Universiteit Leiden
Governance and Global Affairs

How can regulation enhance transparency of AI facilitated content moderation

Reshma Pandohi Mishre
S2503387

Master Thesis, Executive Master in Cyber Security
Supervisors
Dr. T. Tropina
Dr. S. Picek

Leiden University
Faculty of Governance and Global Affairs
Cyber Security Academy

January 17, 2021

Abstract

Content moderation is about optimizing the equilibrium between two important values: freedom of speech and a safe and secure digital space. The main tasks are defining what is admissible content and assuring that inadmissible content is not allowed into the digital public space. Commercial digital platforms cannot be expected to carry this responsibility on their own without any incentives or obligations. They have their own commercial goals to serve. Tightened and more precise regulation is necessary. Overfitting the regulation will compromise freedom of speech. Underfitting the regulation will compromise the security of the digital space. An important aspect of assessing this balance is transparency.

In this thesis we looked at the historical timeline of drafted regulation and the rise of social media. The three layer-model of cyberspace was used to analyse AI facilitated content moderation. Transparency requirements on each level have been identified and existing and upcoming regulation on content moderation and AI has been assessed to identify gaps.

Current regulation on transparency in content moderation lacks clarity, enforcement, and consistency, partly because the E-commerce Directive was drafted before the explosive rise of social media and AI. It is remarkable, however, that the basic requirement for notice and takedown still serves a very relevant purpose.

An increased focus of regulation of the technical layer is required with the introduction of artificial intelligence tools in content moderation. Although regulation on artificial intelligence is fragmented and still in an early stage of development, the Digital Services Act and the EU White Paper on Artificial Intelligence include promising measures, such as record keeping and auditing. The overlap and mutual synergy between both regulations should be closely monitored.

The last conclusion is on transparency of terminology. Terminology regarding transparency in the world of AI technology, often relates to insight into the technical functioning of algorithms and to the ability to predict the outcome of an artificial intelligence model. In the governance world, transparency is linked to accountability and clarity. This gap between the world of artificial intelligence technology and the world of governance will need extra attention when drafting further regulation on AI. There is a need for common terminology.

Key words: Transparency, AI, content moderation, Digital Services Act, E-Commerce Directive

Acknowledgments

I would like to thank my first supervisor, Dr. Tatiana Tropina, for her encouragement and optimism during the writing process and my second supervisor, Dr. Stjepan Picek for his guidance in the world of AI and Python. Thanks to them, I managed to keep my motivation, even during the uncertain times of a global pandemic. To my parents, my brother and sister, I say thanks for patiently listening to my never-ending discussions on this topic and many thanks to my business partner Guido Dam for his supporting words during our joint struggles with our theses.

And lastly, I would like to thank my darling daughter Saachi for her love and support, especially during all the hours when I was glued to my laptop screen.

Content

"By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it." - Eliezer Yudkowsky[1]

[1] Eliezer Yudkowsky Quotes (Author of Harry Potter and the Methods of Rationality) (goodreads.com)

1. Introduction

Many generations have pursued it, some accomplished it by building pyramids or by leaving behind great works of art or essential inventions, but now it is within reach of each person: a way to be immortal. All they have to do is post something, anything at all, on a social media platform. For many people posting content and receiving likes and reactions has become an essential part of their daily routine.

In the early days of the internet, the production of content was in the hands of webmasters who built websites with content for the users' benefit. The traffic direction of content was mostly from the websites to the users. This stage of the internet is known as Web 1.0.[2] The fast growth of social media brought a rapid rise in user-generated content with it, which marked the transition to Web 2.0, where content could as easily be uploaded as downloaded onto platforms to be shared with other users in the network.

This growing amount of user-generated content brought new challenges with it; for example, privacy issues when personal content like photos and films are shared without enough awareness of privacy risks. Another challenge is the issue of dealing with content that is considered harmful.

Harmful content comes in many forms. An obvious example is terrorist content. Propaganda videos aiming to influence people to join ISIS or videos of beheadings of captured adversaries and journalists have been viewed by many people through media like Facebook.[3] Content with the purpose of disinformation has taken a flight in the past years and is very prominently present on social media nowadays, such as several conspiracy theories linking the Coronavirus to 5G networks.[4] One of the most notorious examples of fake news probably is the QAnon conspiracy theory, which, in combination with misleading messages on social media by Donald Trump, led to the occupation of the US Capitol on the 6th of January 2021.[5]

Another type of harmful content is content concerning child abuse. This type of content is illegal, entailing criminal liability. Sharing is done through private networks, sometimes using private groups on social media.

It does not take much imagination to know that these examples of harmful content can profoundly affect social media users and society as a whole. With the increasing amount of harmful content, the efforts to protect social media users increased as well. Over the years, content moderation in some form has become an indispensable necessity.

---

[2] B Cormode, G; Krishnamurthy, 'View of Key Differences between Web 1.0 and Web 2.0 | First Monday', *First Monday*, 13.6 (2008) <https://firstmonday.org/ojs/index.php/fm/article/view/2125/1972> [accessed 20 November 2020].
[3] Gabriel Weimann and others, 'Democratising Online Content Moderation: A Constitutional Framework', *Perspectives on Terrorism*, 6.1 (2020), 53–64 <https://doi.org/10.1177/2053951719897945>.
[4] Axel Bruns, Stephen Harrington, and Edward Hurcombe, '"Corona? 5G? Or Both?": The Dynamics of COVID-19/5G Conspiracy Theories on Facebook', 177.1, 12–29 <https://doi.org/10.1177/1329878X20946113>.
[5] *https://edition.cnn.com/2021/01/07/us/insurrection-capitol-extremist-groups-invs/index.html*

In the past, content moderation was performed by human labour. In some cases, a (social media) company would have employees of its own for this and in other cases, a commercial content moderation company (CCM) would be hired.

Using human labour for content moderation has downsides. The amount of content that is uploaded is enormous. In 2019, 300 hours of video were uploaded to Youtube and 510.000 posts to Facebook every minute[6] and these numbers keep growing. This high rate of content creation makes it impossible to have all the content checked and handled by human teams.

A pressing issue in content moderation is the harm related to the job: content moderators have to look through flagged content to decide whether it can be allowed or not. Some of the content can be of a violent or perverse nature, up to alarmingly disturbing levels —many employees who have viewed such content end up having psychological issues.[7]

Over recent years AI systems have been applied in many fields where humans cannot perform because of a shortage in numbers of humans available[8][9] or because the job requires more intellectual capabilities than humans can deliver.[10] AI is used in fields like voice and face recognition and decision-making processes in the medical and financial sector, like fraud detection and customer services (for example, approving loans). In social media, AI plays an essential role in creating the feed of content and ads for people based on their likes and dislikes.[11]

The use of AI in content moderation can be a solution to some issues in this field. AI is capable of dealing with large amounts of data and AI systems, although being able to mimic emotions cannot become subject to mental issues due to traumatic events. However, the use of AI comes with challenges. AI systems can be a black box where decisions are hard or almost impossible to trace back, which is why transparency is named a notorious problem in AI.[12]

In this thesis, we will research how to improve transparency of AI for the scope of content moderation. After diving deeper into the transparency-issues of AI in content moderation, the problem definition will be formulated in the next chapter.

[6] *How Much Data is Created on the Internet Each Day? | Micro Focus Blog*
[7] Sarah T Roberts, 'Content Moderation', 2017 <https://escholarship.org/uc/item/7371c1hf> [accessed 20 November 2020].
[8] Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale', *Big Data and Society*, 7.2 (2020) <https://doi.org/10.1177/2053951720943234>.
[9] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance', *Philosophy and Technology*, 2020 <https://doi.org/10.1007/s13347-020-00429-0>.
[10] Marcel Salathé, Thomas Wiegand, and Markus Wenzel, 'Focus Group on Artificial Intelligence for Health', *ArXiv*, 2018.
[11] Cobbe.
[12] Stefan Larsson and Fredrik Heintz, 'Transparency in Artificial Intelligence', *Internet Policy Review*, 9.2 (2020), 1–16 <https://doi.org/10.14763/2020.2.1469>.

2. Exploring the transparency problem in AI facilitated content moderation

In this chapter, we will have a closer look at the societal relevance of content moderation and the important role of transparency in it. From there a connection with transparency of artificial intelligence tools will be made. Bringing these together will lead us to the research question.

2.1 Societal relevance

Societal relevance of transparency of content moderation

For many people, social media have become one of the most important tools for information sharing, newsgathering, and opinion forming. A safe and secure digital space without harmful and illegal content such as hate speech, display of violence, or child abuse materials is essential.

The multitude of companies from 20 years ago has made place for fewer large digital platforms that are now called tech-giants. This "platformization" according to van Dijck[13], puts a significant amount of economic and societal power and influence in the hands of these leading technological platforms. Many examples are showing the effects of these online media.

During the elections of 2016 in the United States of America, there was a massive spread of fake news and hate speech about the candidates on social media, however much higher for the republican candidate than for the democratic candidate. The nearly unimpeded propagation of these posts on social media impacted voters' opinions and choices to the level that, according to some major news media, it resulted in Donald Trump winning the elections.[14, 15]

The next US elections of 2020 put content moderation in the spotlight again. This time social media platforms were better prepared for their task of minimising the spread of misinformation. Platforms like Twitter and Facebook engaged in actively moderating content, including messages posted by the president of the US himself. According to the New York Times, half of President Trump's posts were flagged for including disputed or misleading information.[16] After the Capitol storming in the US on January the 6th, the inciting effect of some social media posts was considered so high that some of the larger social media platforms decided to take more drastic measures. Twitter and Facebook were the first to suspend Donald Trump's account because their policy had repeatedly been violated and the risk for harm still existed.[17]

---

[13] José van Dijck, 'Governing Digital Societies: Private Platforms, Public Values', *Computer Law and Security Review*, 36.xxxx (2020), 10–13 <https://doi.org/10.1016/j.clsr.2019.105377>.
[14] https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies
[15] Hunt Allcott and Matthew Gentzkow, 'Social Media and Fake News in the 2016 Election' <https://doi.org/10.1257/jep.31.2.211>.
[16] 'How Twitter Policed Trump During the 2020 US Election - The New York Times' <https://www.nytimes.com/2020/11/06/technology/trump-twitter-labels-election.html> [accessed 13 November 2020].
[17] 'Twitter *Permanently* Bans Trump After Years Of Aggressive, Violent Rhetoric On Platform | HuffPost' <https://www.huffpost.com/entry/trump-twitter-ban_n_5ff733bbc5b61a92a8c08b8c> [accessed 10 January 2021].

Closer to home in 2020, the famous Dutch rapper Lange Frans, a firm supporter of the well-known QAnon conspiracy theories, actively spread fake news through his own Youtube channel. The decision by Youtube to shut down the channel caused a small uproar in Dutch society.[18] This QAnon theory induced many fake news-items denying the existence of the Coronavirus, posing a real threat to the health of many people.[19]

The large digital platforms do not have journalistic objectives. The mission of Facebook, for example, is "to give people the power to share and make the world more open and connected."[20] Facebooks revenue comes from advertisements targeting specific user groups with particular characteristics. The more users they have, the better.[21] The content on these platforms has moved from being socially relevant, as intended at the birth of these platforms, towards being societally relevant. It affects people's health and lives. Balancing the task of content moderation with maximising user engagement and user satisfaction has to be a struggle for these companies.

There are many questions related to this balance that come to mind. How can we be sure that these companies do what is needed in the field of content moderation to prevent harm and that their policies prevent over-moderation and infringements on the right to free speech? What are the exact criteria for human moderators? Could there be hidden agendas or a level of arbitrariness? Having this many questions means that more transparency is needed. Fair content moderation without hidden agendas is required. Transparency on this matter is essential: transparency on the policy, transparency on how the policy is implemented, and transparency on what standards these platforms set for themselves.

How does AI impact transparency in content moderation

Without diving too deep into the nature and different forms of AI (which will be done in the following chapter), for now we will focus on the role of artificial intelligence in content moderation. As mentioned before, decision-making processes in many fields have incorporated the use of AI. In the financial sector, activities like fraud detection and customer services (for example, approving loans) use AI functionalities.[22] In healthcare, AI can assist as a diagnostic tool or as a tool to enhance images taken from scans.[23] On social media platforms, AI systems are used to create the feed of content and ads based on the likes and dislikes of the user and AI also has a role in flagging and processing harmful content.[24]

---

[18] *https://www.nrc.nl/nieuws/2020/10/21/youtube-verwijdert-kanaal-van-lange-frans-a4016785*
[19] https://www.bbc.com/news/blogs-trending-53997203
[20] 'Facebook Mission Statement 2020 | Facebook Mission & Vision Analysis' <https://mission-statement.com/facebook/> [accessed 12 November 2020].
[21] *https://www.businessinsider.com/how-facebook-makes-money*
[22] Jake Silberg and James Manyika, 'Tackling Bias in Artificial Intelligence (and in Humans) McKinsey', *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*, 2019, 1–8 <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>.
[23] Salathé, Wiegand, and Wenzel.
[24] Cobbe.

Research has shown that AI systems can display bias.[25, 26] There are many examples in healthcare,[27] facial recognition,[28] or in predicting criminal activity.[29] Understanding the way AI is built and how it functions is of importance to understand the widespread effects of AI decisions on society and more particular, on content moderation. Development of the algorithm and the used data sets for learning are in the hands of commercial parties who rely on these algorithms for competitive advantage and these are therefore kept secret.[30] Therefore adding AI to the world of content moderation increases the need for transparency.

In 2019 the EU prepared a framework to advance the development of trustworthy AI.[31] This framework describes the ethical principles- respect for human autonomy, prevention of harm, fairness, and explicability- necessary and indispensable in the development, deployment, and use of AI. This last principle - the principle of explicability- means "that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Even though these principles currently are not binding, it is vital to understand the implications of these principles and requirements since content moderation processes involving EU citizens and using AI should be designed according to these requirements from the EU framework.

The need for transparency of AI has also been put forward by human rights organisations, such as the Freedom Online Coalition.[32] Further recognising the importance of transparency in AI, in February 2020, the EU followed up with a white paper on AI.[33] According to this white paper, transparency of AI is a prerequisite for successful uptake by society. In the context of content moderation, the importance of transparency lays the foundation for additional and more precise regulation. It still might be a challenge to translate the transparency requirement required by society to the world of AI.

2.2 Problem definition and research question

Social media platforms can have a profound impact on society and human lives on various levels. Content moderation of user-generated content is needed to provide a secure digital social space, but it

[25] Silberg and Manyika.
[26] Songül Tolan and others, 'Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia' <https://doi.org/10.1145/3322640.3326705>.
[27] Xavier Ferrer and others, *Bias and Discrimination in AI: A Cross-Disciplinary Perspective*, 2020 <http://arxiv.org/abs/2008.07309> [accessed 16 November 2020].
[28] https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app
[29] https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses
[30] Kyle Langvardt, 'Regulating Online Content Moderation', *Georgetown Law Journal*, 106.4 (2018), 1353–88 <https://doi.org/10.2139/ssrn.3024739>.
[31] European Commission, 'Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future', *Ethics Guidelines for Trustworthy AI*, 2019 <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> [accessed 21 December 2020].
[32] Roger C. Schank, 'What Is AI Anyway?' <http://www.aistudy.com/paper/aaai_journal/AIMag08-04-004.pdf> [accessed 26 September 2020].
[33] EU, 'WHITE PAPER On Artificial Intelligence - A European Approach to Excellence and Trust)', *European Commission*, 2020 <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf> [accessed 11 December 2020].

requires a large amount of human effort and the processes are not transparent. Artificial intelligence tools can be used to ease this task. However, these tools bring along more opacity, increasing the need for transparency to assure that no arbitrary or arbitrariness and

Commercial companies primarily serve their commercial goals. Adhering to the content moderation policies will cost resources and commercial objectives probably will prevail when choices have to be made. For society to understand how platforms balance free speech and a safe digital environment, transparency accountability is essential.

