLP today is the fact that most NLP resources and systems are available only for high-resource languages*', such as English, Spanish or Chinese, whereas '*low-resource languages*', such as Indonesian, Punjabi or Swahili, '*…spoken and written by millions of people have no such resource or systems available'*.[41] Due to their lack of training data, NLP tools are challenged by variations in dialect and language usage across demographic and cultural groups of English speakers. A prominent example contains Instagram's DeepText automated hate speech filter that in internal tests incorrectly identified the following sentence as hate speech: '*I didn't buy any alcohol, this weekend, and only bought 20 fags. Proud that I still have 49 quid tbh*.' [42] The word '*fags*' was evidently identified as a slur, as in American English it is a derogatory term for gay men, whereas in this context it has clearly been used to refer to cigarettes in colloquial British English.[43] To avoid these failures in NLP tools, researchers must manually correct the biases and warn for content moderation decisions that disproportionally censor certain minorities.[44]

As slang is just one example of the contextual difficulties inherent in NLP, it is likewise the machine's (dis)ability to recognize cultural backgrounds. What is defined as hate speech in one country does not have to be perceived as such in another. Furthermore, cultural norms and the understanding of certain statements may develop over time which poses significant challenges for the algorithm. For instance, Russians and Ukrainians for a long time have been calling each other the slang word 'moskal' or 'khokhol'. After a conflict started in 2014, these slang words started to be used as hate speech. In a similar case in Myanmar, the Burmese word 'kalar' was historically kind and friendly. The term could however be used as a provocative slur and was used as a term to promote attack by Buddhist nationalist against Muslims.[45] In both cases, Facebook did not detect those words as 'hate speech' until it was reported by users from the respective countries. The company still struggles with enforcing new policies to remove those expressions in a threatening context. This phenomenon has also been observed in Europe, when in Germany the influx of migrants arrived in 2015 and Facebook after receiving feedback by users developed new guidelines to remove calls for violence against migrants or dehumanizing

[41] Hirschberg, 'Advances in Natural Language Processing', 349.

[42] Thompson, 'Instagram Unleashes an AI System to Blast Away Nasty Comments.'

[43] Duarte, 'Mixed Messages?', 15.

[44] Bolukbasi, 'Man is to Computer Programmer as Woman to Homemaker?', 5.

[45] Allan, 'Hard Questions.'

references to them. Thereby they still left in place the ability for people to express their views on immigration itself.[46]

However, hate speech detection significantly depends on cultural norms and personal sensibilities and should not be considered a binary yes or no question.[47] A further obstacle to the algorithm is the obscurity of a uniform definition of hate speech. Although, the filtering process is individual to each platform provider and based on its specific community standards, a clear and consistent definition could improve the algorithm's application independent from its location.

2.4 The Human in Content Moderation

Despite the technical progress, human oversight remains essential to fill the gap between algorithmic and human ability to parse the nuanced meaning of communication. Nevertheless, increasing concerns should arise when considering the individual's working conditions:

Regardless of the platform, estimated 100,000 human content moderators ('human moderators', 'moderators' or 'reviewers') are staffed globally to assess user-generated content for their compliance with social media's terms of services and community guidelines.[48] Solely Facebook boosted its human content moderators from a total amount of twelve in the year 2009 to 15,000 in the year 2018 urged by an enormous growth of 120 million to 2.3 billion monthly active users.[49]

Most of that labour is not operated by the platform's own employees, but rather by spanning internal reviewers, contract workers from third parties or by outsourcing to online labour.[50] The main difference lies in the employees' payment: While those moderators that are contractually obliged by a third party, earn from $1,404 per year as in India and Bangladesh to $28,800 per year as in the U.S., Facebook's in-house employees receive a regular salary at an average of $240,000 a year.[51]

[46] Allan, 'Hard Questions.'

[47] Ross. 'Measuring the Reliability of Hate Speech Annotations', 1.

[48] Steiger, 'The Psychological Well-Being of Content Moderators', 1.

[49] Feldmann, ,How does Facebook moderate content.'

[50] Roberts, 'Commercial Content Moderation: Digital Laborers' Dirty Work, 2.

[51] Feldmann, 'How does Facebook moderate content.'

Regarding the moderators' job in more detail, non-compliant posts range from copyright infringement to disinformation and obscenity laws, also including depictions or actual acts of gore or lethal violence,[52] sexual abuse, child or revenge pornography[53] and more.
Based on that several studies undeniably prove that repeated, prolonged exposure to specific content, coupled with limited workplace support, can significantly impair the psychological well-being of human moderators.[54] This in its worst case might lead to a form of posttraumatic stress disorders (PTSD), also known as vicarious trauma. In this context, three former content moderators at Facebook sued the company in a Californian superior court for failing to create a safe work environment. In May 2020 the lawsuit has been settled by an agreement covering a financial compensation of $1,000 for each of 11,500 claimants, if PTSD is diagnosed.[55]

Beside the few that hazard to sue a company as Facebook, moderators are often limited to publicly speak about their working conditions. In its recent open letter to the Irish Parliament, published the 22nd of July 2021, more than 100 content moderators made three demands to Facebook: First, the company 'must end its culture of fear and secrecy', meaning that no non-disclosure-agreements (NDAs) or training sessions the moderators were obliged to sign or attend, are restricted to 'user data' and 'personal information' and that they can freely criticize their working conditions. Second, the letter claims that the mental health support Facebook provides to its moderators is 'woefully inadequate'. Proper psychological assistance was urgently needed to mentally process the harmful content. As a third demand, content moderators want Facebook to stop second-class citizenship by being outsourced to third parties. They rather want to 'be brought in house' and 'receive the same pay, benefits and employment conditions' as regular employees at Facebook.[56]
Regarding the psychological burden on the individual content moderator, not only an improvement of working conditions is urgently needed, but also an enhancement of machine learning approaches becomes essential to relieve human beings.

[52] Deniz, 'Fast violence detection in video', 478.
[53] Sae-Bae, 'Towards automated detection in child pornography', 5332.
[54] Halevy, 'Preserving Integrity in Online Social Networks', 25.
[55] Newton, 'Facebook will pay $52 million in settlement.'
[56] Banerjee, 'Facebook's content moderators demand for an end to culture of 'fear and secrecy'.

3 What are the Challenges of Shaping a New Strategy in Content Moderation?

With a special focus on strengthening platform users' rights but also on reducing human's involvement, a new strategy must face diverse challenges such as technical shortcomings in platforms' technology or legal requirements applicable within the EU.

As the first challenge it will be considered the impact of algorithms' inaccuracies on the users' fundamental rights, namely the right to non-discrimination and the freedom of expression. More precisely, in a model case it will be analysed what practical consequences it might have for the user and for the human moderators' involvement when lowering or extending the filtering scope of a content moderating algorithm.

Taken as the second challenge, a collaboration between human and algorithmic content moderation must meet several regulatory requirements regarding content liability, implemented by EU or national legislators and elaborated by the cases of the European Court of Justice (CJEU). Moreover, recommendations will be provided on how a legislation should be framed to hold platform providers appropriately accountable and to ensure a certain legal protection for human content moderators.

3.1 Technical Challenges: A Model Case on Algorithms' Inaccuracies Impacting Platform Users' Rights

Provided that even in the next decade content moderation's algorithms will not become sophisticated enough to entirely detect and filter any hate speech related post appearing on a platform, it will be investigated how algorithms should be programmed to protect the platform users' rights and concurrently reduce human involvement.

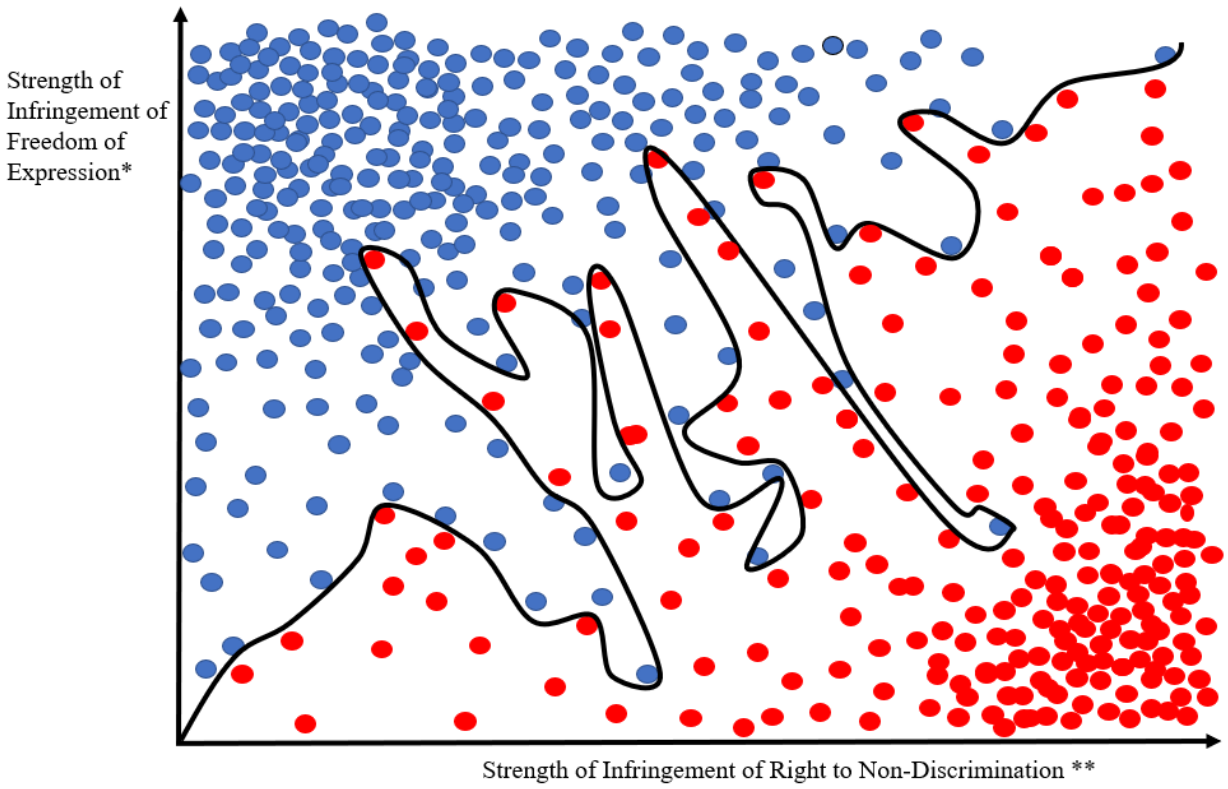
Starting point for the following model case will be an illustration of the optimal congruency of algorithmic and human classification wherein the algorithm drew a decision boundary based on its already known training data. In a second illustration, the algorithm is challenged by previously unseen data, which consequently leads to inaccuracy in its detection rate. Based on this problem, two algorithmic approaches programmable by a human coder can shift the decisions boundary to either strengthen the right to non-discrimination or the freedom of expression. What impact one or the other solution might have on the user's fundamental rights, as well as on the role of human reviewers will be the focus of this model case.

3.1.1 A Simplified Two-Dimension Model with Optimal Accuracy

Starting point to the model case on algorithms' inaccuracies impacting the platform users' rights, is the illustration of a two-dimensional diagram with an optimal congruency of

algorithmic and human classification. As one of many algorithms used for content moderation and diverging in scale, criteria and training data but not related to Facebook or any other company, the illustration reflects an algorithm that is trained based on a given data set and classifying exclusively known content.

Figure 1. Optimal Congruency of Algorithmic and Human Classification.[57]



Before describing the illustration in more detail, one must notice that for simplicity reasons and without any empirical proof the algorithm in this diagram measures and assesses content based on only two criteria assigned to its x-and y-axis. In its real-world application an algorithmic decision contains thousands of different criteria and dimensions.

[57] Illustration by the author.

Specific to this diagram the algorithm measures content based on its strength of infringement of the user's right to non-discrimination (x-axis) and of the user's freedom of expression (y-axis) by considering the following questions:

- *How strong does the content infringe the right to non-discrimination, if it remains visible on the platform? (x-axis)*
- *How strong does the content infringe the freedom of expression, if the content is erased from the platform? (y-axis)*

Thereby, it is important to know, that the algorithm measures each criterion independently and allocates each content to a score. Once content is scored in every criterion, the algorithm places its representing dots (no matter the colour) into its right position in the diagram.

In a next step the algorithm assesses the scored dots on whether they are categorized as 'hate speech' or not. To illustrate its evaluation, the algorithm draws a decision boundary, represented as the black line.

