2021 ISCA Symposium on Security and Privacy in Speech Communication
10-12 November 2021, Online

Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation

Vera Schmitt[1], *Veronika Solopova*[2], *Vinicius Woloszyn*[1], *Jessica de Jesus de Pinho Pinhal*[1]

[1] Technische Universität Berlin, Germany
[2] Freie Universität Berlin, Germany

vera.schmitt@tu-berlin.de, solopov97@zedat.fu-berlin.de, woloszyn@tu-berlin.de
j.dejesusdepinhopinhal@tu-berlin.de

Abstract

In April 2021 the European Commission (EC) proposed a new regulation to establish a regulatory structure for the risk assessment of Artificial Intelligence (AI) systems and applications. The intended goal of initiating a harmonised legal framework for the European Union (EU) poses new challenges in developing countermeasures for hate speech and fake news detection. This analysis investigates the implications of the proposed regulations on different automatic content moderation approaches such as flagging, blocking and filtering. The fuzzy nature of the risk categories causes major challenges for the risk categorisation task and leaves room for future improvements of the proposed regulations.

1. Introduction

Fake news and hate speech have been around for centuries but the emergence of the internet and social platforms has facilitated the spread of false information and hate speech globally [1]. Whereas the drivers of hate speech are multifaceted and range from personal insults to politically motivated spread of certain ideologies [2, 3], one of the most important motivations for the distribution of false information is financial gain [4] as well as the influence on elections [5]. Hereby, the profit comes primarily from advertisement services such as Google's AdSense. Moreover, the spread of false information created a more profound concern that the prevalence of fake news has increased political polarisation, undermined democracy and decreased trust in public institutions [6, 7].

The impact of hate speech is less clear as the desensitising impact of the frequent exposure to online hate speech on bystanders fuels prejudice against the victimised groups and decreases public sympathy towards them strengthening the *victim-blaming* phenomenon [8, 9]. Moreover, social psychology explains the emergence of hate and prejudice with the concept of out-groups as a potential existential and cultural threat. Thus, the motivation is to defend and preserve the existing social norms to keep the in-group itself intact [10]. Thus, both, hate speech and false information, have major implications on social and news platforms and online communication.

Over the past years, industry, governmental institutions and civil society have worked on developing policies, automated detection tools, and enforcement frameworks to tackle deceptive actors and false content online. Also, researchers have proposed techno-centric solutions to detect fake news [11] and hate speech content [12, 13, 14]. Although AI has evolved drastically over the past years, it is still prone to errors [15, 16], attacks [17, 18, 19], and biases [20, 21]. For example, [18] has shown that state-of-the-art technologies are vulnerable to simple adversarial attacks such as character-swap-based methods (e.g. changing "Trump" to "Trupm"). In response, in 2019, the EU released a report about automatic moderation of content on the internet which discourages countermeasures being purely based on AI without any human supervision. Further regulations for tackling fake news and hate speech were proposed in September 2020. The EU Commission announced their aim to expand the list of crimes to cover hate speech on the grounds of race, religion and national or ethnic origin. Later this year, the EC planned to release a common definition of hate speech even though some member states have expressed concerns that cultural differences may result in a threat to the right of freedom of expression under a common definition [22].

More recently, the EU has proposed the first legal framework on AI in April 2021 [23]. The proposal intends to create a uniform legal framework for AI within the EU. Here, safety-critical applications are addressed by a risk-based approach in order to classify AI systems and applications into four different risk categories. Each category has its own set of obligations which the providers of AI systems have to follow to protect users' fundamental rights. However, the categorization procedure is not a very transparent task as the risk categories are not distinct and there is no independent entity defined which supervises the classification procedure. Therefore, the contribution of this paper is the analysis of the applicability and consequences of the proposed regulations on AI in the domain of automatic content moderation (ACM) to verify the feasibility of the proposed risk-based approach.

2. Terminology and legal background

In the following, the terms used in this analysis will be defined and an overview of the existing ACM methods will be given.

2.1. Legal definition of fake news and hate speech

The definition of fake news and hate speech is a non-trivial task as both phenomena depend on each country's constitutional and legal structure, culture, political situation and level of public awareness on the problem of fake news and hate speech. Especially, the definition of fake news is challenging as fake news is often situated within a grey area of political expression encompassing both mis- and disinformation [24]. Moreover, the term is often abused to label opinions and information as fake when they do not comply with their viewpoint. Therefore, it is challenging to find appropriate countermeasures to tackle fake news while protecting fundamental rights, including freedom of expression (Article 11, defined in EU Charter of Fundamental Rights), data protection (Articles 7 and 8) and media pluralism.

Broadly, the term fake news can be distinguished between disinformation and misinformation whereas the latter is unin-

10.21437/SPSC.2021-10

tentionally false and inaccurate information shared by private persons [25]. In line with the EC High-Level Expert Group (HLEG), the term disinformation is defined as *"verifiable false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm"* [26]. This definition is taken as the basis for further analysis.