Making sure that content moderation is performed in a way that respects human rights should not be put as an exclusive responsibility on the shoulders of these platforms. Regulators have an essential role to play.

In this thesis, we will have a more in-depth look into different facets of this issue. The goal is to find out what is needed to improve transparency of AI facilitated content moderation and how to achieve this higher level of transparency with the help of regulation. This leads to the following research question:

- How can regulation enhance transparency of AI facilitated content moderation?

This research search question can be dissected into the following sub-questions:

1. What are the current issues with transparency in content moderation?
2. What type of transparency is required?
3. What measures are currently used to ensure transparency of AI facilitated content moderation?
4. Which measures are further necessary to improve transparency of AI facilitated content moderation?

This subdivision will help to understand the structure of the paper as will be made clear in the following paragraphs.

Goal

In this research, we aim to bring the world of governance (EU regulation) and artificial technology closer together. There will be focus on transparency requirements in policies, literature, and models from a governance perspective. From the standpoint of technology and the use of technology, there will be focus on understanding the technology and focussing on what aspects of transparency are possible and in what terms aspects of transparency are described. By connecting these two sides, gaps, and possible solutions to the gaps will be derived.

Approach

The theoretical framework in paragraph 3.1 and 3.2, presents a closer look into content moderation and the notion of transparency of content moderation, based on existing literature and other documents, such as for example the EU White Paper on Artificial Intelligence.

After this, in paragraph 3.3, the nature and definition of AI will further be examined. Since AI is a container for many types of technology, the most relevant of these types will be discussed. Following this, in chapter 3.4, research on transparency in AI will be discussed to provide a clear understanding of the challenges in the underlying technology. The transparency aspects of AI will be addressed for different parts of the content moderation.

In chapter 4, a process-model for content moderation will be presented for the purpose of shared understanding. The framework for analysis will be discussed and some critical reflection on the approach will be given.

In the analysis phase, in chapter 5, several aspects that have been looked into will be brought together with the help of a conceptual model for cyberspace. Regulation of content moderation and regulation on AI will be assessed on transparency requirements for AI facilitated content moderation. Identified gaps and possible solutions will be discussed. The thesis will end with conclusions and recommendations for further research.

3. Theoretical Framework

This chapter will outline the theoretical framework that will be used. The theoretical framework will first focus on content moderation and then look at related work on transparency of content moderation. Following this, we will look into understanding AI and then discuss research on transparency of AI.

3.1 Content moderation

The technological possibilities of widespread information sharing and connecting by individuals are an accomplishment, bringing with it many new dimensions to how society functions. However, we also find many security-experts emphasizing the new types of risks that are introduced by these new platforms.[34, 35]

In this paragraph, we will briefly discuss the rise of these platforms and look deeper into the risks mentioned above. After that, developments in content moderation will be discussed.

Contextual background: the rise of social media

Since the beginning of this millennium, the number of people using social media has multiplied, mostly fuelled by young people seizing this new opportunity for self-realisation or building a validated self-image [36]. The internet transformed from the one-way-communication Web 1.0 to the two-way Web 2.0[37] and became more and more valuable to the social media platforms.

In 1997, Six Degrees was the first platform operating on a large scale, followed by Friendster, LinkedIn, and MySpace. Nowadays, Facebook is the most popular social networking platform, with more than 2 billion users worldwide.[38] The number two in this list is Instagram, with over half a billion users.[39]

Digital platforms these days facilitate that anything can be posted online and probably nobody should be surprised hearing that almost anything is. The numbers are dazzling: on a single day 350 million pictures are uploaded on Facebook [40] and every minute more than half a million comments are posted

[34] Gwenn Schurgin O'Keeffe and others, 'Clinical Report - The Impact of Social Media on Children, Adolescents, and Families', *Pediatrics*, 2011, 800–804 <https://doi.org/10.1542/peds.2011-0054>.
[35] Pieter Nooren and others, 'Should We Regulate Digital Platforms? A New Framework for Evaluating Policy Options', *Policy and Internet*, 10.3 (2018), 264–301 <https://doi.org/10.1002/poi3.177>.
[36] Sonia Livingstone and David R. Brake, 'On the Rapid Rise of Social Networking Sites: New Findings and Policy Implications', *Children and Society* (John Wiley & Sons, Ltd, 2010), 75–83 <https://doi.org/10.1111/j.1099-0860.2009.00243.x>.
[37] Cormode, G; Krishnamurthy.
[38] Zia Muhammad, 'A Timeline of Social Media (Infographic)', *Digital Information World*, 2019 <https://www.digitalinformationworld.com/2019/10/social-media-history-infographic.html> [accessed 28 December 2020].
[39] Muhammad.
[40] Jacquelyn Bulao, 'How Much Data Is Created Every Day in 2020?', *TechJury*, 2020

on Facebook [41]. With the growth of popularity of these platforms, the effort needed for content moderation has grown as well.

A complicating factor with content moderation is that moderation-rules help define the identity of the platforms and therefore serve as a marketing tool. An illustration is the recent migration of many Facebook-users to other platforms.[42] Facebooks stricter policy on flagging misinformation around the US presidential elections led to many users feeling that this was an infringement of their rights to freely express their opinions. They decided to switch to a different platform, with other guidelines that allow them more space to exchange ideas.[43] One of these alternative platforms growing in popularity is Parler, advertising with the promise of being a free space for the discussion of ideologies banned from other platforms like Twitter.[44] Using their content moderation policy as a beacon, they were able to gain users.

Content moderation: a closer look

Just as the platforms have undergone a profound transition in scale and use, the same has occurred to content moderation activity. A proper way to understand this change is by comparing an article published in 2015 by Grimmelmann[45] on content moderation and a dissertation published in 2017 by Roberts[46].

James Grimmelmann gives the following definition of content moderation:

> *"[T]he governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse.*

In his article, there is an emphasis on using content moderation as a tool to facilitate cooperation and community building. The agent could be a site administrator of groups of paid workers. The main concern seems to be keeping the platform or community a pleasant space. There is no mention of fake-news.

The dissertation by Sarah Roberts is published with research on the working conditions of content moderation agents. The following definition of content moderation is given:

<https://techjury.net/blog/how-much-data-is-created-every-day/> [accessed 20 November 2020].
[41] Bernard Marr, 'How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read', *Forbes*, 2018, 1–5 <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=3b0d112460ba> [accessed 20 November 2020].
[42] '"Stop the Steal" Supporters Shift from Facebook to Parler to Peddle False Election Claims | The Independent' <https://www.independent.co.uk/news/world/stop-the-steal-facebook-parler-election-b1720473.html> [accessed 28 December 2020].
[43] *Parler: the social network that's winning conservative recruits | Social media | The Guardian*
[44] *Parler: the social network that's winning conservative recruits | Social media | The Guardian*
[45] James Grimmelmann, 'The Virtues of Moderation', 42 (2015), 154–76 <https://doi.org/10.5810/kentucky/9780813169057.003.0008>.
[46] Roberts.

> *"Content moderation is the organized practice of screening user-generated content (UGC) posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction. The process can result in UGC being removed by a moderator, acting as an agent of the platform or site in question."*

Topics like hate speech and images depicting cruelty are extensively discussed as well as the psychological impact on the agents. We see that the platforms' growth and development have called for a new level of effort required for content moderation.

The second definition seems to be more elaborate and fitting to the issues addressed in this thesis, such as viewing the content moderation process as an organized practice dealing with matters like appropriateness and locality. According to this definition, screening is performed not only for the site's sake, but different locations can also require different rules. For example, rules allowing nudity in the Netherlands would be different from those in Saudi Arabia. Jurisdiction is an important aspect too.

Content moderators sometimes are part of the social media platform; however, this activity is often outsourced to companies who specialize in this line of work. Outsourcing is often done in the Philippines or India, but also to companies in Germany and the US (Iowa). In outsourcing, the content moderators (also called the agents) must understand the language and the cultural peculiarities, which is undoubtedly necessary if one has to judge the implications of posts. There also is limited time for each decision, which results in a certain level of inconsistency in the output.[47]

After reading some of the accounts by people who have been doing this job, the one thing that stands out is that this job is very demanding on the psychological level. The worst in behaviour and intention is encountered: hate, criminality, violence, and traumatising images of the abuse of children and animals. The moderators work under a lot of time pressure, and they are often poorly paid; psychological support is required by many and sometimes offered by the companies.[48, 49]

Having their employees sign non-disclosure agreements on the policies and rules used for decision-making is common among companies. Not only working conditions but also decision-making in content moderation remain shielded for the outside world.

The use of automation technology in content moderation

The primary driver to start using automation technology in content moderation has been scale. The amount of content that has to be reviewed has grown exponentially, the required workforce would be so large that this would no longer be a realistic possibility.[50]

In earlier days, when technology was first being used, algorithms were applied to match a text with records on a blacklist to determine whether the text was allowed or not.[51] For this purpose, blacklists

[47] Sarah T Roberts Dissertation, *BEHIND THE SCREEN: THE HIDDEN DIGITAL LABOR OF COMMERCIAL CONTENT MODERATION*.
[48] Roberts Dissertation.
[49] Langvardt.
[50] Gillespie.
[51] Emma Llansó and others, 'Artificial Intelligence, Content Moderation, and Freedom of Expression', 2020, 1–30

with language that is not permitted are maintained on websites [52] [53]. Another well-embedded use of technology in the years afterward is hashing. Imagery that is not allowed (Terrorist promotional content or depiction of abuse, for example) is hashed, and in the screening process, these hashes are compared to newly uploaded content to decide whether the content should be allowed or not.

In recent years more intelligent forms of automation have been used in content moderation, such as machine learning and deep learning. In paragraph 3.3, we will have a more detailed look into these technologies and their impact on transparency.

3.2 Content moderation and transparency

Content moderation has been called a black box by many researchers.[54] There is interaction between humans and machines, where humans are led by policy and instructions that are considered proprietary, and machines have been programmed and trained in a way and with materials that are also kept behind closed doors—there are many transparency issues. For individual users, it is not always clear what content-item precisely was the reason for moderation or what rule was violated. Consistency in applying the rules across countries is also named as an issue, as well as the application of the rules with enough knowledge of the context. [55]

The level of transparency that is provided by digital platforms is mostly on the level of individual cases triggered by user-disputes. These are answered (in the optimistic scenario) with an explanation of which rule was violated and why the content was flagged or deleted. Another level of transparency that is provided by some companies is on the aggregated level, in reports on the results of their content moderation actions[56].

What lacks is transparency on a systems level[57], so that there can be a better understanding of the impact of commercial content moderation on millions of users, allowing an informed discussion on how to regulate social media. Commercial companies, how good their intentions may be, cannot be expected to prioritize transparency. Regulation will be needed to ensure this transparency.

---

[52] Hate-speech-and-offensive-language/refined_ngram_dict.csv at master · T-davidson/hate-speech-and-offensive-language, 'Hate-Speech Lexicon', *Github* <https://github.com/t-davidson/hate-speech-and-offensive-language/blob/master/lexicons/refined_ngram_dict.csv> [accessed 22 November 2020].
[53] Github LDNOOBW, 'List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words' <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words> [accessed 22 November 2020].
[54] Shagun Jhaver, Amy Bruckman, and Eric Gilbert, 'Does Transparency in Moderation Really Matter?: User Behaviour after Content Removal Explanations on Reddit', *Proceedings of the ACM on Human-Computer Interaction*, 3.CSCW (2019) <https://doi.org/10.1145/3359252>.
[55] Nicolas P. Suzor and others, 'What Do We Mean When We Talk about Transparency? Toward Meaningful Transparency in Commercial Content Moderation', *International Journal of Communication*, 13 (2019), 1526–43.
[56] Facebook, 'Community Standards Enforcement', *Facebook Transparency*, October 2018, 2019 <https://transparency.facebook.com/community-standards-enforcement> [accessed 23 November 2020].
[57] Suzor and others.

3.3 What is AI?

A brief leap back in history

In 1950, Alan Turing wrote an article about whether machines were able to think[58]. In this article, he describes the Turing test, an imitation game, which could decide whether the behaviour displayed by a computer could be qualified as "intelligent." Turing predicts that it will be normal by the end of the century to say that a computer is intelligent. In this article, he also proposes that in the steps towards creating an intelligent computer instead of trying to develop the equivalent of an adult human mind we could start by creating the equivalent of a baby's brain which he calls the initial state of the mind. Afterward, an education phase would have to follow where the mind is developed and trained and there may be other sources of influence than education.

Around that time several scientists were investigating the matter intelligence in computers that were to be developed, although some were reluctant to use phrases like "thinking computers " and "intelligent machines" since the terms "intelligence" and "thinking" were supposed to be reserved to human-activity or conscious beings [59].

One of the scientists who was inspired by the idea of creating an intelligent machine was John McCarthy. He is also seen as the founding father of AI since the Dartmouth summer research project on Artificial Intelligence, where a significant advancement on AI was made by a selection of scientists. This event in 1956 is seen as a milestone - some call it the birth - of artificial intelligence [60, 61].

What is artificial intelligence?

John McCarthy applies the following definition of intelligence: "Intelligence is the computational part of the ability to achieve goals in the world." According to him, artificial intelligence (AI) is "the science and engineering of making intelligent machines, especially intelligent computer programs".[62] To understand the general idea underlying the development of AI, we reach back to the idea that was posed by Turing, namely creating a baby's brain and allowing it to develop by learning. Looking at the way AI is developed nowadays, we can say that Turing's idea still forms the essence of how Artificial Intelligence is being developed. The basis of AI is formed by an algorithm with the ability to learn. This is further developed in a learning phase using training data which then results in a model.

There are many well-known application of AI, such as robotics, natural language processing (NLP), facial recognition and voice recognition. A famous example in an AI application is Sophia, the robot who amazed the world during her presentation to the world in 2017 as a result of her life-like behaviour, facial expressions, and communication.[63] She (as the robot is commonly referred to) even

---

[58] A M Turing, *Computing Machinery and Intelligence*, *Computing Machinery and Intelligence. Mind*, 1950, XLIX.
[59] Peter Millican, 'The Philosophical Significance of the Turing Machine and the Turing Test', *Alan Turing: His Work and Impact*, 2013, 587–601.
[60] V Rajaraman, 'John McCarthy – Father of Artificial Intelligence', March, 2014, 198–207.
[61] Gil Press, 'A Very Short History Of Artificial Intelligence (AI)', *Forbes*, 2016 <https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/?sh=3b5954386fba> [accessed 19 November 2020].
[62] Rajaraman.
[63] *https://www.nationalgeographic.com/photography/proof/2018/05/sophia-robot-artificial-intelligence-science/*

pushed the discussion on the matter of intelligence in lifeless objects to a new level by having been granted citizenship to Saudi Arabia. Concern on this development among some scientists was expressed in an open letter to the European Commission on the topic of granting humanoids a separate status of “electronic persons".[64]

Underlying subfields are machine learning and deep learning.[65] Machine learning is applicable to problems like:

- classification: differentiating image of a cat from that of a dog.
- regression: predicting the next move of a system when you know its behaviour over a period of time.
- clustering: creating groups of similar items

Machine Learning

The basis of machine learning is a learning algorithm, which is developed using a training data-set in the learning phase. This data-set could for example be a set of (for each class) equal, large numbers of records with the classes (cat or dog) in the first field and features belonging to the class in the other fields. After processing this data the algorithm will have learned which set of features determines whether an object is a dog and which set determines whether the object is a cat. Testing with images that the algorithm has never processed before will produce an outcome with a certain error-margin. Training with a more elaborate data-set will improve accuracy.[66]

This type of learning is known as supervised learning[67]: the required output (cat or dog) is known beforehand as a function of the input variables (the features). The dataset for supervised learning is split into a training set, used to train the machine, and a test set, which is used to measure the accuracy of the system. You can compare this to a supervisor who knows it all, teaches the system what it should know, and afterward tests if the training was sufficient.