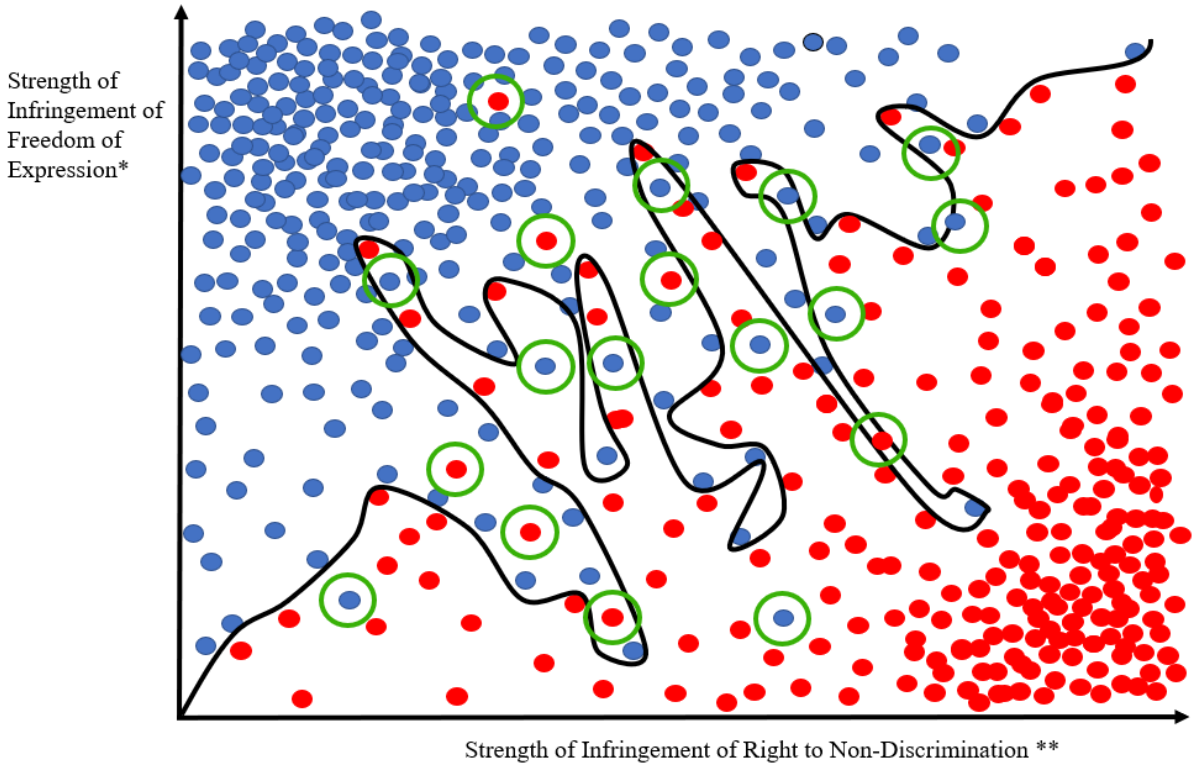
Regarding the colour of dots, they are only related to the decisions made by human beings. While the blue dots represent content that human reviewers assessed as harmless, the red ones were assessed as hate speech.

Since the algorithm has assessed previously known data that reflects the human evaluation of content, one can observe an optimal congruency of algorithmic and human decision.

3.1.2 Evaluation of New Content causes Inaccuracy

However, any optimal congruency between algorithmic and human decision making comes at its cost when new and previously unseen content is uploaded to the platform. The second diagram of the model case illustrates the appearance of new content and how the algorithm struggles to categorize it correctly.

Figure 2. The Problem: New Content on the Platform leading to Inaccuracy in the Algorithm's Detection Rate.[58]



The diagram compared to the above has not changed in its measurement. It is still based on the same criteria the algorithm scored and placed the dots into the diagram and separated them by its decision boundary into hateful and harmless content. Even the human evaluation has not changed. The only difference is the appearance of new, by the algorithm previously unseen content.

[58] Illustration by the author.

In this regard a divergence between human and algorithmic review can be observed and is marked with green circles. Whereas for example, a human content moderator assesses a new comment or post as sarcasm, and therefore keeps it visible on the platform, an algorithm might misconceive the sarcastic context and consequently classify the post as hate speech. This also applies to the reverse case when the algorithm fails to recognize hateful content.

Even though, this diagram is just an illustration, it becomes clear by what failures platform providers are challenged when programming algorithms for the real-world application. Although, many companies such as Facebook constantly improve their AI tools and boast its increasing numbers of proactively detections of hate speech to the public,[59] it remains questionable if those absolute numbers can be traced back to an improvement of technology or if it is rather the more rapid upload and spread of content leading to the raise of detected cases. Even in percentual numbers, when Facebook in early 2021 reported that more than 97% of content has been categorized as hate speech,[60] the denominator of the equitation is what Facebook's AI took down and not the total amount of harmful content available on the platform. Although, platforms self-reported numbers, either absolute or percentual, might insinuate a progress, an algorithm solving the problem has not found. Why else should Facebook still employ about 15,000 human content moderators?

Taken this for granted, a solution to bridge the current algorithmic limitations must be found in a most considerate way for the users' fundamental rights and for the role of human moderators. In a next step two possible approaches for algorithms to be programmed will be analysed in more detail.

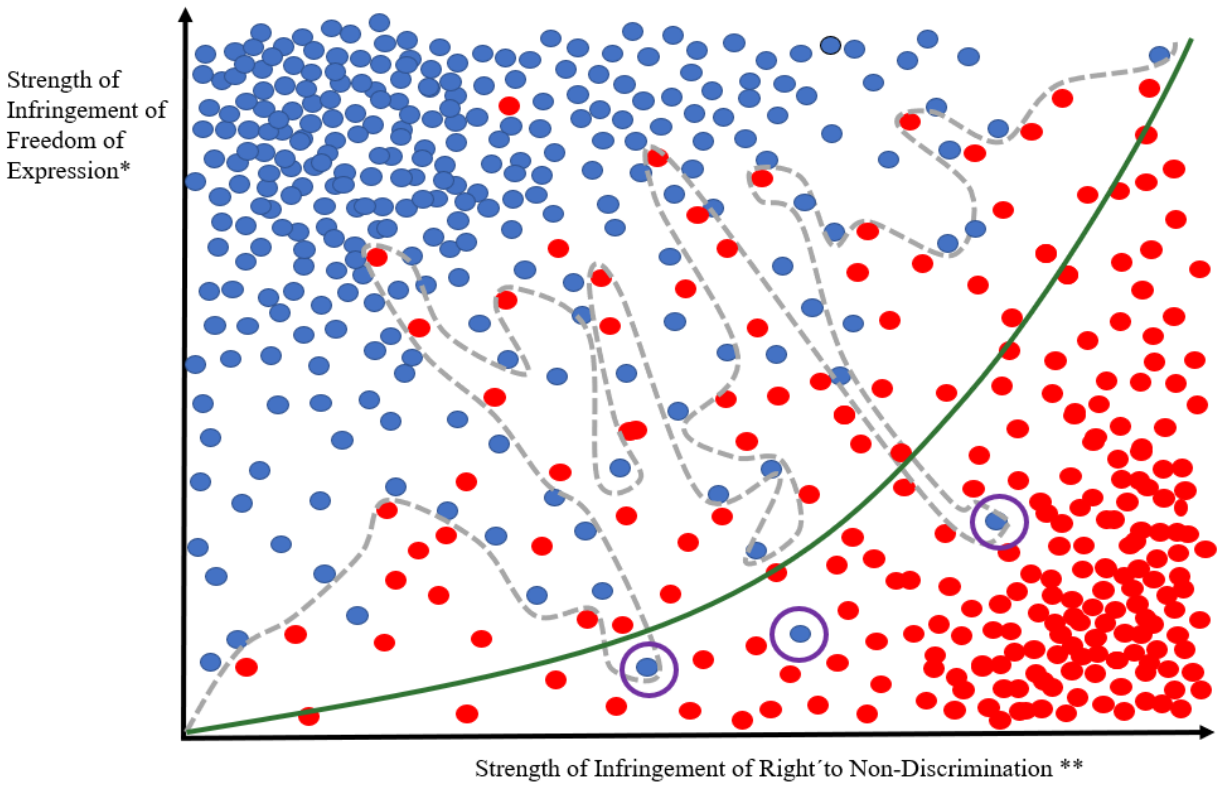
3.1.3 Two Approaches of Hate Speech Detecting Algorithms

In the first approach, the algorithm is programmed to filter hate speech within a narrower scope compared to the original algorithm by shifting its decision boundary towards the x-axis.

[59] Wiggers. 'Facebook's improved AI isn't preventing harmful content from spreading.'

[60] Sonderby, 'Our Continuing Commitment to Transparency'.

Figure 3. Approach 1: Narrow Scope of Hate Speech Detecting Algorithm.[61]



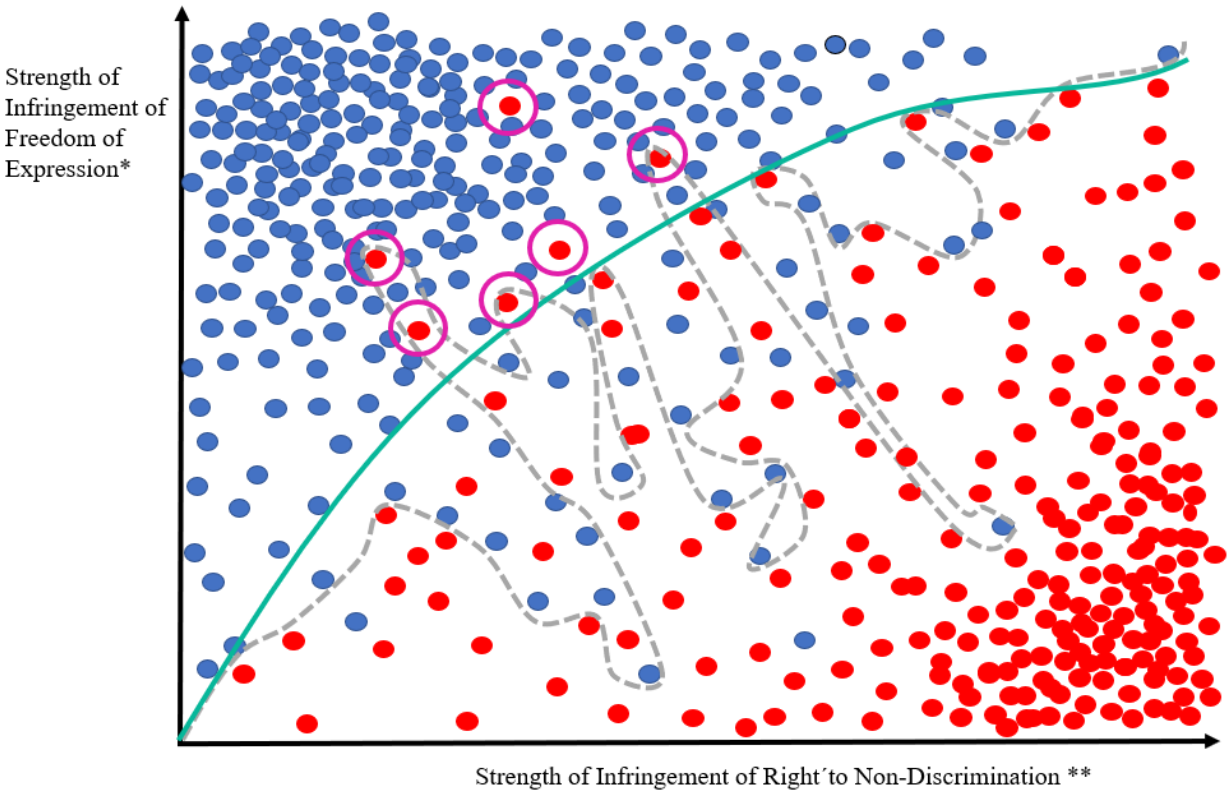
What is kept in this illustration compared to the original illustration is that the algorithm measures content both known and unknown in the same correct and incorrect way as in the previous diagrams. It independently allocates any content a score in each criterion and places it between the axes. What is new, is that its decision boundary has been shifted towards the x-axis alias the criterion of the strength of infringement of the right to non-discrimination. This means the algorithm has changed in its assessment of content by lowering the level of what is categorized as hate speech.

[61] Illustration by the author.

Irrespective of the technical implementation in detail, the consequences of this approach are visible: While the amount of harmless content that incorrectly has been taken down by the algorithm significantly dropped due to the lower decision boundary (blue dots marked in pink circles) in turn, the amount of harmful content remaining visible on the platform has raised by the same level (all red dots above the green boundary).

In a second approach, the algorithm is programmed to detect hate speech within a wider scope compared to the original algorithm by shifting its decision boundary towards the y-axis.

Figure 4. Approach 2: Wide Scope of Hate Speech Detecting Algorithm.[62]



[62] Illustration by the author.

In contrast to the first approach, in the second the algorithm's decision boundary has been shifted towards the y-axis alias the criterion of the strength of infringement of the freedom of expression. As same as in the above illustrated diagram, the algorithm again did not change in its measurement of allocated scores, but rather in its assessment of content by now raising the level of what is categorized as hate speech.