The only binding instrument for hate speech is the Counter-Racism Framework Decision [27] with obligations to make racist and xenophobic speech punishable under criminal law [28]. However, they do not consider online platforms and according to the Treaty on the Functioning of the European Union, online speech only falls into its legislative scope if it constitutes terrorist or child sexual abuse content [29]. The EU Code of Conduct on countering illegal hate speech online is a 'non-binding' document being voluntary for the online platforms to sign. It presupposes self- and co-regulatory monitoring measures and routinely publishes statistical information on hate speech successfully taken down on different platforms. Hate speech and disinformation are mainly regulated through individual regulation in the respective member states. For hate speech, Austria, Czech Republic, France, Italy, Spain and Germany passed specialized legislation to tackle hate speech online. The legal frameworks address fake news only within France, Spain and Germany. Germany's Network Enforcement Act (NetzDG) [30] states that service providers are responsible for illegal content shared on their platforms and are obliged to remove it within 24 hours when receiving a complaint. Hereby, disinformation and hate speech are both covered by the concept of illegal content. Content is deemed illegal if it can be classified in one of the many offences listed in the Panel Code covering domains such as state security, public order, sexual freedom and incitement to hatred or dissemination of unconstitutional symbols or groups. NetzDG also inspired French "Online Hate Observatory" [31] and Austrian "the Communications Platform Law" or KoPl-G [32].

One positive outcome is that the EU and other countries acknowledged that online disinformation and hate speech are severe matters and need to be addressed with respect to human rights. Even within the EU the approaches to tackle false information and hate speech online by different countries have few similarities due to cultural context, political situation, legal structure and the level of public awareness of the respective member states. This emphasises how difficult it is to find a common ground even on the EU level. Therefore, the risk-based approach is one promising direction to assess the harm an AI system can cause on an individual level irrespective of cultural differences and the legal context.

2.2. Automatic content moderation

Within machine learning (ML) approaches that are advancing towards AI, ACM technologies are audio-visual and textual analysis programs that are trained to moderate suspicious content online [24]. The recognition process itself is crucial to detect hate speech and false information to assist human judgement but does not imply any further action. Content recognition is an important component for content moderation tasks but will not be considered in the following analysis as the type of moderation mainly applies to the infringement of human rights. Thus, the focus is on ACM approaches and the classification and assessment of the respective risk category. The main ACM approaches encompass filtering, blocking, (de)prioritisation, flagging and disabling of spreaders which are described in the following [24].

Filtering or removing content is the most effective but also the most invasive countermeasure to tackle disinformation and hate speech. Hereby, filtering content online is an *ex ante* countermeasure that providers adopt in order to cope with the upload and posting of suspicious content. For this purpose, the content is scanned prior to the upload or post. The YouTube Content ID is one example that deploys *ex ante* filtering in order to protect the copyright and to give owners the possibility to decide to either block, monetise or track the video containing their work. On the other hand, removing content is an *ex post* countermeasure which is often a reaction to user requests also known as *notice and action procedure* [33].

Blocking of content is a widely used countermeasure that can be applied by users, email providers, search engines and social media platforms alike. Herewith, it differs from *ex ante* filtering such that the content is not removed while only the user's access is blocked. Blocking can be applied both, *ex ante* and *ex post*, where the user's awareness and ability to review the content and information provided is often a key component. Blocking can be applied in manifold ways such as *browser-based* blocking which is often used to block advertisements or cookies [24].

(De)prioritisation includes internal algorithmic downranking of user's content (e.g. shadow banning), reduction of advertising and prominence of content in users' feeds (e.g. Facebook's options to hide certain ads), demonetisation (Youtube and Twitch policy against keywords in the videos), and certain internet protocols (e.g. P2P). (De)prioritisation in the context of disinformation gives less prominence to content that contains false information, e.g. when content from media organisations or fact-checkers is given preference or shown next to false information [24]. For instance, deprioritisation is often used by platforms to tackle false information concerning COVID19 and vaccine hesitancy.

Flagging of content can be understood as the process of reporting doubtful and insulting content by other users, trusted flaggers, moderators and algorithms to the system which issues the flag after having verified the validity of the content. Another approach is visual tagging or blurring of the content which is potentially harmful. Facebook, Twitter and Youtube have both, machine and human-driven flagging systems, implemented at the core of their content moderation policies.

Disabling and suspension of accounts are temporary and permanent solutions to deal with users' abuse of terms of service and legislation (e.g. email providers, social media platforms, cloud services, and multi-player games). This is known as *jamming* and also applies to public and private groups and is often applied on Reddit and Facebook. The process is usually gradual with users first receiving *Appeals* and warnings. However, the reaction becomes more punitive and permanent reflecting the severity, frequency and persistence of violation.