The dataset as mentioned here is called structured data[68]; it is structured in fields of which the meaning is known. This type of machine learning needs much preparation. For machine learning, data must be prepared in a structured way, including the feature extraction that will decide the outcome.

Another type of learning is unsupervised learning[69]. Here the outcome is not known beforehand. The system decides what categories (classes) will be created, and it does this by maximizing the similarities within a class and minimizing the similarities between the classes.

---

[64] ‘Robotics Openletter | Open Letter to the European Commission’ <http://www.robotics-openletter.eu/> [accessed 19 November 2020].
[65] Stefano A. Bini, ‘Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?’, *Journal of Arthroplasty*, 33.8 (2018), 2358–61 <https://doi.org/10.1016/j.arth.2018.02.067>.
[66] Bini.
[67] https://www.ibm.com/cloud/learn/machine-learning
[68] Bini.
[69] https://www.ibm.com/cloud/learn/machine-learning

The third type of learning is reinforcement learning[70]. The system has a choice of actions, and each action comes with a reward or punishment. By trying to maximize the reward, the system will "learn" to perform specific actions. An example of this is teaching a machine to play a game of checkers.

Machine learning is applied for classification problems, for example, distinguishing cats from dogs, and for regression, which is deciding what the next steps in a system's behavior would be.

Deep Learning

The next level in AI is deep learning, which is a subfield of machine learning. Deep Learning involves the use of an artificial neural network (ANN) which is built up of several nodes in layers, inspired by the way a biological neural network functions. The nodes are the connecting units where the computing is done.[71]

The main advantage of neural networks is that data does not have to be structured and prepared as was necessary for machine learning. It can be multidimensional and unstructured data can be fed to the system and the system will examine the data, find interdependencies, and by adding weights to these interdependencies it will structure the data by itself.

The network can be quite complex, many (hidden) layers deep, which is what the name of this technology refers to, see figure 1.

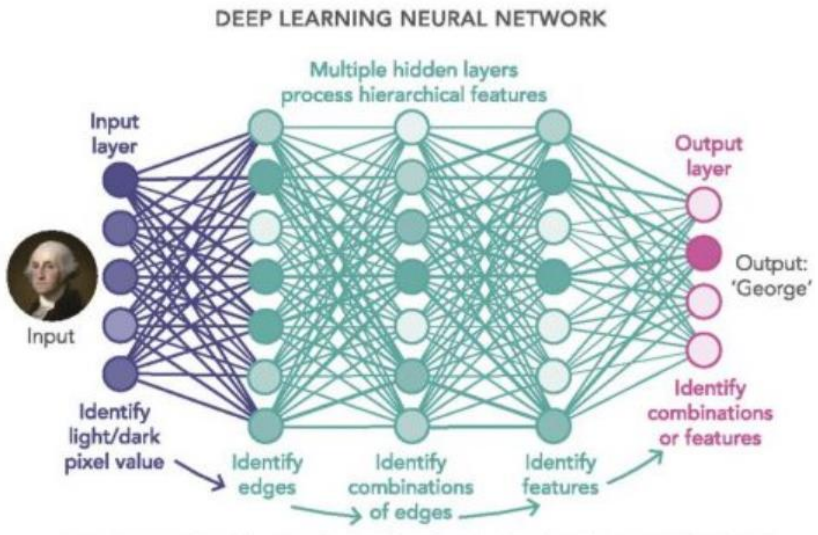


Figure 1: Depiction of a deep neural network [72]

Input data for neural networks can be unstructured data, for example, large pieces of text or big data gathered from different sources in different forms.[73] The complex network accommodates feature extraction and building the model in which the input is mapped onto several hidden layers that

[70] https://www.ibm.com/cloud/learn/machine-learning
[71] Bini.
[72]Copied from: What is Deep Learning, Nature of Machine Learning and Beauty of Deep Neural Networks ? - New World : Artificial Intelligence (newworldai.com)
[73] Bini.

represent abstract aspects of the input. For example, in pattern recognition, such abstract elements can be the difference between darker and lighter pixels to determine the contours of the object. The algorithm will decide what the most relevant aspects are that will be part of these hidden layers, based on their contribution towards determining the desired output.[74]

Algorithmic content moderation

Gorwa[75] defines algorithmic content moderation as "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geo-blocking, account takedown)" and discusses the following primary methodologies that are used for flagging and filtering in content moderation:

- Matching: this involves creating a hash (a calculated unique value matching the content) which is compared to existing hashes of illegal content in a database, to determine whether the content is admissible or not. An example is the Global Internet Forum to Counter Terrorism (GIFCT)[76, 77]. This is a database containing hashes of images used for terrorist propaganda, set up by some of the large cooperating technology companies such as Facebook, Twitter, and YouTube.

  Some alternative forms of hashing are used to prevent images that are slightly different (cropped flipped, slightly changed borders) to circumvent the matching. These are called "fuzzy hashing or perceptual hashing"[78]

- Classification: this involves the use of Machine learning to determine the category of a new content item, for example an uploaded picture.

- Regression: predicting whether some content item is likely to be illegal or not, for example hate speech. Developments in Natural Language Processing account for many advanced applications of AI in the fields of hate-speech-detection, spam, and bullying. [79, 80]

The use of these technologies does not stand in itself, some human interaction is part of this process to make the final decision whether the content needs to be removed. In the next paragraph, we will have a closer look at transparency aspects of these technologies.

---

[74] Bengio and COurville Goodfellow, *Deep Learning*, *The MIT Press* (The MIT Press, 2016).
[75] Robert Gorwa, Reuben Binns, and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data and Society*, 7.1 (2020) <https://doi.org/10.1177/2053951719897945>.
[76] GIFCT | Global Internet Forum to Counter Terrorism
[77] Brittan Heller, *Combating Terrorist-Related Content Through AI and Information Sharing*, 2019 <www.annenbergpublicpolicycenter.org/twg> [accessed 22 November 2020].
[78] Gorwa, Binns, and Katzenbach.
[79] Gorwa, Binns, and Katzenbach.
[80] Chikashi Nobata and others, 'Abusive Language Detection in Online User Content', in *25th International World Wide Web Conference, WWW 2016* (International World Wide Web Conferences Steering Committee, 2016), pp. 145–53 <https://doi.org/10.1145/2872427.2883062>.

3.4 Transparency of AI

Now that some basic concepts in AI have been discussed we can have a closer look at transparency in AI. Transparency will be researched by reviewing existing literature and documents, including for example the EU White Paper on Artificial Intelligence – A European approach to excellence and trust.

Transparency has a widespread use in a multitude of disciplines, ranging from a physical meaning like transparency of objects to transparency as a quality of fairness in systems to overcome information asymmetries.[81] Larsson[82] links the concept of openness ('open data', 'open source', 'open code' and 'open access') and explainability (xAI) to the concept of transparency.

There are many reasons why we need transparency of AI. For example for verification of the system - does it do what it is supposed to do? It is also needed if you want to improve the system. In the case of self-driving cars, for example, it is very important to know what factors led to certain decisions. Data on the performance of the system can be used to implement improvements. A third reason is accountability. When decisions have thorough implications or maybe even legal implications on human lives, explanations are needed to understand what factors led to the system taking certain decisions.

Transparency in AI does not come easily or naturally. The structure of multiple layers and the nested complexity of the nodes makes it hard to trace back how a certain output has been achieved. As strange as it may sound, additional technology is needed, including additional fields of research. Two main concepts are important in this area.[83] The first is explainability, which is more narrow and based on technology. The second concept is interpretability, which is concerned with explaining the behaviour of an AI system in terms that a user can understand.

Larsson advocates a multidisciplinary approach to transparency[84] since merely xAI would not suffice: "This xAI-notion of transparency is narrower and more algorithmic model-oriented than for example the necessary transparency (and "explicability") expressed by AI HLEG (2019) to achieve an ethically sound and trustworthy AI." In this article six additional aspects to complement transparency are discussed.[85] The four most important ones for the scope of this thesis are:

- "legal aspects of proprietorship": social media operate in a competitive market and openness of their systems may not be a desirable strategy.
- "data and algorithm user literacy": a certain level of literacy will be required from governance bodies as well.
- "the symbols and metaphors used for communication, that is, mathematically founded algorithms may be dependent on translations to human imaginaries and languages": the outcome of explainability models may perhaps not fulfil the need from the users and society.

[81] Jens Forssbaeck and Lars Oxelheim, *The Multi-Faceted Concept of Transparency The Multi-Faceted Concept of Transparency*, 2014.
[82] Larsson and Heintz.
[83] Amina Adadi and Mohammed Berrada, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*, 6 (2018), 52138–60 <https://doi.org/10.1109/ACCESS.2018.2870052>.
[84] Larsson and Heintz.
[85] Larsson and Heintz.

- "the obscuring effects of distributed, personalised outcomes that create challenges not the least for supervisory agencies with limited access and overview attempting to detect structural discrimination or other unfair outcomes": transparency requirements exist on the level of the system, as well as on the individual level.

Samek, Wiegand and Müller discuss methods for algorithmic transparency on a systems level.[86] The first method is called sensitivity analysis and it is based on knowing how sensitive an AI model is by understanding what the effect of a change in the input will be on the result. Applying this to the intelligent systems involved in content moderation could for example lead to understanding how these systems work and why a particular word in a post leads to flagging of the post.

The other method discussed in this paper involves decomposing and modelling the AI system. This technique is called Layer-Wise Relevance Propagation (LRP in short). In this model the contribution of each pixel to the end result is taken into account, thereby simplifying the exact computation that is performed in each node in the several layers.

[86] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Robert Müller, 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models', *ArXiv*, 2017.

4. Methodology

In this chapter, the methodology for analysis and discussion will be discussed. The scope for analysis will be presented in paragraph 4.1.

Before diving into analyses and discussions, it is essential to create a shared understanding of the field of content moderation. A visual presentation of a generic model for content moderation will be developed in paragraph 4.2.

The Framework for analysis will be covered in paragraph 4.3, describing the data sources that will be used and the methodology for analysis.

4.1 Scope

The scope for analysis in this thesis will be the European Union's regulations and policies related to content moderation. Content moderation of user-generated content by social media platforms is in scope. There will be little emphasis on the curation of news feed, although some of the transparency issues will apply to that field as well since it likewise involves the use of AI.

4.2 Common understanding: a generic model for content moderation

In the analysis phase, we will look at several existing measures across the full width of content moderation, aiming to better understand their effectiveness, scope, and depth. For this purpose, apart from having a definition describing the essence of content moderation, we need to model the several steps of content moderation. This model will serve as the basis for analysis.

The method used to develop this Model will be process-decomposition, starting with the trigger of content being created and ending with the activity of content being deleted or being viewed. In this method, the aim is to break down the (usually) complex process to achieve sub-steps that are overseeable and that lead to a proper understanding of interdependencies and complexity. The modelling of the content moderation process is based on insight gained from the literature review and it is completed with items that have not been explicitly named in literature, but after critically reviewing completeness of the Model, were found to be missing. Such steps are for example "content presentation", "content viewing" and the steps related to policy.

From what we already know we can identify the following process-steps involved in/related to content moderation:

i. Content creation - The first step is the creation of content by humans uploading texts, pictures or films onto the social media platforms, using the functionalities that are provided by these platforms.

ii. Content flagging by deployed AI - In the second step, the deep learning models can flag some content items which are found to be inappropriate or illegal. This is done using the trained, specific models applied by these platforms.

iii. Content assessment by humans or AI - Content items that have been flagged by AI will be assessed by human moderators for final judgment on allowing or deleting the content, or perhaps add a tag if the content is misleading in any way.

iv. Content handling - After judgement by the human moderator, content can be deleted, hidden, or de- flagged according to the policy of the social media platform.

v. Content presentation - The platform presents the content to the user.

vi. Content viewing - Strictly speaking, this aspect is not part of content moderation. It however is one of the results of content moderation, namely having the content viewed by the user.

vii. Content flagging by users – After being viewed, the content can be flagged by the users. Functionality is provided by social media platforms.

To use AI in content moderation, the underlying neural network-models need to be developed and adequately trained. Bearing the abovementioned purpose of this Model in mind, these preparations can be modelled as follows:

viii. Developing a starting algorithm or neural network for deep-learning

ix. Training this algorithm or neural network by applying training data-sets.

x. Provide feedback from human moderation to deep-learning Model to improve the performance of the models.

In the following diagram, the interrelated process-steps are brought together to visualise the content-moderation-process as performed on social media platforms.

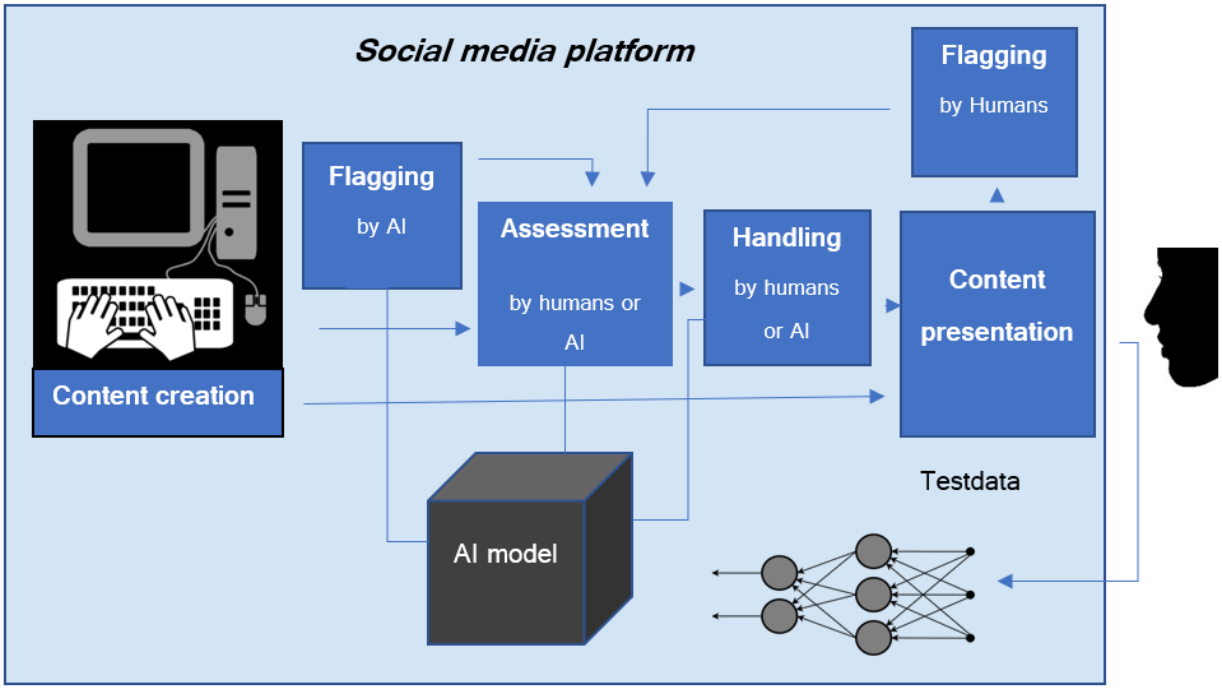


Figure 2: A visual representation of the content moderation-process

The goal of this model is to build a common understanding of content moderation for the scope of this thesis.