The consequences reflect the opposite of the above: The higher boundary significantly raises the level of what is filtered from the platform as potentially harmful, but likewise covers various content by a human being considered as harmless.

3.1.4 Consequences for the User's Fundamental Rights

Comparing the two approaches it becomes obvious that by extending the algorithm's decision boundary towards one right, the other right will be strengthened. For example, when in the first approach the algorithm's boundary is graphically extended towards the strengths of the infringement of the right to non-discrimination, less content will be taken down by the algorithm and therefore less takedowns can infringe the users' freedom of expression. In turn, when in the second approach the boundary is extended towards the strengths of infringement of the freedom of expression, more content potentially including hate speech will be taken from the platform and less people will be infringed in their right to not be discriminated.
However, what approach to choose?

When it comes to the implementation of a regulatory measure potentially causing a conflict of interests/rights, one might refer to the 'proportionality principle'[63], which at EU level governs all actions by requiring a legitimate purpose, necessity, and suitability as well as proportionateness in a narrower sense. Although both measures at hand are necessary and suitable to achieve the legitimate purpose of strengthening a fundamental right, those purposes are contradicting and require a balance of interests when it comes to their proportionateness in a narrower sense. Before doing so, one might notice that no fundamental right can be rated generally higher than another. Its evaluation highly depends on various circumstances of the individual case.
Nevertheless, by balancing the right to non-discrimination with the freedom of speech in algorithmic content moderation, one might determine at least a few objective criteria to

[63] European Commission, 'Better regulation toolbox, #18'.

compare, such as the number of people affected as well as the intensity of infringements. Even in this regard uncertainties remain due to a lack of statistics on either how much and how fast (hateful) content is shared and spread on platforms or on how many users' content is misconceived as hate speech.

Since news and media tend to bias the picture by reporting rather on platform provider's failure to detect hate speech than on cases of overblocking and censorship, [64] most people involved would probably assess the right to non-discrimination as the right to pay more attention to. But even the freedom of speech is a fundamental right that deserves intensive protection. The digital space offers especially individuals and minorities a broad audience to express their opinions, critics, or thoughts, to share information or even to agree and disagree to those in power or express their opinions in peaceful protests.[65]

More precisely, a user who creates or shares a hateful post might direct it to an individual person, to a group of people or to a whole society by causing consequences reaching from individual discriminations to appeals for violence. The sharing option of many platforms contributed to intensify this spread and enormously widens the audience for harmful content. In addition, even the evocation of further hateful comments and reactions makes hate speech spread. According to Astuti, it is especially the younger users that feel provoked by those comments, cannot just neglect it. Carried away by their emotions they react with their own hateful comments and goad each other.[66]

To prevent any content from being spread, either by sharing or reacting, the most efficient way is to avoid its origin of being uploaded. In this regard and by referring to the initial question of which algorithmic approach to choose, the second one performs best when offering an extended scope of content being filtered. A further argument for an early and generous takedown might constitute the potential reduction of irreversible mental harm for especially younger or mentally weaker users. The victims of hate speech are often psychological impacted by negative emotions, such as feeling angry, uncomfortable, sad, depressed, embarrassed, afraid, insecure, and hurt[67] and often require external assistance to combat them.

Also, in favour for the second approach is that harmless content must not be prevented finally from being uploaded. In a second evaluation another algorithm or human being should have a

[64] See e.g., Baggs, 'Online hate speech rose 20% during pandamic'.

[65] Amnesty International, ‚Right to Freedom of Expression'.

[66] Astuti, 'The Hate Speech Behavior of Teenagers on Social Media Instagram'. 257-258.

[67] Ibid. 258.

second review for ambiguous cases and deciding whether it should be uploaded belated. This constitutes the advantage that authors of these post would just be limited in their freedom of expression for a certain time.

Considering the above mentioned, it is the second approach that convinced to choose. An algorithm filtering content within a wider scope consequently covers more hate speech and reduces its spread and the users' risk of mental harm. By additionally implementing the option to re-upload incorrectly filtered content the freedom of speech can be restored.

3.1.5 Consequences for the Human Content Moderators' Involvement

Regarding the consequences for the human moderators' involvement a closer look shall be taken into the specific content the moderators predominantly review in each approach: When less content is filtered by an algorithm (approach 1), factual more hate speech remains visible on the platform. If now the algorithm does not flag this content, more users will do. In almost any platform users have the option to report content that potentially infringes its rights. Regarding Facebook those flagged posts are directly addressed to the human moderators to review. Even if it does not contain hate speech, it logically reflects what the average user assesses as hate speech. This might increase the pressure on human reviewers, its quantity of reviews as well as in its intensity of hate.[68]

In contrast to this, when an algorithm filters more potentially hateful content, the moderator would have less material to review. Although, it can be argued that the number of incidents reported by users feeling infringed in their freedom of expression might raise, the fact that the user must become proactive and likewise be convinced by its infringement, indicates to become less material to be reviewed in total. This indication applies also to the intensity of hate, that might be reduced for cases where the user proactively argues for its freedom of speech.

Although this model theoretically seems to relieve the human moderator, in a real-world application platform provider must guarantee a fast review of reported incidents to keep any infringements of the freedom of speech as short as possible. Therefore, one might consider requiring the flagging user to precisely categorize and describe its content before it will be visible for the human reviewer.

Consequently, the second approach of this model case contains even for the human moderators the more considerate way of reviewing potential harmful content.

[68] Caplan, 'Content or Context Moderation?', 14.

3.1.6 Conclusion Model Case

Resuming the model case illustrated, content moderation algorithms in the nearer future will continue to identify newly uploaded content incorrectly. To protect the users' fundamental rights, two approaches have been demonstrated each tending to strengthen one or the other fundamental right. Under consideration of the respective advantages and disadvantages of each approach, the second convinced both to protect its users' rights and relive the human moderators in a more appropriate way. Therefore, this model case recommends to any platform provider to rather extend than lower their scope of hate speech filtering algorithms.

3.2 Legal Challenges for Platform Providers in their Hate Speech Removal Process

Although, the technical analysis has considered the second algorithmic approach filtering content within a wider scope as more appropriate for platform providers to protect their users' fundamental rights and to reduce human content moderators' involvement, the development and successful real-world implementation of any algorithmic approach strongly depends on the requirements provided by the applicable law. Necessarily, platform providers before offering their services within the EU must comply with both current EU law and, if primary applicable, with Member States' national law.

In the first part of the legal analysis, the EU law will be investigated to what degree platform providers are hold liable for the content hosted by their services ('*de lege lata'*). Transferred into the model case this might affect their choice of algorithmic approach, when urged to either widen or narrow the scope of their algorithmic decision boundary. Based on the knowledge gained, the second part will question the role of EU legislation in the future ('*de lege ferenda*') by asking whether the law should become the role of a gatekeeper for content moderation technology or rather open doors for innovation through liberal legislation?

Finally, it will be considered the degree of human involvement already required within current EU legislation as well as its need for future legislation.

3.2.1 To what Extent are Platform Providers hold Liable for the Content they host under current EU Law?

To properly assess and classify the current liability regime (‘*de lege lata’*) for platform providers within the EU (also referred to as ‘Internet/online intermediaries’),[69] the applicable law will be analysed in chronological order. Starting at EU level with the introduction of the first intermediary liability regime provided by the E-Commerce Directive, it will be investigated further how certain Member States implemented the Directive in national law. Followed by some EU initiatives aiming to incentivize platform providers to voluntarily adopt proactive measures, the loop will be closed with the recently proposed Digital Services Act, a new regulation introducing a revised intermediary liability regime.

3.2.1.1 The Liberal US Legislation

Before starting the analysis at EU level, a brief degression will be taken into the liberal US law, which is important to comprehend the growth of today’s biggest internet platforms as Facebook, YouTube or Twitter are. More precisely, it is Section 230 c of the Communications Decency Act (CDA), as amended by the US Telecommunications Act of 1996, that constitutes of two immunities protecting online intermediaries from being hold accountable for taking steps to restrict illegal and other forms of inappropriate content.[70] Even though both immunities are provided with exemptions (see Sec.230e), Sec. 230c 1. CDA establishes that no user or provider ‘shall be treated as the publisher or speaker of any information provided by another information content provider’, whereas Sec.230c 2. A. protects the same subjects from

> ‘*any action voluntary taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected (…)’.*

Although, both immunities are labelled as the so-called ‘*Good Samaritan principle*’, the most common use and focus of this research is the latter immunity (Sec.230c 1.A.), meaning immunities for taking down, as distinct from immunities for leaving up.

---

[69] ‘Internet’s intermediaries [..] give access to host, transmit and index content, products and services originated by third parties on the Internet or provide Internet based services to third parties.’ OCED, ‘The economic and social role of internet intermediaries’, 6.

[70] See also §512 of the Digital Millennium Copyright Act (DMCA) that under certain conditions protects services providers from monetary liability for copyright infringements based on allegedly infringing activities of third parties.

Moreover, the Good Samaritan Principle, as concluded by Barata, (1) prevents intermediaries from any liability for third-party content they share, even from illegal content they fail to detect or assess, (2) provides them with the freedom to set their own content policies, apart from any applicable legal requirements regarding the content or nature of policies, and (3) encourages platform providers to ban, police and remove not only presumed illegal posts, but also lawful, yet still harmful content. [71] This enables providers to invite their users to flag inappropriate content, without a fear that such notifications will create a liability risk for the provider.
As fundamentally characterized by the Good Samaritan principle, US legislation constitutes one of the most liberal regulatory frameworks for platform providers and is recently criticised within the United States but will not discussed further within this study.[72]

Allocation to the Model Case

According to the previous technical analysis, platforms are provided the freedom to choose one or the other algorithmic approach provided in the model case by either narrowing or extending the scope of their filtering algorithms. Enabled to independently develop their own detecting methods pursuant to the content policies they face, they must neither fear the risk of overblocking when filtering too much content, nor the risk of liability when filtering too little content. Although, there might be an increasing pressure from the platform users' side to better protect their rights and to become more proactively, at least from the legal side intermediaries (until now) are protected from liability enabling many of them to grow fast.
Due to the low level of liability within the US, it is more likely that platform providers as soon as they pursue to offer their services abroad, must comply with stricter regulations.

3.2.1.2 The E-Commerce Directive

Two decades ago, the Council and Parliament of the European adopted the E-Commerce Directive (ECD)[73] to remove obstacles to cross-border online services and to establish an intermediary liability regime within the EU. The so-called 'safe harbours' or 'immunities' grant intermediaries a liability protection when they engage in the provision of 'mere conduit'(Art.12)[74], 'caching' (Art.13) and 'hosting' services (Art.14) and were significantly relevant for the growth of e-commerce within the EU. Due to each safe harbour requiring

[71] Barata, 'Positive Intent Protections', 6.
[72] Cordeiro, 'Free Speech in the Internet Era', 57-59.
[73] Direction 2000/31/EC.
[74] All following Articles belong to the ECD, unless not otherwise declared.

intermediaries to meet several conditions to benefit from the liability exemption, Art. 14 addresses hosting services that

(1) perform the role of 'a mere technical, automatic, and passive nature' (recital 42),

(2) do not have actual knowledge of illegal activity or are not aware of facts from which illegal activity is apparent (Art. 14 1.a) and

(3) act 'expeditiously' once knowledge or awareness are acquired (Art. 14 1.b).

This liability exemption regime is accompanied by Art. 15 which will be discussed in more detail later in this Chapter and generally prohibits Member States from introducing general monitoring obligations on intermediaries. Although both specific and voluntary monitoring is allowed,[75] especially the latter could lead to the awareness or knowledge of facts or circumstances from which an illegal activity is apparent and bears the risk for hosting services to lose their liability exemption.[76]

In other words, the more active platform providers monitor the content hosted by their services, the higher is the probability of acquiring 'actual knowledge' and the chance to lose their liability exemption. Despite the discouraging impact on platform providers to take voluntary measures,[77] Art.14 additionally lacks any precision, especially regarding a definition of 'actual knowledge' or which intermediaries' intervention can clearly be classified as active or passive.[78] Some clarification is barely achieved through the effort of a few individual cases: Regarding the interpretation of the Court of Justice of the European Union (CJEU) in the L'Oréal case,[79] the intermediary becomes liable in situations where it achieves actual or specific knowledge of the illegality 'as result of an investigation undertaken on its own initiative'[80] or by receiving proper notification that allows intermediaries to become 'actually aware of facts or circumstances on the basis of which a diligent economic operator should have identified the illegality.'[81]

---

[75] As frequently explained in case law (C-324/09 L'Oreal v ebay, C360/10 SABAM v. Netlog) and broadly discussed in literature e.g., Senftleben, 'The Odyssey of the prohibition on general monitoring obligations on the way to the Digital Services Act.'.

[76] Kuczerawy, 'The Good Samaritan that wasn't'.

[77] Barata, 'Positive Intent Protections',1.

[78] Schwemer, 'Legal analysis of the intermediary service providers of non-hosting nature', 32. See also Hoboken, 'Hosting intermediary services and illegal content online', 7.

[79] C-324/09.

[80] C-324/09. Rec. 122.

[81] Ibid. Rec. 122.

Moreover, it is significant to perceive that knowledge and awareness do not equal the need to act upon any kind of notice to avoid liability under Art.14. Recital 46 suggests that intermediaries need to take proper and balanced decisions by bearing in mind the 'observance of the principle of freedom of expression and of procedures established for this purpose at national level'.[82]

This leads to the assumption that the ECD does not adequately promote the adoption of voluntary and proactive content moderation policies by hosting services, but rather the opposite. As Sartor explains, 'making the protection conditional on passivity would induce a hands-off approach that would result both in an increased quantity of online illegalities and in the failure to satisfy the users that prefer not to be exposed to objectionable or irrelevant material'[83] (as these motivations explain the US Good Samaritan principle).

Allocation to the Model Case

The regulatory framework of the ECD, which is applicable since 2001, rather represents the first instead of the second algorithmic approach provided in the model case. In contrast to the US regulatory framework, it does neither include a 'Good Samaritan principle' nor does it provide platform providers the freedom to extend their algorithmic decision boundary without fearing liability consequences. The ECD's requirement to act expeditiously once knowledge or awareness is obtained, rather encourages platform providers to remain passive by filtering less content within a narrower algorithmic decision boundary to not bear the risk of losing liability.

3.2.1.3 Proactive Measures

EU Initiatives

Whereas the ECD within the safe harbour of Art. 14 tends to discourage intermediaries to proactively monitor content, the European Commission within several initiatives aims to stimulate the contrary: For example, the Commission's 2016 'Code of Conduct' on hate speech, the 2017 Communication on Tackling Illegal Content[84] and the companion 2018 Recommendation on Measures to Effectively Tackle Illegal Content Online,[85] face to incentivize platform providers to voluntary monitor, remove or otherwise decrease the visibility

[82] Barata, 'Positive Intent Protections', 9.
[83] Sartor, 'Providers Liability', 27.
[84] COMM (2017) 555 final.
[85] C (2018) 334 final.

of unwanted content. In particular, the idea of 'so-called 'Good Samaritan' actions[86] was already expressed in the Communication, when the EU Commission argued that 'taking such voluntary, proactive measures does not automatically lead to the online platform losing the benefit of the liability exemption provided for in Art. 14 of the E-Commerce Directive.' Whereas this statement might suggest an adoption of Good Samaritan measures by concurrently covering the immunities provided under Art.14, the obligation for platform providers to 'act expeditiously to remove or to disable access to the information in question upon obtaining such knowledge and awareness', remains in place.[87]

Although, the non-binding initiatives taken by EU institutions clearly demonstrate an increasing awareness for platform providers to become more actively involved in the monitoring process, the ECD's liability exemption regime combined with the risk of losing it, still scares hosting services to leave their 'safe harbour' and become proactive.

Dynamic Injunctions

A shift towards a more proactive involvement of hosting services is not solely forwarded by EU initiatives, also the CJEU recently changed its interpretation of the 'prohibition on general monitoring obligation' of Art. 15 ECD with significant consequences for hosting services:

In addition to Art.14's, the liability exemption is supplemented by Art. 15 providing that intermediaries may not be obliged to monitor their services in a general manner to detect and prevent illegal content:

Member States shall not impose a general obligation on providers, when providing [mere conduit, caching, and hosting services], to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity.

However, the prohibition of monitoring obligations only refers to the monitoring of a '*general nature'* but does not concern monitoring obligations in a '*specific case'* (Recital 47). As its distinction lacks any clarification, the prohibition of 'general' monitoring obligations has significantly limited the permissible scope of the measures that can be imposed on online intermediaries for the enforcement of third parties' rights and the prevention of future infringements. Even though such measures seem required under the ECD: Whereas Rec. 45 allows injunctions on intermediaries to both terminate and prevent infringements, also Rec. 48

[86] COMM (2017)555 final, 3.

[87] Barata, 'Positive Intent Protections', 12.

indicates that the imposition of duties of care specifically on hosting providers to detect and to prevent should be allowed.

However, from an online intermediary's perspective the most suitable option to fulfil the obligation of taking preventive action, is the adoption of filtering systems, i.e., algorithmic content moderation systems. In how far is the imposition of filtering duties compatible with the prohibition on general monitoring obligations?

According to Senftleben, who recently described this issue as an 'Odyssey',[88] the two following interpretations of 'general monitoring' within Art.15 (1) have emerged in the context of policy and academic debates:

- Option A: Any general monitoring is prohibited and can never be classified as monitoring in a 'specific case'(Rect.47). To find violations of a specific right it is prohibited as it would require the service provider to sift through all content on the platform. Therefore, filtering duties would not be allowed under this interpretation.
- Option B: General Monitoring is allowed as in search of infringements of a specific right. The generality and speciality of monitoring is not determined by what is being monitored, but by the generality and speciality of the subject matter which the monitoring seeks to identify in upload content.[89] Separated in two sub-categories under this option, general monitoring of violations of a specific right by means of automated filters is not prohibited if (1) ordered by a court or (2) ordered by a court or triggered by a notification.[90]

In contrast to the controversial debates, the CJEU, at least until recently, has interpreted Art.15(1) more consistently: For example, in the case 'Scarlet Extended' the court concluded that the obligation to implement a filtering system, which entails active observation of all communication on the platform, constitutes a general monitoring obligation.[91] The CJEU based its decision on whether a measure constitutes a general monitoring obligation by assessing two factors: (i) the relative amount of traffic to be monitored per user; and (ii) the relative number of users that must be monitored.[92] The CJEU concludes that an obligation to monitor becomes

[88] Senftleben, 'The Odyssey of the prohibition on general monitoring obligations on the way to the Digital Services Act.', 8.

[89] Ibid.9.

[90] Ibid, 8.

[91] C-70/10, rec. 37–39

[92] C-70/10 rec. 38–40; C-484/14, rec. 87

general in nature, where it covers *all* data of *all* users.[93] This was confirmed in the case 'Sabam/Netlog', where the Court found that an injunction against a hosting service provider, that required an undefined and unlimited filtering system, would be incompatible with EU law and would strike an unfair balance between the applicable fundamental rights.[94]

According to Senftleben's two interpretational options on Art.15 (1), the CJEU had appeared to unambiguously embrace Option A, especially in areas of copyright and trademark law. However, within the area of defamation, the Court has shifted towards Option B (1),[95] widening the scope of Art. 15(1) within its landmark decision in Eva Glawischnig-Piesczek v. Facebook:[96] The CJEU held that an injunction ordering a host to remove content which includes not only future *identical* infringements, but also *similar* infringements to the infringement in the initial proceedings, herein referred to as 'dynamic injunctions'[97], would be compatible with Art. 15 (1) ECD.
In the case, a Facebook post related to Eva Glawischnig-Piesczek was found to be harmful, so that the Austrian politician requested Facebook to disable access or remove it, and furthermore, to prevent any further posts that had an equivalent meaning. Understandably argued, an injunction would only include the exact identical defamatory wording that was initially used, but the infringer could simply re-formulate the harmful message and post it again. Although, Facebook argued that this would lead to monitoring all passing data on its platform, the court concluded that the dynamic injunction covers both 'identical and equivalent' content, by not imposing any general monitoring obligation on host providers.[98]

The landmark decision obviously shifts the court's understanding of 'general monitoring' from Senftleben's Option A to Option B (1) and brings anything but clarity to the debate: On the one hand the court's decision seems comprehensible regarding the effectiveness of filtering and the risk of re-uploading the same content by the same party not changing its infringing character. On the other hand, the court did not provide any definition on 'equivalent' and leaves it open to the hosting provider to assess a post as similar defamatory. Regarding the latter the Court

[93] Van der Donk, 'How dynamic is a dynamic injunction?', 608.
[94] C-360/10.
[95] Senftleben, 'The Odyssey of the prohibition on general monitoring obligations on the way to the Digital Services Act.', 8
[96] C-18/18.
[97] Van der Donk, 'How dynamic is a dynamic injunction?', 608.
[98] Ibid.

added yet, that hosting providers are not required to 'carry out an independent assessment of the content' and that, a dynamic injunction should contain specific elements which are properly identified in the injunction, such as the name of the person concerned, the circumstances in which that infringement was determined, and what constitutes equivalent content.[99] Important to add is that the Courts rested on the prior assessment by the national court that has declared the content as illegal, and according to Senftleben's options rather offers support for Option B (1), without allowing notification on a mere rightsholder without any Court involved (B (2)).[100] However, further questions arise in terms of practical feasibility, especially for small- and middle-sized hosting services, and whether the court presumes a level of sophisticated technology that until now, or likely never, will exist.[101]

Concludingly, the CJEU with its decision in Glawischnig-Piesczek v. Facebook, broadens the scope of Art. 15 (1) ECD by including blocking injunctions that can oblige hosting services to proactively filter not only identical, but also equivalent infringements, unless those injunctions are effective, proportionate and strike a fair balance between fundamental rights. This decision demonstrated the court's increasing awareness for the urgent need for a more proactive involvement of platform providers, but also built-up pressure on the EU legislator to provide clarification within the upcoming Digital Services Act.

Allocation to the Model Case

Based on the low level of algorithmic decision boundary allocated to the ECD, platform providers through the EU initiatives but also through the revolutionary change in the CJEU's interpretation of Art. 15 shall be encouraged to voluntary raise their decision boundary or at least to allow some extensions towards content that is 'identical or equivalent' to content previously declared as illegal. Nevertheless, one must keep in mind that any voluntary participation does not exempt hosting services from the liability under (Art.14) ECD.

3.2.1.4 Member States' National Law

As the ECD constitutes a non-binding directive, Member States have been gone beyond the minimum requirements of the ECD and implemented stricter regulatory measures during the

[99] C-18/18, rec. 45.

[100] Senftleben,' The Odyssey of the Prohibition on General Monitoring Obligations on the way to the Digital Services Act', 15.

[101] Daskal, 'A Europan Court decision may usher in global censorship'.

last decade. E.g., Germany has adopted its Network Enforcement Act (NetzDG) in 2017, requiring platforms to block illegal content within 24h and to react on complaints within 48h. The law urges platform providers with fines, up to 50 million euros,[102] to avoid any systematic failure in deleting illegal content. Taken as referee, further (former) Member States, such as the UK and Austria adopted similar regulations.[103] Even France proposed a law[104] aiming to fight hate speech on high-visibility social media platforms and search engines. The obligation proposed to remove 'clearly illegal' hateful content within 24 hours of being notified of it, should have been enforced by imposing fines up to 20 million euros or 4% of the company's global annual revenue. Finally, the law was adopted in June 2020 without the provisions of hate speech removal, as the Conseil constitutionnel agreed with an opponent voice that the threat of high fines combined with a short time frame to evaluate the illegal nature of flagged content were unconstitutionally harmful to the freedom of expression.[105]

Allocation to the Model Case

Although, it is still a minority of EU Member States that has implemented their own regulatory approaches in combating hate speech online, one might observe an increasing awareness for stricter regulation within the EU. While some national regulation, such as the German and Austrian, urge platform providers to proactively remove more content within shorter timeframes, they tend to face the second algorithmic approach by urging platform providers to widen the scope of decision boundary. Thereby and as feared by the French Conseil constitutionnel, they subordinate the freedom of expression to the right to non-discrimination. Nevertheless, the EU Commission anticipated the adoption of its proposed Digital Services Act as a binding regulation. Once it is in force, the Member States' national law must be adjusted.

3.2.1.5 The Digital Services Act

In response to the increased pressure to proactively involve platform providers and in order to harmonize the patchwork of Member States' own regulatory initiatives, the Commission on the 15th of December 2020 published a proposal for a binding regulation, the Digital Services Act (DSA).[106] Aiming to replace the ECD, the proposal retains the ECD's key principles, in

[102] §4 (2) NetzDG.

[103] Kommunikationsplattformen-Gesetz.

[104] Loi Avia.

[105] Boring, 'Constitutional Court Strikes Down Key Provisions of Bill on Hate Speech'.