4.3 Framework for analysis

The framework for analysis will include:

- Discussion of the level and type of transparency that are needed
- Analysis of the level and type of transparency that is required by existing and upcoming regulation,
- Analysis of measures that have been (voluntarily) implemented by the private sector,
- Identification of the gap and discussion of additional measures to further improve transparency of content moderation, especially for social networks and other platforms that involve user-generated content.

Existing regulatory initiatives and current debate on this topic will be studied as well as transparency solutions that are developed in the field of artificial intelligence technology. The three-layer Model by van den Berg[87] will be used to analyse the state and type of (needed) regulation concerning content moderation. Both, endeavors to create transparency and transparency issues will be depicted on this

[87] Jan Van den Berg and others, 'On (the Emergence of) Cyber Security Science and Its Challenges for Cyber Security Education', *NATO STO/IST-122 Symposium in Tallin*, 2014, 1–10.

"canvas," as well as developments in the technical and socio-technical layer. From this overview, we will attempt to build up a closer understanding of what measures are required in the governance layer.

Analysis will be conducted in a few logical steps, each building upon the previous step towards the research question. Following are the steps:

Step 1: What sort of transparency is needed to enhance AI facilitated content moderation?

In the first step the basis for further analyses will be laid out: a mapping of the content moderation process on the three-layer Model by van den Berg.[88] This will help in pointing out the level and type of transparency that is needed.

Step 2: What measures are required by regulation and what measures are implemented?

In the second step, we will dive into the different regulations that apply to content moderation. Following this, various implemented measures will be analysed and these measures will be categorised by looking into the primary reason for implementing this measure.

Step 3: What regulation on AI is there?

The regulatory developments covering AI and transparency of AI will be explored in this step.

Step 4: what is the gap?

The implemented measures for transparency will be assessed against the initial purpose that they serve in terms of preventing harm and providing a safe and secure space for social media users. This analysis will allow us to point out the gap in this step.

Step 5: What are the relevant technological/ regulatory developments?

In the last step, technological and regulatory solutions that are being developed will be discussed and their value in addressing transparency-issues in content moderation. What are these solutions, how well would they aid in closing the gap? Are technological measures required? Or is it regulation that is needed? The expectation is that a combination of technological and regulatory measures will be needed to close the gap further and enhance transparency.

Data from the following sources will be collected to perform this analysis:

- Literature review of scientific articles
- Policy analysis
- Analysis of reports on private companies

[88] Van den Berg and others.

- Analysis of technological reports
- Existence of legally binding frameworks

4.4. Reflection on the research

The research focuses on governance measures., The research to set up the theoretical framework covers some topics in the area of artificial intelligence, but this is merely a scratch on the surface. In the preparation phase, we engaged in some small machine learning exercises to better understand what it all is about. These exercises provided practical understanding and connection to complicated technicalities of the subject. The time-frame of this research and my basic technical knowledge of the matter restricted me from further diving into experiments with neural networks and deep learning.

This research's validity is guarded by relying on findings from literature, policies, and technical reports from renowned institutions. To clarify societal relevance, other relevant documents, mostly from qualified journalistic sources, have been included. When needed reports from the platforms are used to demonstrate examples.

With regard to literature, that part of literature is included, that is “interpretable” to readers who have not fully mastered the technological and mathematical sides of AI. Therefore perhaps some technical depth is missing. Research documents were selected based on the prevalence of governance-related and non-technical points of view with a link to a certain level of understanding of the relevant technology.

The focus of transparency enhancing measures in this research will be on non-technical policy and regulatory options. Socio-technical options have not been taken into account because of the different underlying nature of mechanisms (for example, psychology).

5. Analysis

In this chapter, the analysis according to the steps explained paragraph 4.3 will be presented. Each step will be covered in a separate paragraph.

5.1. What kind of transparency is needed?

As mentioned in the framework for analysis, the three-layer model by van den Berg [89] will be used as a "canvas" to plot transparency issues and development. This model was developed to conceptualise cyberspace which is built up by the different layers that interact and complement each other. The centre of this model consists of a layer of technology and technical infrastructure, called the technical layer. This is enveloped by a layer of social interaction with the technology where benefits and harm are manifested, called the socio-technical layer. The third layer is the governance layer, consisting of regulation necessary to control the harm by influencing the other two layers' developments. Social media and content moderation consist of the technical infrastructure en social interactions. Governance (regulation and policies) aims to influence both layers in order to minimise risk and harm.

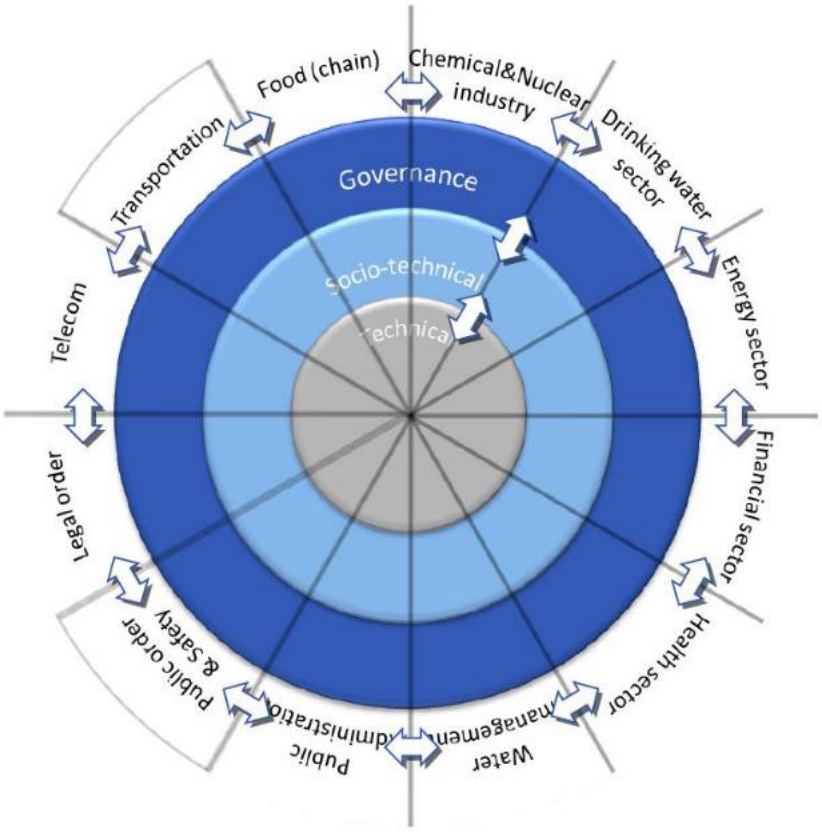


Figure 3: Conceptualization of cyberspace in layers and (cyber) sub-domains.

In figure 3 the three-layer Model is presented. The inner layer is the technology layer, consisting of the technology, technical infrastructure and all software. The socio-technical layer contains the

[89] Van den Berg and others.

interactions between technology and humans, which are called "cyber-activities". According to the three-layer model by van den Berg these *"include many types of (a) data & information exchange, (b) information search & retrieval, (c) e-watching & -listening, (d) IT-enabled transactions, (e) remote control (of e.g. complex ICSs), (f) cyber protesting & cybercrime, up to (e) cyber warfare"*. [90] And the final layer is the governance layer, containing all norms, rules and regulations. This model will help to structure content moderation and the nature of the requirements for transparency.

Content moderation in the three-layer model

The process steps identified in the generic Model for content moderation in chapter 4.2 now be linked to a layer in the three layer model. To start with the socio-technical layer, the following elements of content moderation that involve human interaction can be identified:

i. Content creation – humans create content by uploading text, pictures, or films onto social media platforms, using the functionalities provided by these platforms.

iii. Content assessment by humans of items that have been tagged by AI

iv. Content handling - deleting or hiding or de- flagging content after judgment by the content moderator according to the policy of the social media platform.

v. Content presentation - The platform presents the chosen content to the user.

vi. (Content viewing) - Strictly speaking, this aspect is not part of content moderation. It however is one of the results of content moderation, namely having the content viewed by the user.

vii. Content flagging by users – After being viewed, the content can be flagged by the users. Functionality is provided by social media platforms.

Some other aspects of content moderation are embedded in technology, so these are part of the technical layer, such as:

ii. Content flagging by deployed AI – deep learning models can filter and flag content items by classification or prediction of texts.

viii. Developing a starting algorithm or neural network for deep-learning

ix. Training this algorithm or neural network by applying training data-sets.

x. Using feedback from human moderation to deep-learning Model to improve the performance of the models.

Content moderation in current times has more and more become a process that is embedded in the technical layer with the use of algorithms for flagging and curation.

Using the three-layer model to analyse and plot the process steps makes it evident that the governance activities are missing. The governance layer was not addressed in the process decomposition, but it is relevant to the analysis in this thesis.

Adding the governance process steps is necessary to have a policy in place for content moderation. This

[90] Van den Berg and others.

adds the following elements:

xi. Content moderation policy development and implementation by the social media platforms

xii. Content moderation policy development and implementation by governments

xiii. AI policy development and implementation by the social media platforms, including transparency aspects

xiv. AI policy development and implementation by the governments, including transparency requirements

xv. Policy oversight activities

These elements, plotted in the three-layer model, are presented in the following figure:

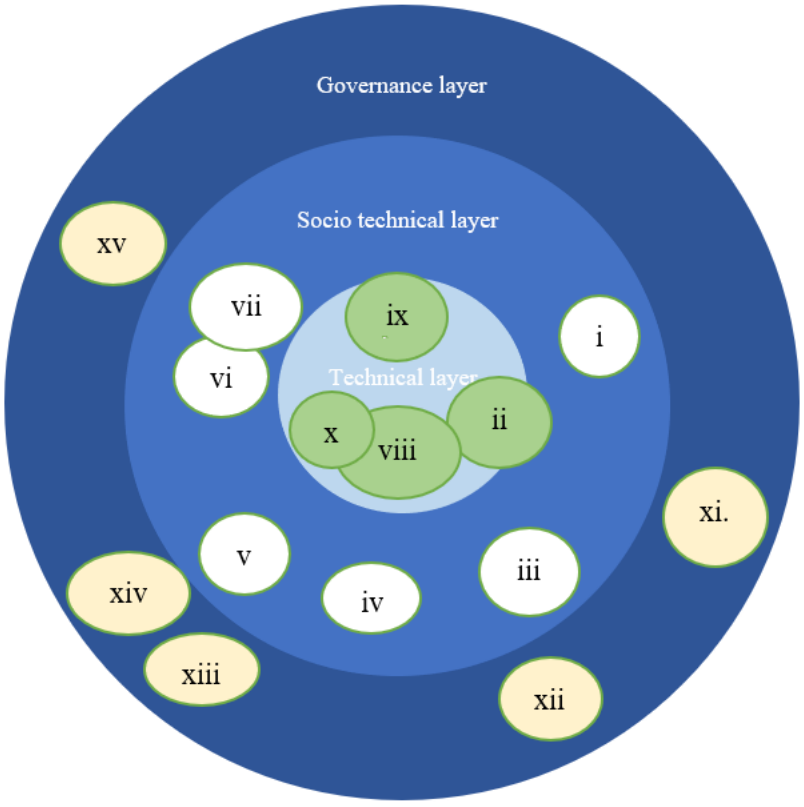


Figure 4: Content moderation activities plotted onto the three-layer-model of cyberspace

The spread of all activities shows that content moderation involves all three-layers of cyberspace. Content moderation in earlier days, before the application of AI, required a minimal amount of technology. Governance of content moderation was focussed on the activities in the socio-technical layer. Recent developments in the field of AI have placed key content moderation activities like filtering and flagging, in the heart of the technical layer.

Transparency in the three layers

Before we move to the next step of assessing whether the current regulatory standards are adequate to safeguard necessary transparency needs of AI facilitated content moderation, we have to understand what these needs are. These needs or the requirement(s) for transparency can be determined for each step in figure 4.

On the technical layer, the introduction of AI tools brings up a new dimension in opacity and therefore new transparency requirements on this layer are necessary, both to serve the needs on systems level as the needs on an individual level. On systems level, the use of explainability tools might be necessary, alongside transparency on the selection and development of algorithms and training of the models. For accountability towards individual users of the platform transparency on flagging and notifying and performance of artificial intelligence tools is required.

On the socio-technical layer, transparency is needed in the area of content moderation (assessment and handling of flagged items) by human moderators. As discussed before the performance of AI models can benefit from feedback from decisions taken by humans who are better at interpreting contextual information. Therefore transparency is also required in this interaction with artificial intelligence tools.

In the governance layer, the need increases for regulation by governments to secure transparency. This regulation need comprises both the policies on transparency in the use of AI in content moderation itself and the policies regarding transparency on strategies on AI and explainability of AI. These will help to understand to what purpose these tools will be applied and what their intended reach will be. In the policy-oversight measures, there will also have to be focus on transparency. The larger technology platforms put effort into developing these deep learning models. Regulation bodies will need to invest in AI knowledge and AI capabilities as well to be able to understand and assess these developments by for example performing in depth analyses of their AI landscape and implemented measures to safeguard free speech.

In figure 5 an overview of these transparency requirements is depicted. In the following step of analysis, these transparency requirements will be held against current regulation.