[106] Meant by this in any case is the proposal of the DSA.

particular the requirement of acting expeditiously once knowledge or awareness of illegal activity ('*or illegal content*' as included in Art.5 DSA, former Art.14 ECD) is obtained but also the prohibition of the imposition of general monitoring obligations (now Art.7 DSA, former Art.15 ECD). Some clarification is added in e.g., Recital 18 stating that the liability exemption should not apply, where the provider plays an active role of such kind as to give it knowledge of, or control over, that information.' According to Kuczerawy, that confirms previous interpretation and the ruling of the CJEU in e.g., L'Oréal v. eBay, that providing 'active' services may lead to the loss of immunity if knowledge could be established.[107]

Moreover, the DSA faces to include a 'Good Samaritan principle' in its Art. 6 clarifying that intermediaries may not lose their liability protections '*solely because they carry out voluntary own initiative investigations or other activities aimed at detecting, identifying and removing, or disabling of access to illegal content […]'*. Further specified in the Recitals, providers can acquire knowledge and awareness through '*own-initiative investigations or notices […] in so far as those notices are sufficiently precise and adequately substantiated*' (Rec. 22), whereas the mere fact that providers undertake investigations activities '*does not lead to the unavailability of the exemptions from liability set out in this Regulation, provided those activities are carried out in good faith and in a diligent manner*' (Rec. 25). According to Kuczerawy this applies to activities taken to comply with the requirements of EU law, including those set out in the DSA *'as regards the implementation of their terms and conditions*.' The researcher emphasises the benefits for hosts, that can easier and faster remove content based on its terms and conditions since that awards more discretion to remove content that is unwanted but not necessarily illegal. A further advantage of the inclusion of platforms' self-regulatory regimes is that they are not necessarily being considered as 'active' and could broaden their scope of filtering mechanisms, which in turn assuages policy makers.[108]

However, the ECD's main problem carried forward by the DSA is a lack of precise definitions as well as the differentiation between active and passive monitoring. In particular questions arose such as, what happens to activities other than own-initiative investigations? When Rec.22 demands notices to be 'adequately substantiated', does it imply that unsubstantiated claims are not a sufficient basis for knowledge?[109] Especially regarding the acquisition of 'actual

[107] Kuczerawy, 'The Good Samaritan that wasn't'.

[108] Ibid.

[109] Barata, 'The Digital Services Act and the Reproduction of Old Confusions'.

knowledge' the DSA remains uncertain. For example, if a filtering mechanism is trained to detect one specific type of illegal content (e.g., terroristic content), but oversees another type of content (e.g., hate speech), would the remaining content result in liability because the host 'knew' or 'should have known' about the illegality?[110]

Frequently named as the 'Good Samaritan principle'[111], but potentially having a different effect than the US Sec. 230 CDA, Art. 6 DSA might lead to more content removals. This in turn raises experts' concerns regarding over-blocking and the protection of users' freedom of expression,[112] which is one of the main goals of the DSA (e.g., stated in Recital 22 that the '*removal or disabling of access should be undertaken in the observance of the principle of the freedom of expression')*. Exemplarily, Peukert questions, why the DSA is appropriate to all types of illegal content, including copyright infringements as well as defamatory content, and demands to allow automated decision-making only in cases of manifestly illegal content for which no independent assessment is needed. [113]

Considering the previous, the DSA faces two main objectives: On the one hand it aims to incentivize hosting services to voluntarily adopt own-initiative investigations to comprehensively detect undesirable content, but on the other side it prevents them from removing too much content to protect the users' freedom of expression. This conflict confirms that the challenge of balancing the users' fundamental rights within the content moderation process not only arises from a technical perspective (as analysed in Chapter 3.1.) but is also highly discussed by the EU policy makers. It remains questionable whether the EU Commission within its final version of the Digital Services Act will bring clarity to the debate.

Allocation to the Model Case

Finally, the DSA tends to follow the first algorithmic approach in almost the same way as the ECD does. Still accompanied by a particular degree of legal uncertainty, the DSA encourages platform providers to rather keep their decision boundary low to avoid a loss of liability. Nevertheless, a low raise of the filtering scope might be observed by including platforms 'terms and conditions' as well as extensions through its Art.6 enabling 'own-initiative investigations.'

---

[110] Kuczerawy, 'The Good Samaritan that wasn't'.

[111] See Kuczerawy, 'General monitoring obligations', Barata, 'The Digital Services Act and the Reproduction of Old Confusion' or Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation'.

[112] See i.a. Peukert, 'Five Reasons to be Sceptical About the DSA' or Harrison, 'Freedom of Expression & Hate Speech in the EU's Digital Age,' 36.

[113] Frosio, 'Taking Fundamental Rights Seriously in the Digital Platform's Liability Regime'4.

3.2.1.6 Conclusion

Regarding the regulatory approaches within the EU, especially compared to the liberal US legislation, one can observe a shift towards increased awareness for regulatory approaches aiming to integrate platform providers more responsibly into the content moderation process. Although, national and European legislators seem to face the same goal, their choice of algorithmic approach diverges. While the DSA as well as its predecessor rather follow the first approach (with some potential extension in the DSA's Good Samaritan Clause), the e.g., German or Austrian national legislations, in contrast to e.g., the French, rather tend to the second approach. How the different approach shall be brought in line and what changes should be made to the DSA proposal, will be analysed in the next part.

3.2.2 How should EU Law be Framed to hold Platform Providers appropriately liable?

Although, from a technical perspective the extension of an algorithmic decision boundary might be the appropriate way to protect platform users' rights, from a legal perspective the assessment might be different. As demonstrated above, different national legislators as well as the EU legislator itself set different incentives for platform providers to programme their algorithms.

However, what criteria should be considered from a legal perspective when shaping a new strategy in content moderation within the EU? Shall the law ('de lege ferenda') assume the role of a strict gatekeeper or rather open doors for innovative technology through liberal legislation?

To appropriately categorize the new approach, one might take the liberal US legislation and the strict German national approach as two extreme points of a regulatory scale. Somewhere in between, a new regulation aims to balance the conflicting interests of appropriately protecting platform users' rights and of ensuring an adequate freedom for innovation and attractiveness for foreign providers to offer their services within the EU. Regarding the latter, the limited liability provisions of the ECD have essentially promoted the growth and prospering of innovation and entrepreneurship within the EU. Keeping that in mind, a new legislation might undermine these protections by introducing new responsibilities that potentially impact small- and middle-sized companies or start-ups disproportionately and could shrink the diversity of platforms and hosts available to support a broad range of expression online.[114]

[114] CDT, 'Nine Principles for Future EU Policymaking on Intermediary Liability'.

According to the huge diversity and worldwide reach of platform providers, one might go a step further by provocatively questioning whether regulation of intermediaries should still come from governmental side. Or, since platform providers a long time ago left national borders, if they should rather be replaced by intermediaries' self-regulatory regimes. As the answer to this question goes beyond the scope of this study, one might reduce it to the question whether the determination of illegal actions should be done by intermediaries or rather left to courts. As this is more likely a question of practicability, an involvement of courts would potentially contribute to a stronger legal certainty within content moderation by providing clear definitions and consistent standards, but otherwise might fail due to increasing timely effort and a cost-intense inclusion of educated lawyers and judges.[115]

However, under the overarching goal of encouraging platform providers to moderate content in a proportionate manner and to devote appropriate efforts to tackle illegal content without unnecessarily burdening users' rights to freedom of expression, the following aspects, provided by Barata,[116] should be considered, when walking back and forth the regulatory scale:
Any structural or systematic obligation shall meet the criteria of *proportionate, transparent, commercially reasonable* and *generally flexible*.
Starting with the first criterium of *proportionateness*, the focus should no longer be on the pure outcome of content moderation processes, but rather on the shift from quantitative towards qualitative removing. Specified by Barata, intermediaries should not be evaluated on whether they have removed 'enough' illegal content or if they have failed to enforce their policies 'consistently' or 'comprehensively'. It would solely create a strong incentive to over-removal of lawful speech and concurrently reduce any space for individual adoptions of their policies. Legal regimes must clearly differentiate between the administrative responsibility related to failure to fulfil regulatory obligations and the loss of immunity regarding hosted content. Sanctions should only be applied in cases of demonstrated systematic failure to respond to valid notifications of illegal content. This especially applies to the proposed DSA, that was already criticized by Kuczerawy for punishing providers that became active but failed to filter one specific type of illegality.[117]
Regarding the second and third criteria listed, a new regulatory framework should encourage intermediaries to *transparently* share their impact of their content moderation systems and to

[115] Langvardt, 'Regulating Online Content Moderation',1368.
[116] Barata, 'Positive Intent Protections', 14-15.
[117] See. fn.110.

develop mechanisms to evaluate their effectiveness. Additionally, the criterium of *commercially reasonable* requires that no penalties or liability should arise from notifications of apparent violations of content policies or 'terms and conditions.'

Finally, all regulatory measures shall keep a *general flexibility* to enable the emergence of different approaches of content moderation varying in nature, function and organizational structure of intermediaries. Under the principle of technological neutrality, meaning that '*legislation should define the objectives to be achieved, and should neither impose, nor discriminate in favour of, the use of a particular type of technology to achieve those objectives*.'[118] no technological solution shall legally be mandated in detail by keeping in mind current deficits of algorithms still limited in e.g., parsing the nuanced meaning of context.

All in all, the listed criteria provide comprehensive guidelines for a new legal strategy in content moderation processes and should especially be implemented in the proposed DSA to find a balanced location on the regulatory scale. By constantly keeping the door open for new technology on the European market, the EU gatekeeper should set new regulatory incentives to rather encourage platform providers to voluntary adopt monitoring measures than to urge them with high fines.

Technically spoken, a generally low level of algorithmic decision boundaries shall be kept allowing innovation and providers' self-driven development, but in any case shall become flexible for extensions.

3.2.3 To what Extend does and should EU Legislation/Jurisdiction require Human Involvement?

Screening through current legislation the awareness for human content moderators tends to zero. Whereas the current ECD as well as the proposed DSA, despite the redress mechanism in Art. 17 (5) DSA[119], not even mention the human content moderators, at least within the Artificial Intelligence (AI) Act[120] the wording is found in Recital 48 demanding that '*the natural person to whom human oversight has been assigned have the necessary competence, training and authority to carry out that role*.'

[118] Merton, 'The Technological Society', 11.

[119] Art.17 (5) DSA demands platform providers to ensure that decisions are not solely taken on the basis of automated means, but do not mention 'human (content) moderators.'

[120] COM (2021) 206 final.

Although this formulation could be interpreted as suitable to apply for human moderators' protection (e.g., by requiring specific (legal) competence to evaluate content or the provision of mental trainings to avoid psychological harm), Recital 48 considers 'high-risk AI systems' and rather includes the provided phrase to appease the user for the technical competence of the human developer.

However, recent ruling by the CJEU in Glawischnig-Piesczek has demonstrated that a clarification regarding the involvement of human moderators is urgently needed: As described above,[121] the case broadens the scope of Art. 15 ECD by injunctions ordering a host to remove content which is identical or equivalent to the content that has previously been declared as unlawful. This leaves many open questions regarding the practical feasibility, but in particular on the assessment of 'equivalent' infringements: When is content considered to be similar? Does it require an independent assessment by the platform provider and consequently the hiring of additional human content moderators to compensate the algorithms deficits?
Without providing any precise definition of 'equivalent', the Court simply denies the need for an 'independent assessment of the content to be carried out by the host provider and refers to the injunction that should precisely describe the specific elements of the infringement's nature. Exemplarily mentioned are the name of the person concerned by the infringement determined previously, the circumstances in which that infringement was determined and what equivalent content is.[122]
What the Court seems to disregard, is the technical limitations of filtering mechanism. As van der Donk explains, an infringement considered as similar in a legal sense does not necessarily constitute a similar infringement in a technical sense.[123] A hate speech related post could simply be rewritten and a video or image including defamatory content could easily be changed in its data or file with the consequence of a completely different appearance to the algorithm. As content filtering systems solely decide in a rational way whether content is identical or non-identical, they are not yet sophisticated enough to recognize the nuanced meaning of context, or even less of 'equivalent' content. This leads to the assumption, that the Courts' interpretation of Art. 15, due to the shortcomings of filtering technology, must consequently be compensated by the additional hiring of human content moderators.

[121] See Chapter 3.2.1.3.

[122] C-18/18, rec. 45.

[123] Van der Donk, 'How dynamic is a dynamic injunction?', 609.

However, an explicit requirement for human involvement in current EU legislation is still missing. Therefore, some suggestions will be provided to be integrated in the proposed DSA and to legally protect human content moderators based on their understanding of Sec. 2.4.:
The overarching goal of a new regulation should be to incentivize platform providers to primarily reduce their total amount of human content moderators and to secondary, improve their individual working conditions. This means that any employment of human content moderators shall be justified by the 'proportionality principle', meaning that their involvement shall pursue a legitimate purpose, be suitable and (at minimum) necessary to achieve this purpose and be appropriate in a narrower sense. If justified, legal requirements according to the demands recently published by more than 100 content moderators[124] shall include:
(1) the prohibition of outsourcing human content moderation to third parties
(2) to ensure an appropriate salary, as well as
(3) the option for requesting adequate psychological support and trainings.

As finally to keep in mind for the implementation of new legislation, is an appropriate balance between regulation and liberty to innovation. This means that no technical tools shall legally be suggested in detail to keep any platform providers' freedom to independently develop technology with the purpose of relieving the human content moderators.

3.2.4 Conclusion

Summarizing the legal challenges to shape a new strategy in content moderation within the EU, an increasing awareness under legislators to involve platform providers proactively in the filtering process can be observed. By balancing an adequate level of the freedom of innovation, but concurrently strengthening platform providers liability, the EU legislator, in contrast to some national legislators, remains reluctant in its proposed DSA. It rather incentivizes platforms to more own-initiative investigations instead of urging them with high fines. Assessed as a good first approach, the analysis especially appreciated the Good Samaritan principle, but even criticized the still blurry distinction of liability and its exemption. However, still to be included is a specific requirement for platform providers to reduce and if necessary, improve human content moderators' working conditions.

[124] Banerjee, 'Facebook's content moderators demand for an end to culture of 'fear and secrecy'.

Compared to the outcome of the technical analysis, from a legal perspective rather the first approach should be chosen to incentivize platform providers to voluntarily raise their algorithmic decision boundary by developing individual measures.

4 Practically thought: How could a Solution for a New Strategy in Content Moderation look like?

According to the previous analysis of the technical and legal challenges for a new strategy for content moderation, its results and recommendations vary in dependence on its addressees. This study does not aim to provide a one-size-fits-all solution satisfying all interests of all stakeholders and bearing the risk of neglecting one or the other aspect. Therefore, the following part provides three independent approaches, focusing on a technical, legal and human moderator's perspective. Whereas the first approach of 'Quarantining online hate speech' addresses the technical recommendations, the second approach of 'Contesting algorithms' provides a solution for the EU legislator. Finally, the 'Crowdsourced Image Moderation' contains a tool that protects human moderators from directly being confronted with violative images.

4.1 Quarantining Online Hate Speech

The first approach of 'quarantining online hate speech' addresses the outcome of the *technical analysis* demanding for a stronger protection of users' rights of non-discrimination by extending the scope of platforms' filtering algorithms and is based on the research of Ullmann and Tomalin in 'Quarantining online hate speech: technical and ethical perspectives.'

By analogy with the quarantining of malicious computer software, the researchers' idea is to (potentially temporary) censor posts that are automatically classified as being harmful and mark them with an alert, which protects the recipient viewing the harmful content in the first instance.[125] Provided with the name of the content's author and the notification that the flagged post might contain a specific type of hate speech, the recipient can avoid a potential infringement of its right to non-discrimination and refuse the publication of the post.

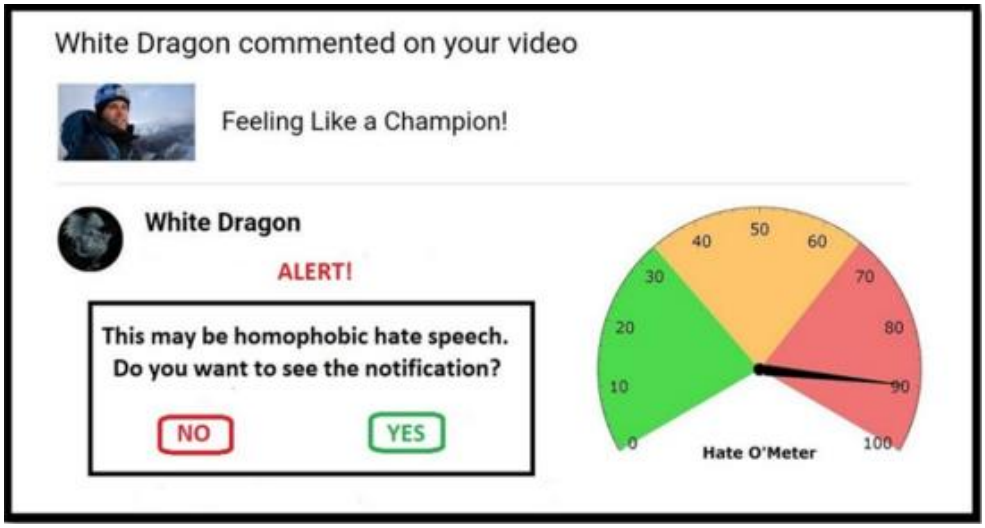
Based on the example used in the research, the American Skier, Gus Kenworthy after coming out as gay in October 2015 was subsequently bombarded with homophobic slurs on his YouTube channel.[126] If quarantining were deployed to these slurs, Kenworthy would have

[125] Ullmann, 'Quarantining online hate speech', 69.

[126] Ibid., 75.

received an alert for each post, requesting him whether he wants to read the post by now knowing the name of the author (e.g., "White Dragon") and that the post potentially contains homophobic hate speech. In addition, the recipient might be provided with an indication of the degree of severity of the post by the value specified on the Hate O'Meter graphic.

Figure 5. Homophobic Hate Speech quarantined and provided with a Graph indicating Degree of Severity of the Post.[127]



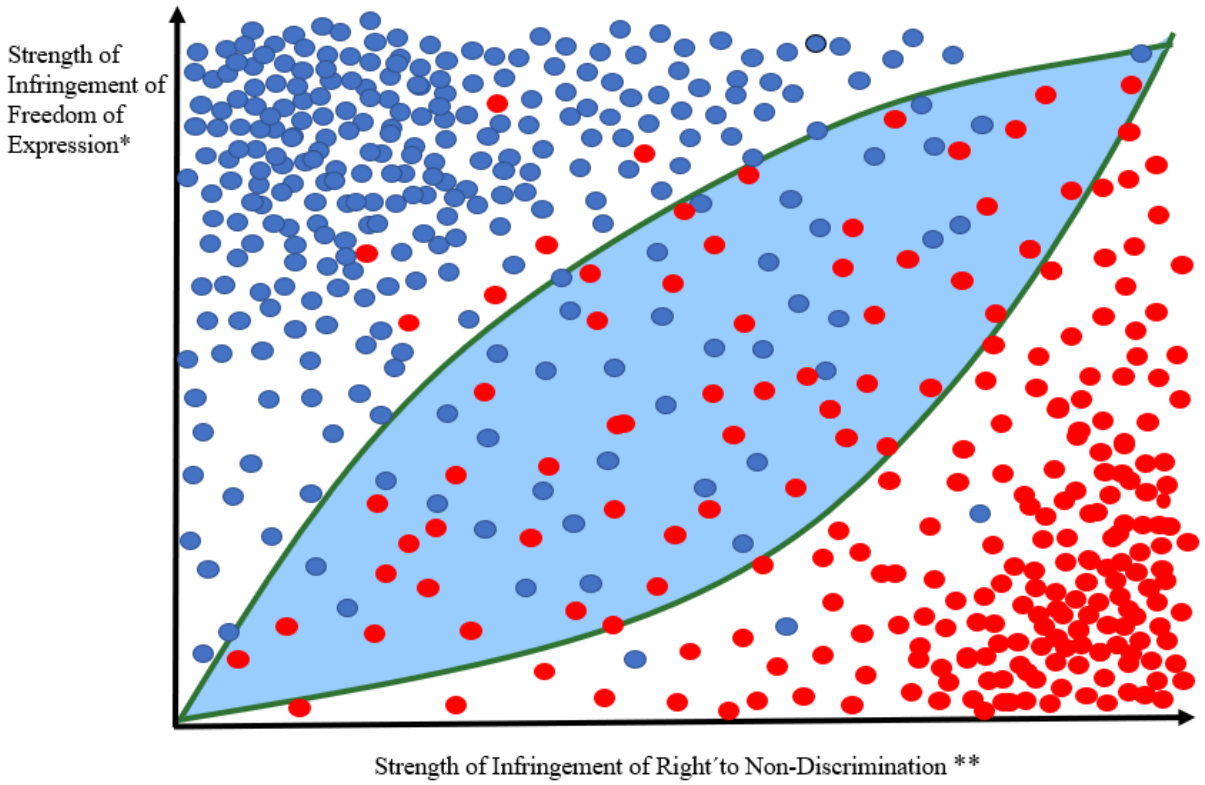
Applied to the model case (Fig.6), a 'quarantining zone' might be imbedded within two decision boundaries filtering potential harmful content within a wider and a narrower scope (also known as approach one and two of the model case).

Whereas all posts underneath the lower boundary would be flagged without any warning, the posts and comments above the wider decision boundary would remain visible on the platform. Specifically, content falling *within* the quarantining zone would be added by a warning that it potentially contains a specific type of hate speech. Compared to an algorithm that filters content ultimately and within a wider scope, the advantage of quarantining algorithms is that posts are neither entirely permitted nor entirely prohibited, but rather held in a limbo for a finite time until they have been assessed by the relevant recipients or moderators.[128] Thereby concerns regarding a potential overblocking and infringement of other users' freedom of expression could substantially be reduced.

[127] Ullmann, 'Quarantining online hate speech'., 75.

[128] Ibid., 75.

Figure 6. Algorithm keeps Ambiguous Content in Quarantine for Double-Check.[129]



Taken as the basis, the concept of quarantining might be extended in various ways: Referring to the idea of Ullmann/Tomalin, it could also be integrated a Hate O'Meter that demonstrates the tendency of content towards one or the other decision boundary. E.g., the more a specific content would tend towards the lower boundary within the quarantining zone, the higher the pointer of the Hate O'Meter raises. The integration of a Hate O'Meter could be helpful for recipients to assess whether they want to allow the content or not.

[129] Illustration by the author.

Another feature one might integrate in the quarantining regards the protection of underage users. For those, any quarantined content should be blocked immediately without any option for the recipient to decide whether it should be visible subsequently. Its assessment should only be taken by a human moderator or by a more specialized algorithm.

Concludingly, the idea of quarantining potentially harmful content addresses both the demand for an algorithm filtering content within a wider scope and thereby strengthening the right to non-discrimination as well as its raising concerns about a potential overblocking in an adequate way. The two-step concept of firstly blocking more potential harmful content but secondly offering the chance of being assessed as lawful, enables comprehensive protection to users and might be extended by additional features, such as a Hate O'Meter or a child protection blocking.

4.2 Contesting Algorithms

Referring to the results of the *legal analysis*, the EU legislator by proposing the Digital Services Act remained reluctant in urging platform providers to extend the scope of their filtering algorithms in order to protect the users' freedom of expression.[130] Indeed, through proactive measures, such as non-binding regulations (e.g. Communication on tackling illegal content online[131]) or the Good Samaritan principle (Art.6 DSA) the EU legislator aims to incentivize platform providers to voluntarily adopt 'own-initiative investigations' to ensure flexibility and room for innovation. To adequately address the development within the EU legislation, the technical approach provided hereunder is based on the research of Elkin-Koren, who introduced so-called 'contesting algorithms':

Within her approach, the researcher suggests a double-check of the dominant platform removal systems through additional, adversarial public systems ('Public AI'). Designed to independently reflect underrepresented social values in platforms' own systems, such as fair use or free speech, Public AI shall run any platform removal decision prior to its final removal. If the Public AI confirms the platform's decision, the removal can be proceeded. If a dispute arises from discrepancies between the scoring assigned by the platform and the Public AI, the removal would be postponed until the conflict is resolved. Any data generated, including content removed or emerging discrepancies, should be gathered for analytical purposes and the system should be updated after any resolution.[132]

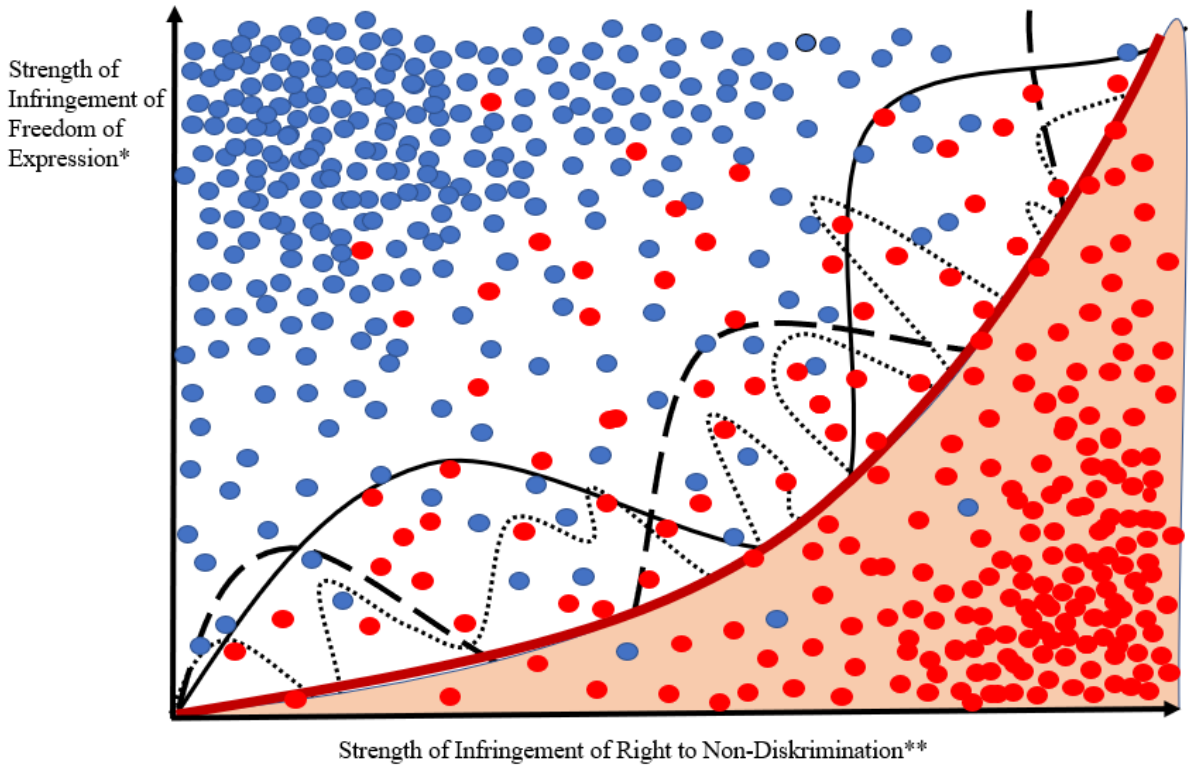
[130] See Chapter 3.2.1.5.