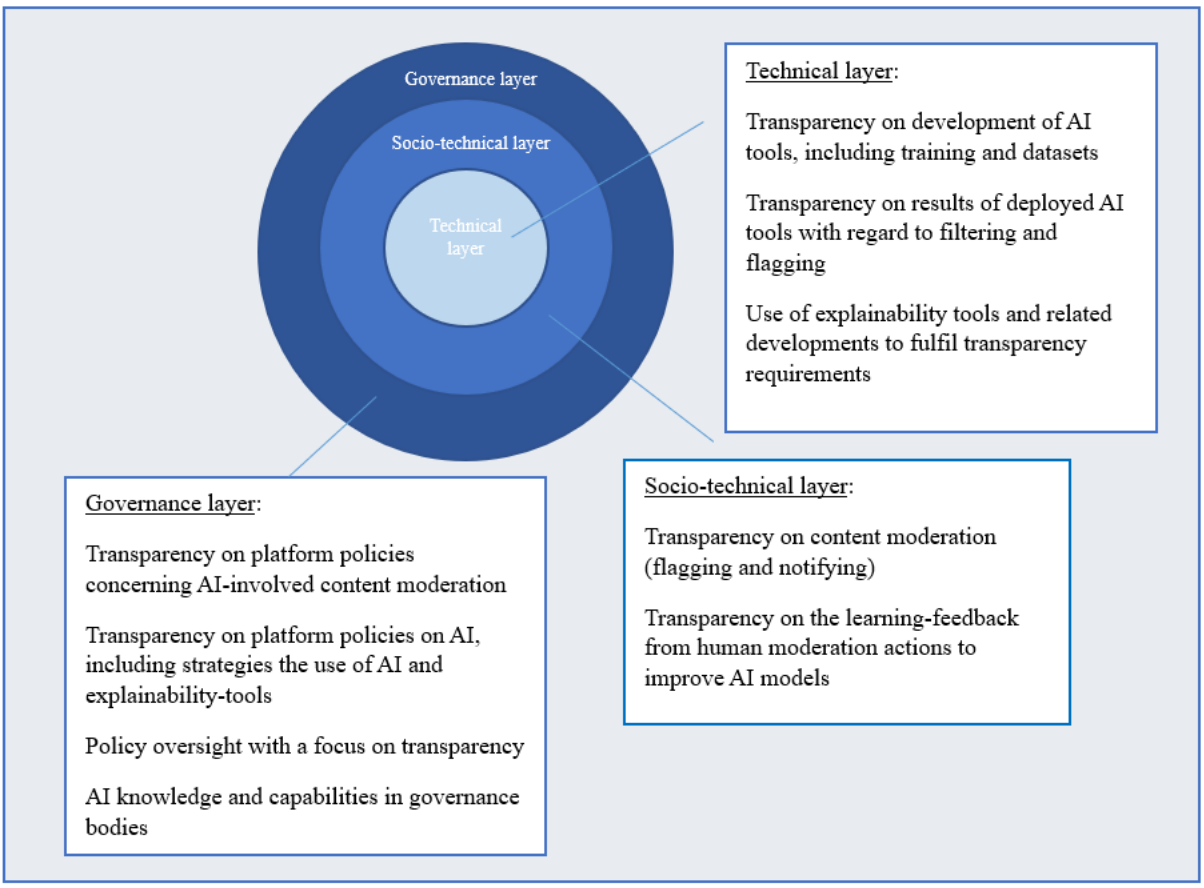


Figure 5: Transparency requirements related to the use of AI in content moderation

5.2. Analysis of current regulation on content moderation

Now that we have a notion of what sort of transparency is required, in this paragraph, we seek to analyse implemented measures based on their effectiveness and reach, their type (self-regulation, co-regulation, or legislation) and transparency aspects.

With regard to content moderation, regulation has followed after the technology is implemented and used, which can be seen in the following figure (figure 5). This timeline provides insight into the evolution of social media and the development of regulation.

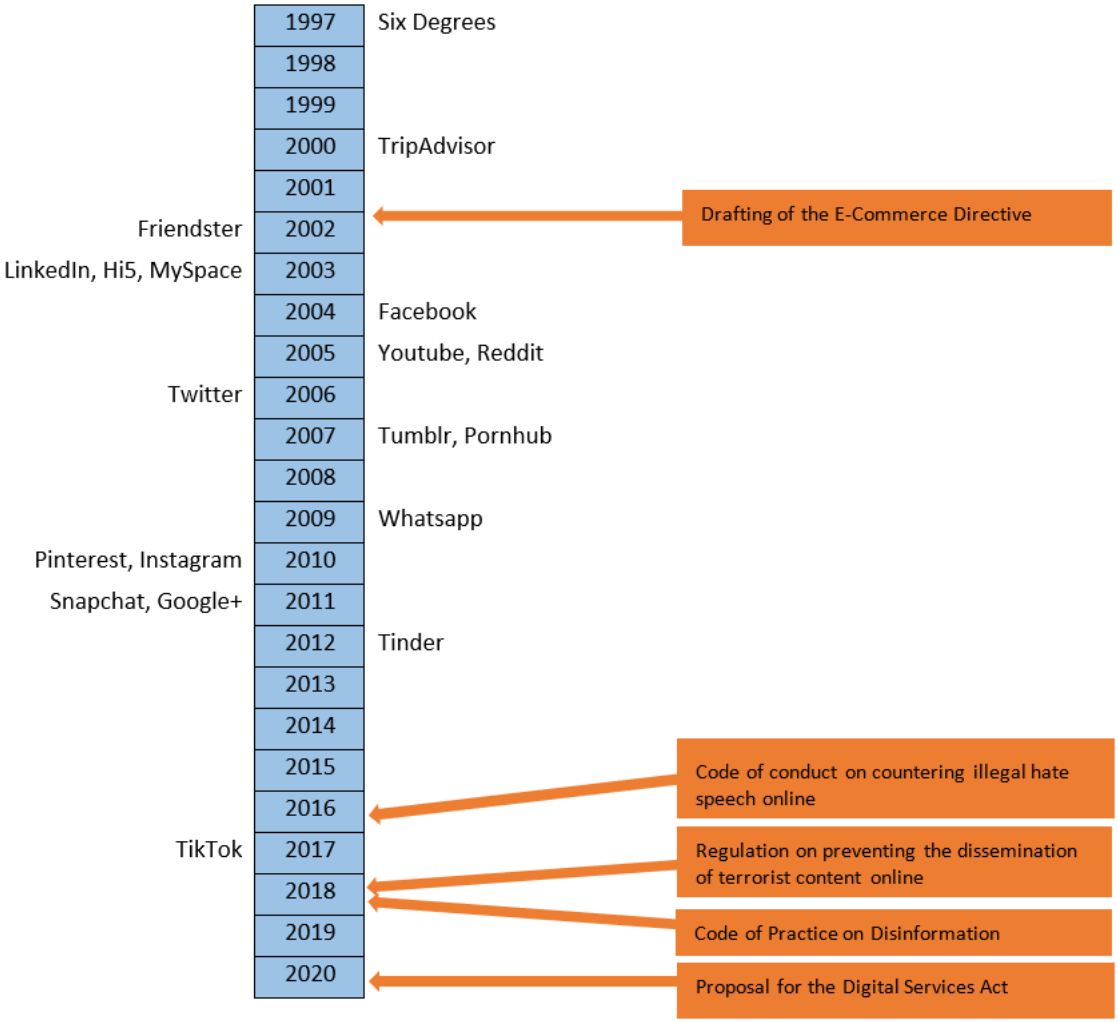


Figure 6: Timeline of social media platforms

(This figure is based on a timeline developed in a paper by S. Humphreys[91]. It has been extended to 2020 for the purpose of this theses.)

This graph presents the launch dates of some of the largest social media platforms and the issue dates of regulation in the EU. This set of regulation is relevant for content moderation in the EU and this will therefore be the scope for analysis.

E-Commerce Directive

To understand the current state of regulation on algorithmic content moderation and transparency we will go back to 2002 when the basis for regulating large technology companies in Europe was laid

[91] Sarah Humphreys, 'Tweeting into the Void?: Creating a UK Library Twitter List and Analyzing Best Practice - Successes and Myths', *Insights: The UKSG Journal*, 32 (2019) <https://doi.org/10.1629/uksg.471>.

down in the E-Commerce directive [92]. This directive aims at stimulating international online commercial activities by regulating online service providers and other online commercial organisations to ensure safety and security for citizens of the EU.

Article 14 of this directive [93] applies to content moderation and it states the following:

> *Where an information society service is provided that consists of the storage of information provided by a recipient of the service, Member States shall ensure that the service provider is not liable for the information stored at the request of a recipient of the service, on condition that:*
>
> *(a) the provider does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent; or*
>
> *(b) the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information.*

For social media platforms, the implementation of the E-Commerce Directive has taken the form of the Notice and Takedown-procedure for illegal content. Only after being notified, the platforms are obliged to take action. There is no obligation to monitor the information that they process, "nor a general obligation actively to seek facts or and used by industry; circumstances indicating illegal activity," as stated in article 15 of this directive states.[94] Proactively monitoring or flagging the content was a requirement, probably because at the time of drafting this directive, there were no means available to do this, and at that time, one could not have foreseen the vital role that these platforms were to play in society. Flagging content by AI is a form of proactive monitoring.

Transparency and evaluation

In this directive, transparency is mentioned in two places: 1) transparency concerning unsolicited advertisements and 2) transparency in relation to commercial promotional offers[95]. Keeping the historical development in mind, it may not be a surprise that transparency in the content moderation is not a topic in this regulation. The E-commerce Directive was drafted many years before the explosive growth of the social media platforms, with Facebook being introduced in 2004 and Twitter following in 2006 (see figure 5). AI was still in its early stages of development and the inherent problems related to transparency did not exist yet. The debate on contemporary issues like fake news, the increase in opacity due to AI and the struggle for balance between freedom of speech and content moderation had not started at that time. The. What also could not have been foreseen was the social and political power that these online platforms would grow into and the special role that transparency and accountability would have as a means to enable regulation of these powers. Therefore the directive does not explicitly cover the use of AI in content moderation or the regulation of

[92] 'E-Commerce Directive | Shaping Europe's Digital Future' <https://ec.europa.eu/digital-single-market/en/e-commerce-directive> [accessed 9 December 2020].
[93] 'E-Commerce Directive | Shaping Europe's Digital Future'.
[94] Article 15 of the E-Commerce Directive
[95] See page 5 of the E-Commerce Directive

transparency. Nevertheless, it is remarkable that this basic requirement for notice and takedown still serves a very relevant purpose, taking into account the development of social media and AI.

Over the years the E-Commerce Directive has been evaluated by the EU regularly. One of the criticisms on the content moderation rules in an assessment from May 2020 [96] is that "the Directive lacks sufficient safeguards to prevent violations of fundamental rights, in particular, freedom of expression". Lack of transparency-requirements in the notice and takedown-process contributes to over-notification and over-removal. As a remedy, increased transparency on the assessment of notifications, fines for over-removal and specific oversight-powers by regulators are recommended. Enforcement options are also missing in this framework.

Gap identification

There are some gaps related to regulation of transparency that can be pointed out.

Technical level

Due to the timeline of drafting this directive and the rise of most social media platforms, this directive does not cover the use of algorithms in content moderation. None of the requirements from the technical layer (as outlined in paragraph 5.1) are part of this directive.

Socio-technical level

The abovementioned report identifies a gap in the socio-technical layer regarding over-removal and over-notification. Increased transparency and enforcement options, such as fines, are recommended.

Governance level

There also is a gap with regard to the four requirements from the governance layer: on platform policies concerning AI-involved content moderation and AI strategies, policy oversight, and governance on building AI knowledge and capabilities in governance bodies.

Code of conduct on countering illegal hate speech online

The E-Commerce Directive has served as a baseline upon which, over time, additional and more elaborate regulation targeting several specific topics has been added. These are for example, the code of conduct on countering illegal hate speech and self-regulatory standards to counter disinformation. In this paragraph, we will further examine the code of conduct on countering illegal hate speech.

All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent, or national or ethnic origin is considered to be illegal hate speech under this code of conduct, [97] which was issued in March 2016.

---

[96] Alexandre de Streel and Martin Husovec, *The E-Commerce Directive as the Cornerstone of the Internal Market: Assessment and Options for Reform*, 2020 <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/648797/IPOL_STU(2020)648797_EN.pdf> [accessed 11 December 2020].

[97] 'The EU Code of Conduct on Countering Illegal Hate Speech Online | European Commission' <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-

The goal of this code of conduct, which was drafted in cooperation with Facebook, Microsoft, Twitter, and YouTube, was prevention and countering the spread of illegal hate speech online.

This code of conduct highlights the collective responsibility to maintain a secure digital space for users of the platforms: both the government and private companies have a role to play. Keeping the digital space clear from illegal hate speech is a task that has to be accomplished on top of the commitment to protect the right to free speech of individuals.

Transparency and evaluation

To balance both these sides, transparency and accountability are essential. It has to be clear what decisions led to the removal of content and how and on what grounds the decisions were made. This crucial role of transparency, however, is not reflected in the code of conduct. A need for clear processes on the assessment of notifications is stated, but only in the last section of the code of conduct transparency is mentioned in the context of an agreement to work on promoting transparency.[98] Artificial intelligence in the context of content moderation is not mentioned at all.

Every few years, reports are published reviewing the achieved levels of compliance as well as the effort that is put into it.[99] Over the period 2016- 2019, the reported assessment rate of flagged content was 89% end the removal rate was seemed to be stable at 70%. There are also numbers on the size of the workforce (15.000 people at Facebook and 10.000 across Google and Youtube) and some insight on the numbers of appeals by users.[100]

Also worthwhile noticing is that since these reports are the instruments for the larger platforms to provide transparency on their efforts and struggles on content moderation, it is interesting to see to what level and extent transparency actually is provided. Facebook, for instance, has a bi-weekly forum with its experts to discuss trends and issues and the impact that these should have on the community guidelines. The meeting minutes are publicly available[101], allowing some light on the inner circle of content moderation.

Following screen-print comes from one of these meeting minutes as an example to understand the topics which are discussed and the level of transparency that is allowed to the outside world (regulators and anyone else who is interested). This example from the bi-weekly forum is about how to deal with the removal of a user who actively engages in hate-speech.[102] This example provides a sense of transparency on Facebook's deliberations on such topics and what recommendations are made by the experts in this forum.

---

xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> [accessed 11 December 2020].

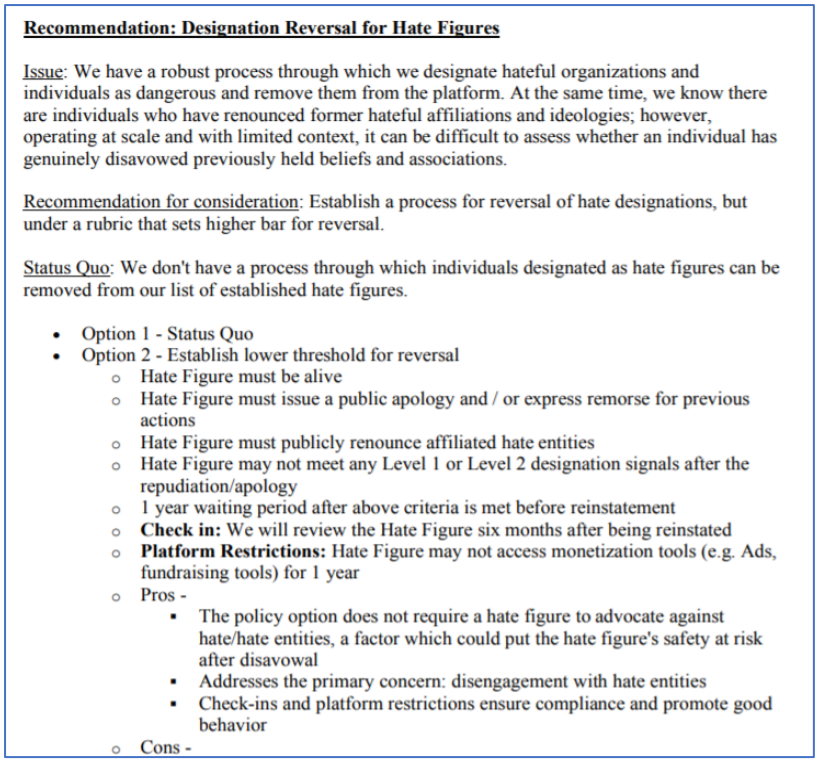
[98] P3 of the Code of Conduct: "They also agree to further discuss how to promote transparency and encourage counter and alternative narratives."

[99] EU, 'Assessment of the Code of Conduct on Hate Speech on Line Stae of Play', *European Commission*, 2019.September 2019 (2019), 1–9.

[100] P6 of the Code of Conduct: Facebook reports having received 1.1 million appeals related to content actioned for hate speech between January 2019 and March 2019, and 130,000 pieces of content were restored after a reassessment.

[101] 'Product Policy Forum Minutes - About Facebook' <https://about.fb.com/news/2018/11/content-standards-forum-minutes/> [accessed 11 December 2020].