[131] COMM (2017) 555 final.

[132] Elkin-Koren, 'Contesting algorithms', 9.

Elkin-Koren's approach of Public AI aims to fill the gap between platforms' algorithmic content moderation regimes that optimize one single outcome or reflect a predetermined trade-off, and the need for deliberation, negotiation and contestation that are essential for governing speech in liberal democracies.[133] Her idea of introducing an adversarial framework to counterbalance the single optimization standard of current content removal systems could be extended to further legal assessment, e.g. to the currently proposed EU legislation.

Figure 7. Contesting Algorithm runs any Platform Removal Decision prior to its final Decision and sets a Minimum Standard within the EU.[134]



[133] Elkin-Koren, 'Contesting algorithms', 8.

[134] Illustration by the author.

Transferred into the model case, the EU Commission could set a minimum common standard in the hate speech removal process by implementing an additional 'contesting algorithm', requiring every platform provider to run it over its individual filtering mechanism.

Anyways, with a consistent definition of 'hate speech' or 'hateful content' determined by the contesting algorithm, platforms could theoretically be urged to remove any content falling underneath the EU's decision boundary (red line), but due to the Good Samaritan principle could maintain any approach that goes beyond. In this regard it is important for the EU to keep the minimum standard on a relatively low level to neither urge platform providers to become proactive nor restricted them in their freedom to independently develop own initiatives for their content filtering processes. As this idea is in its early stage, any discussion of a practicability and detailed implementation will be left to another day.

In summary, the 'contesting algorithms' contain an adequate approach to balance the EU's interests of implementing a harmonized minimum standard in hate speech content moderation, but concurrently providing platform providers with the freedom of innovation. Although, addressing two different perspectives, one might consider a combination of quarantining and contesting algorithms in a common model, when each approach has proved its worth.

4.3 Crowdsourced Image Moderation

As content moderation is not there yet to entirely replace human moderators by AI tools and even in the future it remains questionable to what extend human involvement is required to appease moral concerns, a method to relieve the moderators is urgent. The approach provided by this study is based on the research of Dang, Riedl and Lease, who investigated how to reveal the minimum amount of information to a human reviewer such that an objectionable image can still be identified correctly.

Figure 8. Image showing various Levels of Obfuscation.[135]

[135] Dang, 'But who protects the moderators?', 2.

The researchers were seeking to preserve the accuracy of human moderation, while making the moderators' working conditions safer. By focusing on the violence detection in videos and images, they experimented with blurring entire pictures to different extents (Fig.8) such that low-level pixel details are eliminated but the image remains sufficiently recognizable to accurately moderate. [136]

In another experiment they implemented three interactive controls (slider, click and hover) for content moderators to partially reveal blurred regions (Fig.9) to help them success fully moderate images that have been too heavily blurred. In one interactive mode, the moderator can increase/decrease the level of blur via a 'slider' widget. Two other interactive modes allow him to reveal a small region of the original image, either temporarily by mouse over or permanently by mouse click.[137]

Figure 9. Interactive settings let Moderators unblur a small Region by Mouse Over (temporary) or Mouse Click (permanent). [138]

For evaluating their experiments, the researchers considered accuracy and time taken as arguably most critical for both stakeholders (moderators and platforms) in a successful commercial content moderation.[139] They found that static blurring leads to decreased moderator accuracy with increasing blur. In contrast, interactive blur interfaces can reduce emotional impact on moderation without sacrificing accuracy or speed. The simple knowledge one has control can reduce emotional labour, distinct from any benefit from actually exercising that control. [140] Although the three interactive interfaces (slider, click and hover) did not differ in accuracy and completion time, with the key goal of keeping accuracy and reducing emotional impact the

[136] Dang, 'But who protects the moderators?', 2.

[137] Dang, 'Fast, Accurate, and Healthier', 33-34.

[138] Ibid, 35.

[139] Ibid.,40.

[140] Ibid, 41.

researcher recommend hover as the best comfort and lowest negative experience and exhaustion.

As the visualization of violence through images and videos is considered a more harmful expression of hate speech for the human moderators to review, the control mechanism provided by the researcher promises a certain degree of protection and simple feasibility. If considered as contributing to the moderators' well-being, one might extend the interactive blur interfaces to text passages filtered as hate speech by allowing moderators to only reveal parts of it. With the long-term aim of reducing the total amount of human moderators employed, the implementation of an interactive control mechanism should become a short-term goal that is cost-efficient and easy to implement for any platform provider or third-party employing moderators.

4.4 Final Thoughts

Taken the three approaches as incentives for a new strategy, it is not the single platform provider that is addressed to implement one, two or all three of them as a one-size-fits-all solution. The strategy should be implemented in collaboration with all stakeholders involved in the content moderation process aiming to protect the users' rights in the European digital space and to improve the human moderator's working conditions. Whereas each platform could implement their individual quarantining algorithm, the EU could set a minimum standard through a contesting algorithm still enabling platforms' individual approaches. This contains the advantage of harmonized minimum filter requirements within the EU every platform provider must comply with, but otherwise does not patronise them to adopt a specific type of technology, which could especially challenge small- and middle-sized companies' budget and (human) resources. Despite that, each platform provider or any third-party employing human moderators should legally be obliged by a European framework or voluntarily implement interactive control mechanisms such as 'crowdsourced image moderation'. Although, this study is limited to the category of 'hate speech', those technical tools should be enhanced to other categories of unlawful content having a similar health-damaging effect on the human moderators, such as 'violative content', 'sexual child abuse' or 'terroristic content'. The approaches provided merely aim to set incentives for platform providers but should be extended or replaced by similar, but minimum same efficient technologies.

However, regarding future development of content filtering technology and the need for human moderators, one final thought should be shared: Although, the aim to reduce human reviewers' involvement due to their health-damaging working conditions seems comprehensible, platform

users become increasingly concerned about their rights in the digital space solely being protected by machines. In contrast to human beings guided by emotions and empathy, algorithmic decision-making is based on training patterns and the data that is fed to the machine learning model.[141] Although, they meanwhile significantly relieve human moderators with the 'easier' cases in hate speech removal processes, their decision contains a simple yes-or-no questions. Hence, when it comes to ambiguous content leaving room for interpretation, the algorithmic precision is insufficient and human judgement and morality is needed to fill the gap. Even researchers, such as Hasselberger criticize that human morality is not a well-defined domain with a finite list of identifiable features whose importance and relationships can be precisely analysed in advance. Every concrete moral situation is a new situation, and some moral features of situations cannot be coherently defined in terms of fixed and precise set of value-free physical data. [142]

Despite any technical enhancement, it today remains questionable to what degree human moderators must be kept in the content moderation process to ensure a certain degree of human morality. The more it becomes significant to protect the individual human reviewer and improve its working conditions by having in mind their great value for the society's well-being in the digital space. It should become the platform providers' job to increasingly place human reviewers in the background of content moderation processes to perform mere control function and not being confronted with harmful material in the first instance.

5 Conclusion

There is no doubt, that a new strategy for algorithmic and human content moderation in the category of 'hate speech' is urgently needed to strengthen the users' fundamental rights and to ensure a stronger protection for human moderators. The study's analysis has pointed out several challenges arising from a technical and legal perspective, addressing both platform providers and the EU legislator. To successfully master these challenges, the stakeholders have been provided with concrete recommendations as well as practical approaches.

Based on the model case, the technical analysis has recommended platform providers to widen the scope of their filtering algorithms (second approach) to ensure a fast removing of content

[141] See Chapter 2.2.

[142] Hasselberger, 'Why we can't (and shouldn't) replace human moral judgement with algorithms?', 987.

flagged as harmful. Considered as an advantage, this prevents recipients from viewing potential harmful content in the first instance, but also enables, after a second evaluation, the re-upload of content assessed as covered by the freedom of speech. Even human moderators would benefit from this solution in a twofold manner: Firstly, the broad filtering scope could reduce the total amount of hate speech related material for the moderator to review, and secondly, could mitigate the harm within the content to review when the user is required to proactively argue for its freedom of expression.

In contrast to the technical analysis primarily addressing the platform providers within the EU, the legal analysis recommends the EU legislator to follow the first approach of the model case and to not urge providers to raise their decision boundaries. As investigated, the E-Commerce Directive grants hosting services a liability exemption if they do not acquire actual knowledge of illegal activities. This is especially important to keep for the upcoming Digital Services Act, as it enables room for innovation and increases the EU's attractiveness for providers to offer their services within the EU. Nevertheless, the freedom leads to increasing concerns among the legislative and judicial power within the EU, as it prevents platform providers from becoming proactive in the content filtering process. To counteract the private companies' restraint, the upcoming DSA and recent case law have set first incentives: Whereas the CJEU widened the interpretational scope of Art.15 ECD by covering proactive detection of content that is 'identical or equivalent' to content that has previously been declared unlawful, the EU legislator firstly implemented a 'Good Samaritan principle' within the proposed DSA to enable own-initiative investigations of hosting services aimed at detecting, identifying and removing illegal content.

The technical and legal recommendations, although focusing on two different approaches of the model case, are anything but mutually exclusive: Whereas the upcoming DSA, despite its deficits, provides hosting services an appropriate balance of innovative freedom and incentives to adopt own-initiative investigations, the latter are free to voluntarily raise their decision boundaries and to adopt measures such as a 'quarantining' or 'contesting' algorithms.

Nevertheless, both stakeholders should be required to improve the human content moderators' working conditions, either within a legal framework or with technical initiatives, such as 'crowdsourced image moderation'.

Finally, the recommendations provided rather complement than exclude each other by contributing to a joint strategy for algorithmic and human content moderation in the category of 'hate speech' and should encourage platform providers and the EU legislator to build on in the future.

Table of reference

Books and Articles

Abbasi, Ahmed, Ammar Hassan and Milan Dhar. 'Benchmarking Twitter Sentiment Analysis Tools.' *Proceedings of the 9th Language Resources and Evaluation Conference*. (2014):1-6.

Astuti, Firmina and P. Partini. 'The Hate Speech Behavior of Teenagers on Social Media Instagram.' *International Summit on Science Technology and Humanity*. (2019), 257-258.

Barata, Joan. 'Positive Intent Protections: Incorporating a Good Samaritan principle in the EU Digital Services Act.' *Center for Democracy and Technology*. *(CDT*) (2020), 1-13.

Barata, Joan. 'The Digital Services Act and the Reproduction of Old Confusions: Obligations, Liabilities and Safeguards in Content Moderation.' *Verfassungsblog* (2021), https://verfassungsblog.de/dsa-confusions/, doi:10.17176/20210302-154101-0.

Bolukbasi, Tolga, et al. 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.' *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS),* (2016):5.

Brunk, Jens, Jana Mattern and Dennis M. Riehle. 'Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems.' *European Research Center for Information Systems*: 2019 IEEE 21st Conference on Business Informatics.

Caplan, Robyn. 'Content or Context Moderation?' *New York, NY: Data & Society Research Institute*. (2018): 1-14. https://datasociety.net/output/content-or-context-moderation/.

Center for Democracy & Technology. 'Nine Principles for Future EU Policymaking on Intermediary Liability.' (2019). Nine-Principles-for-Future-EU-Policymaking-on-Intermediary-Liability-Aug-2019.pdf (cdt.org).

Cordeiro, Jacob. 'Free Speech in the Internet Era: Reviewing Policies Seeking to Modify Section 230 of the Communications Decency Act of 1996.' *Senior Honors Projects*. Paper 903. (2021): 57-59.

Council of Europe. 'European Court of Human Rights, Fact sheet - Hate Speech.' (2012):1. https://www.refworld.org/docid/4f39419d2.html.

Council of Europe. 'European Court of Human Rights, Fact sheet - Hate Speech.' (2008):2.

Dang, Brandon, Anubrata Das and Matthew Lease. 'Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content.' *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 8, no.1. (2020): 33-41.

Dang, Brandom, Martin J.Riedl and Matthew Lease. 'But who protects the moderators? The case crowdsourced image moderation.' *arXiv preprint arXiv:1804.10999.* (2018):2.