[102] https://about.fb.com/wp-content/uploads/2018/11/11.27.18-Content-Standard-Forum-Minutes-.pdf



Recommendation: Designation Reversal for Hate Figures

Issue: We have a robust process through which we designate hateful organizations and individuals as dangerous and remove them from the platform. At the same time, we know there are individuals who have renounced former hateful affiliations and ideologies; however, operating at scale and with limited context, it can be difficult to assess whether an individual has genuinely disavowed previously held beliefs and associations.

Recommendation for consideration: Establish a process for reversal of hate designations, but under a rubric that sets higher bar for reversal.

Status Quo: We don't have a process through which individuals designated as hate figures can be removed from our list of established hate figures.

- Option 1 - Status Quo
- Option 2 - Establish lower threshold for reversal
  - Hate Figure must be alive
  - Hate Figure must issue a public apology and / or express remorse for previous actions
  - Hate Figure must publicly renounce affiliated hate entities
  - Hate Figure may not meet any Level 1 or Level 2 designation signals after the repudiation/apology
  - 1 year waiting period after above criteria is met before reinstatement
  - **Check in:** We will review the Hate Figure six months after being reinstated
  - **Platform Restrictions:** Hate Figure may not access monetization tools (e.g. Ads, fundraising tools) for 1 year
  - Pros -
    - The policy option does not require a hate figure to advocate against hate/hate entities, a factor which could put the hate figure's safety at risk after disavowal
    - Addresses the primary concern: disengagement with hate entities
    - Check-ins and platform restrictions ensure compliance and promote good behavior
  - Cons -

Figure 7: Screen-print from Content Standards Forum minutes – November 27, 2018

This is just one of the topics that are discussed. We do not know how and by whom the agenda of this forum is set, whether all delicate issues that need discussion are part of this forum, or if there are issues that are only discussed behind closed doors. At best, these meeting minutes provide partial transparency.

Yet another attempt towards more transparency is on the use of technology for content moderation. Statistics are provided on removed content, which was flagged by machines. Facebook reported 65.4% in the first quarter of 2019, which apparently is an increase compared to the number for previous months, 51.5%. Youtube compares the number from 2017, 79 %, to the second quarter of 2019, which was 87%. Humans do the final assessment of all the flagged content; "human-in-the-loop" is standard procedure for these companies. These percentages show that the number of potentially abusive content that is flagged by Artificial intelligence is increasing. The reporting timeframes are ad hoc and there is no reporting structure across the several platforms, which makes it impossible to compare the numbers. Furthermore, no explanation is given for the increase in these numbers.

Looking back to the Code of conduct on countering illegal hate speech online, the following can be concluded regarding transparency. The code of conduct has the form of a public commitment and as such there are no hard (legal) requirements on the efforts or results to be delivered by the platforms. Concerning transparency, the relevant stakeholders 'agree to discuss further how to promote transparency'. The partial transparency provided in the published meeting minutes seems to fit these soft commitments expressed in this code of conduct. There currently is no push for the digital platforms to go beyond this partial transparency.

This lack of transparency is also a topic in an article by the Guardian published in 2018 on the stance of the EU regarding Facebook and the code of conduct on hate speech. The EU expressed that trust underlying this cooperation had to be rebuilt by Facebook following the Cambridge Analytica scandal and the mistakes made in handling content regarding the Rohingya genocide in Myanmar.[103] Transparency could help build this lack of trust.

According to an article from Techcrunch on the latest assessment of this code of conduct, the lack of trust and transparency are still an issue.[104] The continuation of transparency issues and other issues is leading towards the end of public-private collaboration on hate speech-countering. This article mentions that new rules are to be set in the Digital Services Act as part of a broader legislative framework.

Gap identification

Knowing that transparency is a crucial element and seeing the currently the provided details, which are very ad hoc and fragmented, there is a need for more clear and stringent rules. Based on the three-layer model, the following gaps can be identified in regulation:

Technical level

On the technical level, currently, there are no requirements to provide information on the used technology or debates and mistakes leading to the choice of technology. Improvements in numbers of flagged and handled content are presented without a clear explanation of the underlying improvement: are there technological changes or advancements that cause the increase in numbers, or was it due to an expansion of the human workforce? Information about the qualitative performance of the AI model could be important to the regulators. For example, information on the existence of bias or the limitations of the use of AI in content moderation. The current regulation does not explicitly require transparency on this technological level. These requirements could be part of regulation that is being prepared for the future.

Socio-technical level

On the socio-technical level, there is a requirement that users have to be informed on the decisions and actions taken on their content. There, however, is no legal obligation on the side of the big platforms and furthermore, the code of conduct does not apply to all platforms that are being used by EU citizens. Only a few platforms have committed themselves to the code of conduct. If the topic of hate speech is finalised in a legal framework in de Digital Services Act, in the future, all platforms will have to abide by the new rules.

Governance layer

On the governance level, there is transparency on platform policies regarding content that are published to inform the users. What is missing is information on policies concerning the use of AI. In the earlier mentioned article from the Guardian, Facebook mentioned that artificial intelligence was not able to cope with the demands regarding removing hateful speech. Additional human moderators were needed. Such governance issues should be discussed, not only with regulators but also with other digital platforms so they can profit from the lessons learned. In the code of conduct, there is a commitment to intensify cooperation to share best practices. But again, hard obligations to do so are missing. With regard to

[103] EU threatens to crack down on Facebook over hate speech | European commission | The Guardian
[104] On illegal hate speech, EU lawmakers eye binding transparency for platforms | TechCrunch

transparency of AI, there are no obligations to provide transparency.

The level of voluntariness is such that many relevant transparency-needs from the three-layer Model are not covered and they are also not prioritized by the big technology platforms. As long as the full picture of transparency is not provided, it remains unclear what the value of the fragments of reported numbers is. Seeking transparency means setting different reporting standards for the companies and stepping away from voluntariness. It is the regulator's turn to take up this first responsibility for the sake of transparency.

The Code of Practice on Disinformation

The next concern that the EU had to address concerning content moderation, was disinformation. In 2018 the Code of Practice on Disinformation was issued by the EU and it was signed by the largest technology platforms like Facebook, Google, and Twitter, and Microsoft, with TikTok following later. The intention of this code, which is a set of self-regulatory principles, is to implement measures to counter disinformation and to provide more transparency to users on the assessment of fake news and political advertising.[105]

Transparency and evaluation

In this code, transparency is more prominently present. In a more general sense, the code states that transparency is essential and the stakeholders state to committing to transparency. In the more specific context of advertisements, the code dictates that transparency is needed to make consumers understand why they are targeted with certain advertisements, and in the context of empowerment of consumers transparency is mentioned as an aspect that needs to be developed further. The following is stated:

> *"Such transparency should reflect the importance of facilitating the assessment of content through indicators of the trustworthiness of content sources, media ownership, and verified identity. These indicators should be based on objective criteria and endorsed by news media associations, in line with journalistic principles and processes."*

Yearly reporting is done after self-assessment by the platforms as agreed upon by all who signed the code. The reports[106] present the objectives of the platforms, the developments in this field, and they provide some numbers of content that have been flagged and handled.

The 36 pages of the report by Twitter in 2019 contain more details on what they are confronted with (high numbers of possible fake accounts[107] that have to be checked and the struggle with functionalities needed to provide transparency to users.[108] Although Machine Learning is named as a

[105] 'Code of Practice on Disinformation | Shaping Europe's Digital Future' <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> [accessed 11 December 2020].
[106] EU, 'Annual Self-Assessment Reports of Signatories to the Code of Practice on Disinformation 2019 | Shaping Europe's Digital Future' <https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019> [accessed 15 December 2020].
[107] P19 of the Twitter annual report
[108] P10 of the Twitter annual report

key element for Twitter, there is no mention of specific areas where AI systems are being applied. The following is reported on fighting malicious automation:[109]

> *"Twitter fights spam and malicious automation strategically and at scale. Our focus is increasingly on proactively identifying problematic accounts and behaviour rather than waiting until we receive a report. Our primary goal on this front is to identify and challenge accounts engaging in spammy or manipulative behaviour before users are exposed to misleading, inauthentic, or distracting content."*

This points towards the use of AI to detect spam behaviour, but no more details are provided on the type of technology that is being used or what the accuracy of the systems is.

Understandably, such information is considered proprietary and sensitive since Twitter (or the other platforms involved in this Code) is a commercial company, but this provides little room for regulators to critically review the measures they have taken and assess whether this would be sufficient or not.

The annual report by Facebook is elaborate as well, providing numbers on accounts that have to be assessed:

> *"The number of fake accounts disabled spiked up from 1.2 billion accounts in Q4 2018 to 2.19 billion in Q1 2019, largely due to increased automated attacks by bad actors who attempt to create large volumes of accounts at one time."*

Furthermore, Facebook reports on many new procedures around preventing (suspected) fake news from going viral before it is fact-checked or fake news about vaccines and other issues that are important to society.

Gap identification

Technical layer

Much insight is provided, mostly fragmented information, so the issue remains that this reporting on transparency does not allow assessment of the performance of the technology. Since no insight into the range of the problem is provided, nor required (for example, an indication of the magnitude of disinformation-related content), the relative effectiveness of the flagging and handling process cannot be discussed.

Socio-technical layer

This code emphasizes the need to inform the users on moderation activities; however, there are no hard requirements or obligations.

Governance layer

There is no clarity on oversight and enforcement activities. Precise requirements on transparency are missing in the regulation, as well as regulatory efforts to increase literacy on the subject of AI in content moderation.

[109] P15 of the Twitter annual report

In concept: Digital Services Act

In December 2020, the EU published the concept of the new regulation to govern digital platforms. There will be more focus on transparency, increased responsibility for digital platforms, and more insight into how their algorithms work.[110] These rules will be shaped in the Digital Services Act[111](DSA).

Before the publication, there had been much debate on the (expected) content of the DSA. The new rules were expected to be "revolutionary" and " likely to require dramatic changes in business practices and even business models."[112]

A concept version has been published during the finalising phase of this thesis. Sections relevant to this research have been analysed.

Transparency aspects

Analysis of the Digital Services act's concept version clearly shows the intention towards more precise regulation on content moderation and transparency. A comprehensive transparency requirement will apply to the notice and takedown-procedure: users will be provided with full explanations on why and how their posts have been handled. Additional reporting will be required by regulators on several aspects of content moderation yearly.

Algorithmic and artificial intelligence systems of the large platforms will be subjected to independent auditing to assess the accountability of the large platforms. The audit reports of these accountability and transparency audits should "give a meaningful account of the activities undertaken, and the conclusions reached. It should help inform, and where appropriate, suggest improvements to the measures taken by the large online platforms to comply with their obligations under this Regulation."[113]

Platforms that are considered to be 'large' will be subjected to additional requirements, such as regular risk assessments and intensified transparency requirements to manage the systemic risks that they pose to society. Risk management will have a special focus on content moderation.

In accordance with the E-commerce-Directive general monitoring is still prohibited. The risk for infringements on free speech seems to be too much. Also, there is no obligation for the platforms to take proactive measures related to illegal content.

Gap identification

Technical layer

The Digital Services Act puts focus on the development and use of algorithms. The severe measures related to audits and reporting seem appropriate, considering the profound impact of these algorithms on people and society. In-depth obligations on the technical level regarding AI explainability, as mentioned

[110] Siladitya Ray, 'Google, Amazon, Microsoft Must Disclose How They Rank Search Results Under New EU Rules', *Forbes*, 2020 <https://www.forbes.com/sites/siladityaray/2020/12/07/google-amazon-microsoft-must-disclose-how-they-rank-search-results-under-new-eu-rules/?sh=72ed61ca11f3> [accessed 14 December 2020].
[111]EU, 'Digital Services Act (Concept)', 2020 <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en> [accessed 21 December 2020].
[112] Digital Services Act: How the EU is going after Big Tech (cnbc.com)

[113] Article 61 of the concept of the Digital Services Act

in the EU White Paper on AI, are not mentioned in the DSA. Perhaps a complementary role is anticipated for AI regulation.

Socio-technical layer

There is an increase in obligations on transparency reporting.

Governance layer

There still is a need to define a consistent and structured framework for reporting. Enforcement measures are discussed. Furthermore, it seems plausible that knowledge and capabilities in AI will have to be developed to facilitate auditing and interpreting results from these audits.

The topic of monitoring obligations might still need debate. Currently available technology, involved in hash-matching and classifying to filter certain harmful images, is already performing a form of proactive monitoring. It is not clear how these activities fit into the regulatory rules. The debate could provide more clarification on this topic.114

In concept: regulation on preventing the dissemination of terrorist content online

Another set of regulation which is being prepared is the Regulation on preventing the dissemination of terrorist content online. Any content that has been identified as terrorist content will have to be removed within an hour under this regime. Failure to comply could result in a fine of up to 4% of the global turnover for the last business year of the platform. There also will be stricter requirements towards reporting, including transparency requirements and then obligation for the EU monitor results and the impact of measures.

Transparency aspects

Transparency is well represented in this concept regulation. Transparency reporting is required, and transparency towards the users is emphasized. According to Gregorio,[115] this approach can be seen as an example of the more strict approach on transparency and accountability by the EU, moving away from the earlier noncommittal approach.

Gap identification

Technical layer

There is a general focus on more transparency, which applies to the technical layer as well. No explicit mention of algorithmic transparency is made.

Socio-technical layer

Transparency on the socio-technical layer is required in quite an elaborate way. Transparency to the users of take-downs is required and transparency on criteria for flagging and removing content is required.

---

[114] Emma J. Llansó, 'No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem', *Big Data and Society*, 7.1 (2020) <https://doi.org/10.1177/2053951720920686>.
[115] Giovanni De Gregorio, 'Democratising Online Content Moderation: A Constitutional Framework', *Computer Law and Security Review*, 36 (2020), 105374 <https://doi.org/10.1016/j.clsr.2019.105374>.

Governance layer

This set of regulation is more demanding than the predecessors regulating harmful content and disinformation. Regular transparency reporting is required, and fines will be used to enforce regulation.

Conclusion

In conclusion, an overview of the several types of regulation is presented in the table below. We can see that the further we move in time, the more strict and binding regulation becomes. Transparency and artificial intelligence start to take a more prominent place in the regulation.

Name	Type of regulation	Requires transparency in content moderation	Explicitly applies to AI used in content moderation
E-Commerce Directive	EU-regulation	Basic, best effort	No (historical)
Code of conduct on countering illegal hate speech online	Co-regulation	Yes, but still minimal	No, not mentioned
The Code of Practice on Disinformation	Co-regulation	Yes, more prominent	Yes, but minimal
Digital services Act (in concept)	EU-regulation	Yes, clear and elaborate	Yes. Some details are left to upcoming regulation

Table 1: Short overview of regulation on content moderation and transparency

There is an increase in demand for transparency. The reported output, however, is not enough to assess the in-depth quality of content moderation. Therefore steering the behaviour of these platforms concerning the current debate on free speech versus content moderation or on the profound influence these platforms have on what society reads, views, and believes is still a struggle.

It should be no surprise that upcoming regulations (Digital Services Act and the Regulation on preventing the dissemination of terrorist content online) are more demanding in many areas, including transparency.

5.3 Regulation on AI

Another relevant area for this thesis that has not been covered yet is regulation on AI. The EU recognizes the increasing value of artificial intelligence and seeks to promote uptake of AI while trying to manage the risks that come along with AI as a lack of transparency and bias. Preparatory outlines have been presented, such as the Ethical Guidelines for Trustworthy AI and the White Paper on Artificial Intelligence. Both these outlines will be analysed to understand how transparency and transparency related to content moderation are positioned.