Datar, Mayur, et. al. Mirrokni. 'Localitysensitive hashing scheme based on p-stable distributions.' *Proceedings of the twentieth annual symposium on computational geometry* (2004): 253–262.

Deniz, Oscar, et al. 'Fast violence detection in video.' *In 2014 international conference on computer vision theory and applications (VISAPP)*, 2 (201): 478.

Duarte Natasha, Emma Llanso and Anna Loup. 'Mixed Messages? The Limits of Automated Social Media Content Analysis.' *Center for Democracy & Technology* (2019): 8-20. https://perma.cc/NC9B-HYKX.

Elkin-Koren, Niva. 'Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence.' *Big Data & Science,* 7, no. 2, (2020) 6-10, doi:10.1177/2053951720932296.

El Naqa, Issam and Martin Murphy. 'What is Machine Learning?' In *machine learning in radiation oncology*, (2015): 3. doi:10.1007/978-3-319-18305-3_1.

Evans, Richard and Edward Grefenstette. 'Learning Explanatory Rules from Noisy Data.' *Journal of Artificial Intelligence Research* 61 (2018) 1.

Frosio, Giancarlo and Christophe Geiger. 'Taking Fundamental Rights Seriously in the Digital Service Act's Platform Liability Regime.' (2020):4.

Gonçalves, João, et al. 'Common Sense or Censorship: How Algorithmic Moderators and Message Type Influence Perceptions of Online Content Deletion.' *New Media & Society*, (2021):2. doi:10.1177/14614448211032310.

Gorwa, Robert, Reuben Binns and Christian Katzenbach. 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance.' *Big Data & Society*, 7, no.1. doi:10.1177/2053951719897945.

Grimmelmann, James. 'The Virtues of Moderation.' *Yale Journal of Law and Technology*, 17, no.1, Art.2. (2015):47.

Hasselberger, William. 'Ethics beyond Computation: Why We Can't (and Shouldn't) Replace Human Moral Judgement with Algorithms.' *Social Research: An International Quarterly* 86, no. 4 (2019): 977-999. muse.jhu.edu/article/748873.

Havely, Alon et al. 'Preventing Integrity in Online Social Networks.' *arXiv:2009.10311*. (2020):25.

Hirschberg, Julia and Christopher D. Manning. 'Advances in Natural Language Processing.' *Science* 261 (2015): 349.

Hoboken, Joris et.al. 'Hosting intermediary services and illegal content online. An analysis of the scope of article 14 ECD on light of developments in the online service landscape: final report - Study. (2019): 10.2759/284542.

Kuczerawy, Aleksandra. 'General Monitoring Obligations: a new cornerstone of Internet regulation in the EU?' Rethinking IT and IP Law – Celebrating 30 Years CiTiP. (2019): 1-6.

Langvardt, Kyle. 'Regulating Online Content Moderation.' The Georgetown Law Journal 106:1353. (2017): 1368.

Kuczerawy, Aleksandra. 'The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act.' *VerfBlog*, 2021/1/12, https://verfassungsblog.de/good-samaritan-dsa/, doi: 10.17176/20210112-181758-0.

Maind, Sonali B. and Priyanka Wankar. 'Research Paper on Basic of Artificial Neural Network.' *International Journal on Recent and Innovation Trends in Computing and Communication* 2, no.1. (2014): 96.

Merton, Robert, John Wilkinson and Jacques Ellul. 'The Technological Society.' (1964): vi.

Mueller, John Paul and Luca Massaron. 'Introducing Deep Learning.' in *Deep Learning for Dummies,* ed. John Paul Mueller and Luca Massaron. (New Jersey, 2019) 9-24.

Niu, Xia-mi and Yu-hua Jiao. 'An overview of perceptual hashing.' *Acta Electronica Sinica*, 36(7):426-427.

Organisation for Economic Co-operation and Development (OPEC), 'The economic and social role of internet intermediaries.' (2010):6.

Peukert, Alexander. 'Five reasons to be skeptical about the DSA.' *VerfBlog*, 2021/8/31, https://verfassungsblog.de/power-dsa-dma-04/, doi: 10.17176/20210831-233126-0.

Roberts, Sarah T., 'Commercial Content Moderation: Digital Laborers' Dirty Work.' *Media Studies Publications*. 12 (2016):2.

Ross, Björm, et al. 'Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis.' *Proceedings of the Third Workshop on Natural Language Processing for Computer- Mediated Communication* (2016): 6-7.

Sae-Bae, Napa et. al. 'Towards Automatic Detection of Child Pornography.' *IEEE International Conference in Image Processing (ICIP)*(2014): 5332.

Sartor, Giovanni, 'Providers Liability: From the eCommerce Directive to the future.' In-depth analysis for the *IMCO Committee commissioned by the Policy Department for economic and scientific policy, Doctorate-General for Internal Policies, European Parliament* (2017), 27. https://www.europarl.europa.eu/Reg-Data/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf

Schulz, Wolfgang. 'Regulating Intermediaries to Protect Privacy Online – the Case of the German NetzDG.' *Personality and Data Protection Rights on the Internet, Forthcoming* (2018):9-12.

Schwemer, Sebastian Felix, Tobias Mahler and Håkon Styri. 'Legal analysis of the intermediary service providers of non-hosting nature'. Final report prepared for European Commission. (2020), 32.

Senftleben, Martin and 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.' (2021): 6-15.

Steiger, Miriah et.al. 'The Psychological Well-Being of Content Moderators. The Emotional Labor of Commercial Moderation and Avenues for Improving Support.' *CHI Conference on Human Factors in Computing Systems (CHI '21*), (2021): 1. doi:10.1145/3411764.

Thompson, Nicolas. 'Instagram Unleashes an AI System to Blast Away Nasty Comments.' *Wired* (2017).

Ullmann, Stefanie and Marcus Tomalin. 'Quarantining online hate speech: technical and ethical perspectives.' *Ethics and Technology* 22, (2020) 69-80.

van der Donk, Beriden B.E. 'How dynamic is a dynamic injunction? An analysis of the characteristics and the permissible scope of dynamic injunctions under European Law after CJEU C-18/18 (Glawischnig-Piesczek).' *Journal of Intellectual Property Law & Practice*. (2020): 608-611

Web pages

Adf international. 'Response to call for submission by the UN Special Rapporteur on the Protection of the Right to Freedom of Opinion and Expression'. Last modified 21 December 2016. ADF.docx (ohchr.org).

Allan, Richard. 'Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community'? last modified 27 June 2017. https://about.fb.com/news/2017/06/hard-questions-hate-speech/.

Amnesty International. 'Freedom of Expression.' https://www.amnesty.org/en/what-we-do/freedom-of-expression/.

Angwin, Julia, Larson, Jeff, Surya Mattu and Lauren Kirchner. 'Machine Bias.' last modified 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Baggs, Michael. 'Online hate speech rose 20% during pandamic: 'We've normalised it'. Last modified 15.11.2021. https://www.bbc.com/news/newsbeat-59292509.

Banjeree, Prasid. 'Facebook's content moderators demand for an end to culture of 'fear and secrecy'. Last modified 23 July 2021. https://www.livemint.com/technology/tech-news/facebooks-content-moderators-demand-for-an-end-to-culture-of-fear-and-secrecy-11626998967857.html.

Boring, Nicolas, 'France: Constitutional Court Strikes Down Key Provisions of Bill on Hate Speech, last modified 2020. https://www.loc.gov/item/global-legal-monitor/2020-06-29/france-constitutional-court-strikes-down-key-provisions-of-bill-on-hate-speech/.

Cambridge Dictionary. 'Hate Speech'. HATE SPEECH | meaning in the Cambridge English Dictionary. https://dictionary.cambridge.org/dictionary/english/hate-speech.

Commission to the European Parliament and the Council, 'Tool#18. The choice of regulatory instrument'. Better regulation toolbox. https://ec.europa.eu/info/sites/default/files/file_import/better-regulation-toolbox-18_en_0.pdf

Council of Europe. ‘Hate Speech and Violence’. https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence.

Culliford, Elizabeth and Brad Heath. ‘Facebook knew about, failed to police, abusive content globally-documents’. Last modified 26 October 2021. https://www.reuters.com/technology/facebook-knew-about-failed-police-abusive-content-globally-documents-2021-10-25/.

Daskal, Jennifer. ‘A European Court decision may usher in global censorship’ *Slate Magazine*. Last modified October 2019. https://slate.com/technology/2019/10/european-court-justice-glawischnig-piesczek-facebook-censorship.html.

Davis, Antigone and Guy Rosen. ‘Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer’. Facebook Newsroom. Last modified 1 August 2019. https://about.fb.com/ news/2019/08/open-source-photo-video-matching/.

Feldmann, Sarah. ‘How Does Facebool Moderate Content.’ last modified 13 March 2019. https://www.statista.com/chart/17302/facebook-content-moderator/.

IBM Cloud Education, ‘Machine Learning’. Last modified 15 July 2020. https://www.ibm.com/se-en/cloud/learn/machine-learning.

King, Jeff. ‘How We Review Content’. Last modified 11 August 2020. https://about.fb.com/news/2020/08/how-we-review-content/.

Laub, Zachary. ‘Hate Speech on Social Media: Global Comparison.’ last modified 7 June 2019, https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

Newton, Casey. ‘Facebook will pay $52 million in settlement with moderators who developed PTSD on the job.’ The Verge. Last modified 12. May 2020. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health.

Rosen, Guy. ‘Community Standards Enforcement Report, First Quarter 2021’. Last modified 19 May 2021. https://about.fb.com/news/2021/05/community-standards-enforcement-report-q1-2021/.

Sonderby, Chris. ‘Our Continuing Commitment to Transparency’, *VP & Deputy General Coun-sel*, last modified 19 May 2021. https://about.fb.com/news/2021/05/transparency-report-h2-2020/.

Srivas, Anuj. ‘RSS, West Bengal and Dublicate Accounts: What Facebook Whistleblower Compliant Touches Upon’. Last modified 5 October 2021. https://thewire.in/tech/facebook-whistleblower-frances-haugen-complaints-sec-hate-speech-misinformation-india.

Subedar, Anisa. ‘The country where Facebook posts whipped up hate’, last modified 12 September 2018. https://www.bbc.com/news/world-asia-46105934.

Tanz, Ophir. ‘Neural Networks made easy’. Last modified 13 April 2017. https://techcrunch.com/2017/04/13/neural-networks-made-easy/.

TechTarget Contributor. ‘Noisy Data’. Last modified May 2010. https://searchbusinessanalytics.techtarget.com/definition/noisy-data?.

Thomas, Zoe, ‘Facebook content moderators paid to work from home.’ last modified 18 March 2020. https://www.bbc.com/news/technology-51954968.

Wiggers, Kyle. ‘Facebook’s improved AI isn’t preventing harmful content from spreading.’, last modified 19 November 2020. https://venturebeat.com/2020/11/19/facebooks-improved-ai-isnt-preventing-harmful-content-from-spreading/.

EU Law

Council of Europe. European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos.11,14 and 15, supplemented by Protocols Nos.1, 4, 6, 7, 12, 13. November 1950.

European Commission. Proposal for a Regulation on a Single Market for Digital Services (Digital Services Act). COM (2020) 825 final.

European Parliament and the Council. Directive on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive). (2000). Directive 2000/31/EC.

European Parliament and the Council. Regulation on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM (2021) 206 final.

National Law

Austria. Communication Platform Act (Kommunikationsplattformen-Gesetz), 2021.

France. Loi no. 2020-766 du 24 juin 2020 visan à lutter contre les contenus haineux sur internet.

Germany. Network Enforcement Act (Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken Netzwerkwerkdurchsetzungsgesetz – Netz DG), 2017.

United States. Communications Decency Act of 1934, as amended by the Telecommunications Act of 1996.

United States. Digital Millennium Copyright Act, 1998.

Soft Law

Commission to the European Parliament. Code of Conduct on Countering Illegal Hate Speech Online, May 2016 with Facebook, Microsoft and Twitter.

Commission to the European Parliament, the Council, the European Economic and social committee of the regions. Communication on Tackling illegal content online towards an enhanced responsibility of online platforms. COM (2017) 555 final. 36

Commission to the European Parliament, the Council, the European Economic and social committee of the regions. Communication on Tackling online disinformation: a European Approach. COM/2018/236 final.

Commission to the European Parliament, Recommendation on measures to effectively tackle illegal content online. (C (2018) 334 final).

Case Law

Case C 18/18 Eva Glawischnig-Piesczek vs. Facebook Ireland Limited (2018) ECLI:EU:C:2019:458.

Case C-324/09 L'Oreal SA and Others v eBay International AG and Others (2011) ECLI:EU:C:2011:474.

Case C-484/14 McFadden (2016) ECLI:EU:C:2016:689

Case C-360/10 SABAM v Netlog NV (2012) ECLI:EU:C:2012:85.

Case C-70/10 Scarlet Extended SA v SABAM (2011) ECLI:EU:C:2011:771.