EU Ethical Guidelines for Trustworthy AI

The AI HLEG (Artificial Intelligence High-Level Expert Group) with members across Europe has developed and published the Ethics Guidelines for Trustworthy AI[116] in April 2019 with the aim to promote trustworthy artificial intelligence systems. Trustworthy AI is characterised by three elements, namely AI being lawful, ethical, and robust. The guidelines are organised in a framework consisting of the four ethical principles on which the use of AI systems should be founded and the seven key requirements which apply to the realisation of AI.

The ethical principles cover the area of 1) human autonomy ("Functions should follow human-centric design principles") and 2)prevention of harm, both as a result of malicious use or caused by information asymmetry. 3)Fairness is an important topic (no unfair bias). The procedural side of fairness brings new elements: AI decisions should be explicable, the entity making the decisions identifiable, and people should be provided the opportunity to seek redress against AI decisions.

The last principle is explicability, which is about processes is being transparent, capabilities and purposes of the system openly being communicated, and decisions being explainable. In the case of black-box algorithms, extended measures may be needed, such as 'traceability, auditability and transparent communication on system capabilities.'[117]

The Guidelines (will) apply to all AI systems; content moderation is not mentioned explicitly. In another part of the framework, the seven key requirements cover topics like human agency and oversight, transparency, and accountability. The complete framework is presented in figure 8.

[116] High-Level Independent Group on Artificial Intelligence (AI HLEG), 'Ethics Guidelines for Trustworthy AI', *European Commission*, 2019, 1–39.
[117] Page 13 of 'Ethics Guidelines for Trustworthy AI

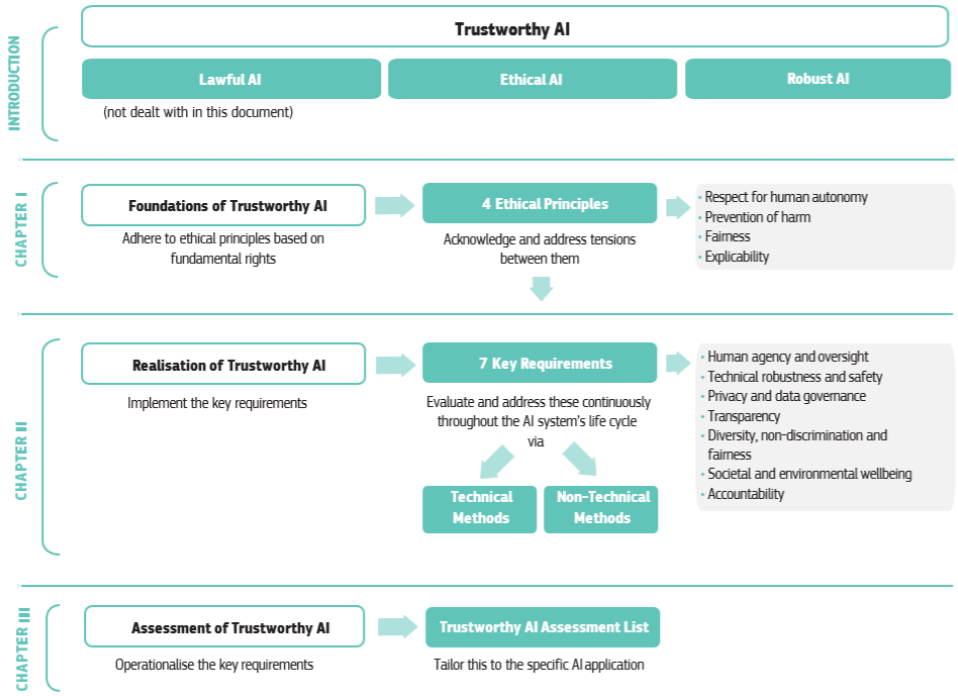


Figure 8: The guidelines as a framework for Trustworthy A

Although transparency is explicitly named as one of the seven key requirements, analysing all the key requirements results in many transparency-related aspects in several of the other requirements. Following are the findings:

Transparency

Human agency and oversight: prior to the development of the AI system, an evaluation on possible infringements on human rights is recommended to do, as well as the following: 'mechanisms should be put into place to receive external feedback regarding AI systems that potentially infringe on fundamental rights.' From a socio-technical point of view, this mechanism could be a flagging or complaint-mechanism. In the technical layer, one could also foresee assessments and simulations with specific test-datasets to test the system for any sort of bias or other types of human-rights infringements. This requirement also includes that humans have to be able to understand the AI-system so they can make informed decisions regarding the system. This could be delivered through interpretability (as discussed in paragraph 3.4): the deployer of the system has to be able to explain the goal and how the system is functioning to the users and society.

Requirements related to technical robustness and safety also implicitly claim the need for transparency: assessing *accuracy* (correct judgments and classifications) and *reliability* (proper functioning of the system, reproducing the same results under the same conditions) require transparency on development and training of the system.

When looking into the key requirement transparency, we find that transparency is divided into three parts. The first is *Traceability*: This concerns the data sets, the methods, and the algorithms used to develop the AI system. Traceability also applies to the ability to trace back the decisions made by the AI

system to facilitate redress. The second part is *explainability*: This is about allowing humans to trace back the technical decisions made by AI. The influence of an AI system on an organisation and the design choices should be explainable. To ensure transparency on a technical level, the option of employing Explainable AI (XAI) is discussed. The last part is c*ommunication*. It has to be apparent to any user that they are dealing with AI systems and not human beings.

Addressing the next key requirement, diversity, non-discrimination, and fairness leads us to transparency as well: 'oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner.'[118] Assessing the systems requirements and decisions may point towards transparency requirements on the technical layer like explainability.

The final requirement that includes transparency is societal and environmental well-being. The impact of AI in the context of society and democratic processes should be monitored and considered. Such a requirement will imply transparency requirements, which have to be formulated more clearly in regulation to come.

Gap identification

Technical layer

Traceability and explainability are identified as measures on the technical layer. A*ccuracy* (correct judgments and classifications) and *reliability* (proper functioning of the system) link to the requirement of transparency on development, training, and deployment of AI tools.

Socio-technical layer

Transparency on the socio-technical layer is addressed extensively in the form of transparency on purpose, constraints and decisions. Mapping this to content moderation will result in transparency to the users on take-downs and on criteria for flagging and removing content is required.

Governance layer

These guidelines are not binding. Bringing in human oversight and accountability relates to transparency on the governance level. The requirements for transparency are explicit at some points. At other points they are hidden under other requirements. Explicit clarity on these requirements should be part of future regulation on this topic. It might also be beneficial to utilise a common terminology to ensure that both regulators and developing parties understand each other.

White Paper On Artificial Intelligence - A European approach to excellence and trust

In February 2020, the EU published a white paper on AI[119] proposing several policy options to stimulate the development of trustworthy AI with the aim of achieving an "ecosystem of excellence." This white paper will also aim to create an ecosystem of trust by ensuring the protection of fundamental and human rights by applying a human-centric approach. This white paper focuses on AI in general but is applicable to content moderation as well. It builds upon the previously discussed ethical guidelines for AI and recognises transparency as essential for explainability. This white paper

[118] Page 18 of 'Ethics Guidelines for Trustworthy AI
[119] EU, 'WHITE PAPER On Artificial Intelligence - A European Approach to Excellence and Trust'.

also explicitly mentions that many of the key requirements are covered by existing regulation or legislation, except transparency, traceability, and human oversight.

Transparency

Bias, amplified by opaque AI technology, is named as one of the dangers to human rights. Apart from adjusting or creating the necessary legal frameworks to address these cases, transparency is needed to facilitate legal accountability. Without detailed insight into the functioning of the systems, it will not be possible to address any infringements on human rights.

Transparency is also required to have insight into the changing nature of operational AI systems. When the AI model changes over time, new risks may be introduced. A mechanism to assess these risks and commence the necessary legal adjustments will have to be set up.

Besides this, transparency on record-keeping of algorithmic development, training data, and the case data is recommended. In the case of high-risk systems, transparency on capabilities and limitations is required. Users should also be provided knowledge of the AI system's existence and purpose in an easily understandable way. In the context of content moderation, this could be translated to users knowing whether their content was removed by a person or by AI.

Gap identification

Technical layer

No explicit mention of content moderation is made. However, transparency requirements on the technical level, such as the requirement for record-keeping and transparency on capabilities and limitations of systems, are proposed.

Socio-technical layer

Transparency on the socio-technical layer is addressed. Users should be made aware that they are interacting with an AI system.

Governance layer

The purpose of transparency is legal accountability. A consistent European regulatory framework will be needed, and effort will be put into analysing whether "current legislation is able to address the risks of AI and can be effectively enforced, whether adaptations of the legislation are needed, or whether new legislation is needed." [120]

[120] EU, 'WHITE PAPER On Artificial Intelligence - A European Approach to Excellence and Trust'.

5.4 Summary of the gaps

Several gaps on several levels have been identified. In this section, a more comprehensive view of these gaps will be provided, discussed against the backdrop of the transparency requirements on the three-layer model, as discussed in paragraph 5.1 and presented in figure 5.

Gaps on the technical layer

Shifting the focus in content moderation towards the technical layer entails the aggravated need for transparency on the technical layer. The inherent opacity of AI brings new transparency challenges.

As discussed in paragraph 3.4, several artificial intelligence technologies are involved in content moderation, varying from hate speech detection to data mining for fake news detection and more. The current regulation does not address transparency of the development of AI tools. However, the Digital Services act and EU White Paper on regulating AI, have included requirements for insight into the tooling deployed by these platforms. Although platforms may be reluctant to share this kind of information since it is proprietary and important for competitive advantage, from the perspective of regulators, it is essential to know which tools are employed to assess how adequate they are and whether additional effort should be required.

Questions that regulators could ask to improve this type of transparency are:

- What are the characteristics of the deployed artificial intelligence technology? (For example, the strengths and weaknesses)Is this the best choice?
- How are the tools trained, with what goals in mind?

The second requirement on the technical layer concerns the results of deployed tools. It is not known how much hate speech or fake news is slipping through the maze of moderation. Regulation trusts that this content will be flagged and removed afterward via the feedback-loop as presented in the model of content moderation. However, no transparency on the results of deployed AI is required in current or upcoming regulation. This category of content, which is illegal or harmful, but is neither flagged by users nor by AI (visualized in figure 9 by the red quarter), will find its way to be viewed by users.

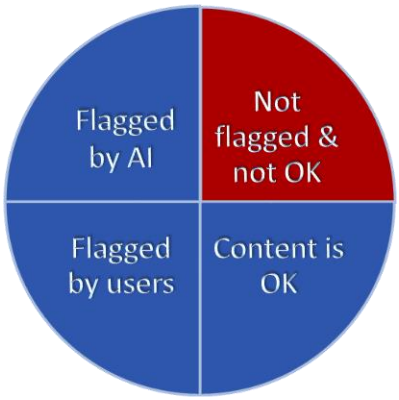


Figure 9: Different categories of user-generated content (not representing a quantitative division)

This information could be used to assess the effectiveness of the content moderation process. It will also aid in interpreting the value of the several fragmented reports on transparency that are provided by the digital platforms.

The questions that can bring more insight are:

- How do these technologies perform in general? Is there a benchmark?
- What is needed to reach a better performance of these technologies?

The use of explainability tools and related developments is the third requirement on the technical layer. Current regulation does not address the explainability of AI tools. The concept Digital Services Act proposes extensive regulation on this topic. The EU Guidelines and White Paper on AI also include this level of transparency, although not explicitly related to content moderation. In some sections, essential links towards the technicalities of AI (development and training) are also incorporated. However, cooperation between governance and developing companies is required in designing explainability to prevent the risk of inmates running the asylum.[121]

Gaps on the socio-technical layer

The requirement regarding having users informed on the decisions and actions taken on their content is present in all regulation sets that have been discussed. Enforcement is not in place. The second requirement concerning transparency on the feedback-loop to improve the performance of AI is not explicitly mentioned in the upcoming regulation, however the requirement in the EU White Paper regarding keeping records on algorithm development may include this. More explicit formulation may be needed to ensure that this is included in the resulting regulation on AI.

An additional gap on this layer is that the levels of required and reported transparency are not consistent, it varies from providing explanations to individuals to transparency reports on numbers of flagged content items that have been acted on. Another gap is that not all rules apply to all the platforms, for example Code of practice on disinformation and the Code of conduct on countering illegal hate speech online apply only to some of the platforms.

[121] Tim Miller, Piers Howe, and Liz Sonenberg, *Explainable AI: Beware of Inmates Running the Asylum*, 2017

Gaps on the governance layer

There is a wide variety of gaps in the governance layer. Existing and upcoming regulation address transparency of platform policies on content moderation. However, none of them mention transparency on policies or strategies on AI. This however, will be important in understanding developments in algorithmic content moderation.

There are some gaps in the application of existing regulation. Not all regulation is applicable to every platform, such as the Code of practice on disinformation and the Code of conduct on countering illegal hate speech online. Consistency across all platforms is missing.

Oversight and enforcement are not implemented in current regulation. In upcoming regulation, provisions on monitoring and enforcing are included. Monitoring AI tools will be a new regulatory challenge. Audits and designated tooling such as explainability tools will be needed on this level to understand how the AI model works. This will be an extensive task for regulators, which cannot be done without experts on AI and more specific experts on explainability and interpretability of AI. Development of expertise and capabilities will be a prerequisite for governance on this topic. The EU White Paper on AI mentions development of AI skills, however, not in the context of governance.

6. Conclusion and recommendations

In chapter 5, several gaps have been identified and discussed. In this chapter, the conclusions will be presented, followed by a reflection on the conducted research and recommendation for further research.

6.1 Conclusions

Content moderation is about optimizing the equilibrium between two critical values: freedom of speech and a safe and secure digital space. The main tasks are defining what is admissible content and assuring that inadmissible content is not allowed into the digital public space. Commercial digital platforms cannot be expected to carry this responsibility on their own without any incentives or obligations. They have their own commercial goals to serve. Tightened and more precise regulation is necessary. Overfitting the regulation will compromise freedom of speech. Underfitting the regulation will compromise the security of the digital space. An important aspect of assessing this balance is transparency, which has been proven laborious to achieve.

In this thesis, we aspired to conceptually understand the transparency issues in AI facilitated content moderation. We looked at the historical timeline of drafting regulation and the rise of social media and the three layer-model of cyberspace that was used to analyse AI facilitated content moderation. Transparency requirements on each level have been identified. Existing and upcoming regulation on content moderation and AI have been assessed. The identified gaps have been discussed in chapter 5. In this paragraph, the final conclusions will be discussed.

Conclusions on transparency of AI in content moderation

With the introduction of artificial intelligence tools in content moderation, an increased focus on regulation of the technical layer is required. Regulating technology comes with the legacy of lagging behind the facts; regulating artificial intelligence technology is no exception. Regulation on artificial intelligence is fragmented and still in an early stage of development. The Digital services Act in concept includes many ways for regulators to increase the level of transparency by being involved in the process of content moderation. The proposed options on reporting and auditing are potentially powerful measures to assess the (strategic) use of AI and essential for accountability purposes.

The specific case of using AI in content moderation is covered by both the Digital Services Act and the regulation on AI, which are under construction. The EU White Paper on Artificial Intelligence mentions exploring if the legal framework will need a change in order to better equipped to regulate developments in AI. The overlap and mutual synergy should be closely monitored.

Currently, reporting on transparency is fragmented. Several indicators are reported over arbitrarily chosen time periods. Regulation does not explicitly demand full transparency, nor does it clearly state what reporting is needed over which time periods. Seeking transparency means setting different reporting standards for the companies and stepping away from voluntariness. It is the regulator's turn to take up this first responsibility for the sake of transparency by using upcoming regulation to be precise and clear about requirements. Hopefully, it will not take too long to finalize regulation on AI. Developments on AI are already taking a flight and regulation is lagging.

Different concepts of transparency

After examining several documents both on the technical side and the government side, the following conclusion can be reached regarding the use of terminology regarding transparency. Transparency in the world of AI technology often relates to insight into the technical functioning of algorithms and to the ability to predict the outcome of an artificial intelligence model. In the governance world, transparency is linked to accountability and clarity. Users have to understand the underlying reasons for certain decisions taken by artificial intelligence. Transparency on the level of AI in the form of explainability does not cover the need for transparency as laid out in governance. This gap between the world of artificial technology and the world of governance will need extra attention when drafting further regulation on AI. There is a need for common terminology. A transparent debate between digital platforms and regulators on technical strategies and the role of human rights and free speech is needed to solve this gap in definitions and terminology.

Reflection on AI in content moderation

For a more nuanced view on the use of AI in content moderation, it has to be mentioned that AI brings many advantages. Without AI, content moderation on this scale would not be possible. Other benefits of AI are diminishing the traumatic side with content moderation for human moderators by pre moderating and blurring extreme images. The use of advanced face recognition systems could be life-changing for victims of sexual abuse of revenge porn. Improving transparency of these tools could stimulate the use of these tools, and thereby help to create a more secure cyberspace.

6.2 Reflection and recommendations

The primary focus of this research was on governance measures, as discussed in paragraph 4.4. For a better understanding of artificial intelligence, some concepts have been researched. Diving deeper into AI applications for content moderation can help develop more precise guidelines in governance to assess the tools. Future research on possible ways to categorise and benchmark these tooling is recommended.

Explainability and interpretability are used in the grey area between technology and interaction with technology by people. Clear definitions and common understanding could be beneficial to both worlds. In a broader perspective future research is recommended on creating common terminology to improve the dialogue between the world of governance and the world of technology. Developing a standard set of wordings and terminology could help to close the distance.

The scope of this research did not allow for examining methods that can be applied to assess the algorithms and datasets that are involved in the creation of an AI model. The Digital Services Act proposes to perform audits. Further research on methods that can be to audit and assess AI models will be necessary.

Further research on the possibility of certification of AI tools is also recommended. ENISA is working on a security certification framework for digital products. Could this certification framework be adapted to include AI technologies?

The final recommendation for further research is on the topic of monitoring obligations. Currently available technologies, such as hash-matching and classifying to filter certain harmful images, are a form of proactive monitoring. It is not clear how these activities fit into the regulatory rules. Further research could provide more clarification on this topic.

Bibliography

Adadi, Amina, and Mohammed Berrada, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*, 6 (2018), 52138–60 <https://doi.org/10.1109/ACCESS.2018.2870052>

Allcott, Hunt, and Matthew Gentzkow, 'Social Media and Fake News in the 2016 Election' <https://doi.org/10.1257/jep.31.2.211>

Van den Berg, Jan, Jacqueline Van Zoggel, Mireille Snels, Mark Van Leeuwen, Sergei Boeke, Leo Van Koppen, and others, 'On (the Emergence of) Cyber Security Science and Its Challenges for Cyber Security Education', *NATO STO/IST-122 Symposium in Tallin*, 2014, 1–10

Bini, Stefano A., 'Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?', *Journal of Arthroplasty*, 33.8 (2018), 2358–61 <https://doi.org/10.1016/j.arth.2018.02.067>

Bruns, Axel, Stephen Harrington, and Edward Hurcombe, '"Corona? 5G? Or Both?": The Dynamics of COVID-19/5G Conspiracy Theories on Facebook', 177.1, 12–29 <https://doi.org/10.1177/1329878X20946113>

Bulao, Jacquelyn, 'How Much Data Is Created Every Day in 2020?', *TechJury*, 2020 <https://techjury.net/blog/how-much-data-is-created-every-day/> [accessed 20 November 2020]

Cobbe, Jennifer, 'Algorithmic Censorship by Social Platforms: Power and Resistance', *Philosophy and Technology*, 2020 <https://doi.org/10.1007/s13347-020-00429-0>

'Code of Practice on Disinformation | Shaping Europe's Digital Future' <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> [accessed 11 December 2020]

Cormode, G; Krishnamurthy, B, 'View of Key Differences between Web 1.0 and Web 2.0 | First Monday', *First Monday*, 13.6 (2008) <https://firstmonday.org/ojs/index.php/fm/article/view/2125/1972> [accessed 20 November 2020]

van Dijck, José, 'Governing Digital Societies: Private Platforms, Public Values', *Computer Law and Security Review*, 36.xxxx (2020), 10–13 <https://doi.org/10.1016/j.clsr.2019.105377>

'E-Commerce Directive | Shaping Europe's Digital Future' <https://ec.europa.eu/digital-single-market/en/e-commerce-directive> [accessed 9 December 2020]

EU, 'Annual Self-Assessment Reports of Signatories to the Code of Practice on Disinformation 2019 | Shaping Europe's Digital Future' <https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019> [accessed 15 December 2020]

EU, 'Assessment of the Code of Conduct on Hate Speech on Line Stae of Play', *European Commission*, 2019.September 2019 (2019), 1–9

EU, 'Digital Services Act (Concept)', 2020 <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en> [accessed 21 December 2020]

EU, 'WHITE PAPER On Artificial Intelligence - A European Approach to Excellence and Trust)', *European Commission*, 2020 <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf> [accessed 11 December 2020]

European Commission, ‘Ethics Guidelines for Trustworthy AI | Shaping Europe’s Digital Future’, *Ethics Guidelines for Trustworthy AI*, 2019 <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> [accessed 21 December 2020]

Facebook, ‘Community Standards Enforcement’, *Facebook Transparency*, October 2018, 2019 <https://transparency.facebook.com/community-standards-enforcement> [accessed 23 November 2020]

‘Facebook Mission Statement 2020 | Facebook Mission & Vision Analysis’ <https://mission-statement.com/facebook/> [accessed 12 November 2020]

Ferrer, Xavier, Tom van Nuenen, Jose M. Such, Mark Coté, and Natalia Criado, *Bias and Discrimination in AI: A Cross-Disciplinary Perspective*, 2020 <http://arxiv.org/abs/2008.07309> [accessed 16 November 2020]

Forssbaeck, Jens, and Lars Oxelheim, *The Multi-Faceted Concept of Transparency The Multi-Faceted Concept of Transparency*, 2014, MXIII <www.ifn.se> [accessed 17 January 2021]

Gillespie, Tarleton, ‘Content Moderation, AI, and the Question of Scale’, *Big Data and Society*, 7.2 (2020) <https://doi.org/10.1177/2053951720943234>

Github LDNOOBW, ‘List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words’ <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words> [accessed 22 November 2020]

Goodfellow, Bengio and COurville, *Deep Learning*, *The MIT Press* (The MIT Press, 2016)

Gorwa, Robert, Reuben Binns, and Christian Katzenbach, ‘Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance’, *Big Data and Society*, 7.1 (2020) <https://doi.org/10.1177/2053951719897945>

De Gregorio, Giovanni, ‘Democratising Online Content Moderation: A Constitutional Framework’, *Computer Law and Security Review*, 36 (2020), 105374 <https://doi.org/10.1016/j.clsr.2019.105374>

Grimmelmann, James, ‘The Virtues of Moderation’, 42 (2015), 154–76 <https://doi.org/10.5810/kentucky/9780813169057.003.0008>

Heller, Brittan, *Combating Terrorist-Related Content Through AI and Information Sharing*, 2019 <www.annenbergpublicpolicycenter.org/twg> [accessed 22 November 2020]

High-Level Independent Group on Artificial Intelligence (AI HLEG), ‘Ethics Guidelines for Trustworthy AI’, *European Commission*, 2019, 1–39

‘How Twitter Policed Trump During the 2020 US Election - The New York Times’ <https://www.nytimes.com/2020/11/06/technology/trump-twitter-labels-election.html> [accessed 13 November 2020]

Humphreys, Sarah, ‘Tweeting into the Void?: Creating a UK Library Twitter List and Analyzing Best Practice - Successes and Myths’, *Insights: The UKSG Journal*, 32 (2019) <https://doi.org/10.1629/uksg.471>

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert, ‘Does Transparency in Moderation Really Matter?: User Behavior after Content Removal Explanations on Reddit’, *Proceedings of the ACM on Human-Computer Interaction*, 3.CSCW (2019) <https://doi.org/10.1145/3359252>

Langvardt, Kyle, 'Regulating Online Content Moderation', *Georgetown Law Journal*, 106.4 (2018), 1353–88 <https://doi.org/10.2139/ssrn.3024739>

Larsson, Stefan, and Fredrik Heintz, 'Transparency in Artificial Intelligence', *Internet Policy Review*, 9.2 (2020), 1–16 <https://doi.org/10.14763/2020.2.1469>

Livingstone, Sonia, and David R. Brake, 'On the Rapid Rise of Social Networking Sites: New Findings and Policy Implications', *Children and Society* (John Wiley & Sons, Ltd, 2010), 75–83 <https://doi.org/10.1111/j.1099-0860.2009.00243.x>

Llansó, Emma, Joris Van Hoboken, Paddy Leerssen, and Jaron Harambam, 'Artificial Intelligence, Content Moderation, and Freedom of Expression', 2020, 1–30 <https://www.ivir.nl/twg/.>

Llansó, Emma J., 'No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem', *Big Data and Society*, 7.1 (2020) <https://doi.org/10.1177/2053951720920686>

Marr, Bernard, 'How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read', *Forbes*, 2018, 1–5 <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=3b0d112460ba> [accessed 20 November 2020]

Miller, Tim, Piers Howe, and Liz Sonenberg, *Explainable AI: Beware of Inmates Running the Asylum*, 2017 <http://home.earthlink.net/>

Millican, Peter, 'The Philosophical Significance of the Turing Machine and the Turing Test', *Alan Turing: His Work and Impact*, 2013, 587–601

Muhammad, Zia, 'A Timeline of Social Media (Infographic)', *Digital Information World*, 2019 <https://www.digitalinformationworld.com/2019/10/social-media-history-infographic.html> [accessed 28 December 2020]

Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang, 'Abusive Language Detection in Online User Content', in *25th International World Wide Web Conference, WWW 2016* (International World Wide Web Conferences Steering Committee, 2016), pp. 145–53 <https://doi.org/10.1145/2872427.2883062>

Nooren, Pieter, Nicolai van Gorp, Nico van Eijk, and Ronan Ó Fathaigh, 'Should We Regulate Digital Platforms? A New Framework for Evaluating Policy Options', *Policy and Internet*, 10.3 (2018), 264–301 <https://doi.org/10.1002/poi3.177>

O'Keeffe, Gwenn Schurgin, Kathleen Clarke-Pearson, Deborah Ann Mulligan, Tanya Remer Altmann, Ari Brown, Dimitri A. Christakis, and others, 'Clinical Report - The Impact of Social Media on Children, Adolescents, and Families', *Pediatrics*, 2011, 800–804 <https://doi.org/10.1542/peds.2011-0054>

Press, Gil, 'A Very Short History Of Artificial Intelligence (AI)', *Forbes*, 2016 <https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/?sh=3b5954386fba> [accessed 19 November 2020]

'Product Policy Forum Minutes - About Facebook' <https://about.fb.com/news/2018/11/content-standards-forum-minutes/> [accessed 11 December 2020]

Rajaraman, V, 'John McCarthy – Father of Artificial Intelligence', March, 2014, 198–207

Ray, Siladitya, 'Google, Amazon, Microsoft Must Disclose How They Rank Search Results Under New

EU Rules', *Forbes*, 2020 <https://www.forbes.com/sites/siladityaray/2020/12/07/google-amazon-microsoft-must-disclose-how-they-rank-search-results-under-new-eu-rules/?sh=72ed61ca11f3> [accessed 14 December 2020]

Roberts Dissertation, Sarah T, *BEHIND THE SCREEN: THE HIDDEN DIGITAL LABOR OF COMMERCIAL CONTENT MODERATION*

Roberts, Sarah T, 'Content Moderation', 2017 <https://escholarship.org/uc/item/7371c1hf> [accessed 20 November 2020]

'Robotics Openletter | Open Letter to the European Commission' <http://www.robotics-openletter.eu/> [accessed 19 November 2020]

Salathé, Marcel, Thomas Wiegand, and Markus Wenzel, 'Focus Group on Artificial Intelligence for Health', *ArXiv*, 2018

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Robert Müller, 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models', *ArXiv*, 2017

Schank, Roger C., 'What Is AI Anyway?' <http://www.aistudy.com/paper/aaai_journal/AIMag08-04-004.pdf> [accessed 26 September 2020]

Silberg, Jake, and James Manyika, 'Tackling Bias in Artificial Intelligence (and in Humans) McKinsey', *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*, 2019, 1–8 <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>

'"Stop the Steal" Supporters Shift from Facebook to Parler to Peddle False Election Claims | The Independent' <https://www.independent.co.uk/news/world/stop-the-steal-facebook-parler-election-b1720473.html> [accessed 28 December 2020]

de Streel, Alexandre, and Martin Husovec, *The E-Commerce Directive as the Cornerstone of the Internal Market: Assessment and Options for Reform*, 2020 <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/648797/IPOL_STU(2020)648797_EN.pdf> [accessed 11 December 2020]

Suzor, Nicolas P., Sarah Myers West, Andrew Quodling, and Jillian York, 'What Do We Mean When We Talk about Transparency? Toward Meaningful Transparency in Commercial Content Moderation', *International Journal of Communication*, 13 (2019), 1526–43

T-davidson/hate-speech-and-offensive-language, Hate-speech-and-offensive-language/refined_ngram_dict.csv at master ·, 'Hate-Speech Lexicon', *Github* <https://github.com/t-davidson/hate-speech-and-offensive-language/blob/master/lexicons/refined_ngram_dict.csv> [accessed 22 November 2020]

'The EU Code of Conduct on Countering Illegal Hate Speech Online | European Commission' <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> [accessed 11 December 2020]

Tolan, Songül, Marius Miron, Emilia Gómez, and Carlos Castillo, 'Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia' <https://doi.org/10.1145/3322640.3326705>

Turing, A M, *Computing Machinery and Intelligence*, *Computing Machinery and Intelligence. Mind*, 1950, XLIX

'Twitter *Permanently* Bans Trump After Years Of Aggressive, Violent Rhetoric On Platform | HuffPost' <https://www.huffpost.com/entry/trump-twitter-ban_n_5ff733bbc5b61a92a8c08b8c> [accessed 10 January 2021]

Weimann, Gabriel, Axel Bruns, Stephen Harrington, Edward Hurcombe, Robert Gorwa, Reuben Binns, and others, 'Democratising Online Content Moderation: A Constitutional Framework', *Perspectives on Terrorism*, 6.1 (2020), 53–64 <https://doi.org/10.1177/2053951719897945>

Appendix A

Following is an interesting read to get a better understanding of the work a content moderator is doing. Although not scientific literature, still worth the time: A first-hand sharing of experience, feelings, and thoughts by a content moderator.

https://sz-magazin.sueddeutsche.de/internet/three-months-in-hell-84381