Despite many propositions against using AI to tackle false narratives and hate speech online, there is also an increasing concern about the potential risks of automatic moderation of content. For example, in a recent study from the European Union Parliament [24], the trade-offs of using AI are examined. The authors raise concerns about the techno-centric solutions that propose automated detection, (de)prioritization, and removal by online intermediaries without human intervention. Moreover, they suggest that more independent, transparent, and effective appeals and oversight mechanisms are necessary to minimize the inevitable inaccuracies of AI. In the following, we will verify whether these advancements are addressed in the

risk-based approach of the proposal to regulate AI applications.

3. Risk assessment of countermeasures for fake news and hate speech detection

Although the opportunities and potential benefits of AI are manifold, certain AI systems lead to potential harm and infringement of rights, especially in the domains of recruitment, education, healthcare and law enforcement. Therefore, the EC formulated a proposal for a new regulatory framework on AI. This framework adopts a human-centric approach with the aim to facilitate the development of AI which ensures the protection of fundamental rights and user safety as well as trust and transparency. In order to adequately assess the influence of the proposed regulations on ACM, the following section introduces the main principles of the regulations and the impact on ACM concerning hate speech and fake news detection.

3.1. Proposal of harmonized rules on AI

The proposed regulatory framework applies to AI applications of both, the public and the private sector, for all systems placed on the EU market or in case EU citizens are affected. Herewith, it aims to provide guidance for AI developers, deployers and users alike by defining clear requirements and obligations regarding specific uses of AI systems. The risk-based approach has been chosen after extensive consultation with multiple stakeholders such as the High-Level Expert Group on AI. Hereby, the risk-based approach recognizes the benefits and potential of AI but at the same time also addresses possible dangers and risks of new AI applications and systems. Within the regulation, a broad definition of AI is given in *Article 3 (Definitions)* which includes any AI system generating outputs such as content, predictions, recommendations or decisions influencing the environment they interact with. More concretely, it applies to ML components, including supervised, unsupervised, reinforcement and deep learning but also logic- and knowledge-based approaches related to inductive logic programming, knowledge representation, inference and deductive engines. Furthermore, it also addresses statistical approaches such as Bayesian estimation and search optimization methods to fall under the definition of AI [34]. Therefore, the previously mentioned concepts of disinformation and hate speech fall into the broad definition of AI of the proposed regulation by the EC.

3.2. Risk-based approach

For the assessment of AI systems, a risk-based approach has been developed including four levels of risk:

1. **Unacceptable risk** includes all AI systems that pose a clear threat to the safety, rights and livelihoods of people. AI systems, ranging from social scoring by governments to toys that use voice assistance encouraging dangerous behaviour, will be banned from the European market.

2. **High-risk** AI systems falling into this category map to one of the following application areas: *critical infrastructure* with the potential to put the life and health of citizens at risk, *educational or vocational training* that might determine the access to education, *safety components of products* such as robot-assisted surgery, *employment* including CV sorting software for recruitment procedure, *essential private and public services* e.g. credit scoring systems, *law enforcement interfering with humans' fundamental rights* including the reliability of evidence, *migration, asylum and border control* including verification of the authenticity of travel documents and *administration of justice and democratic process*. Moreover, all *remote biometric identification systems* fall into the high-risk class and are also subject to strict obligations before they are allowed to be put on the market.

3. **Limited risk** include AI systems that require specific transparency obligations. One example mentioned in the proposal are chat-bots where it is required to show transparently whether the user is communicating with a bot or a human such that the users have the opportunity to make an informed choice to continue or stop their activities.

4. **Minimal risk** according to the proposal, most AI systems fall into the category of minimal risk. Given examples range from AI-enabled video games to spam filters.

3.3. Obligations

The different risk categories are related to certain obligations. AI applications falling into the **unacceptable-risk** category are prohibited. In case they map to the **high-risk** category, they need to follow a list of obligations defined in *Chapter 3* of the proposal before they are put on the market. This list includes, among others, compliance with requirements defined in *Chapter 2*. These demand the development of a risk management system, a data governance structure, technical documentation, record-keeping in the development phase, transparency and provisioning of information to users, human oversight, and accuracy, robustness and cybersecurity measures. Furthermore, the obligations address the development of a quality management system, technical documentation of high-risk systems and apply conformity assessment and compliance with registration obligations defined in *Article 51*. Moreover, service providers of **high-risk AI** need to collaborate with national competent authorities and demonstrate conformity with requirements defined in *Chapter 2* of the proposal by affix CE marking to AI systems to indicate the conformity with this regulation in accordance to *Article 49*. Obligations defined for AI applications falling into the category of **limited risk** need to follow only four obligations defined in *Title IV, Article 52*, namely: AI applications interacting with natural persons or emotion recognition systems are obliged to inform the user that they interact with an AI system. Also, users of AI systems that generate or manipulate images, audio or video content (deep fakes) must be informed about the artificial generation or manipulation of the displayed content.

3.4. Impact of risk categories on ACM

Applying the risk-based approach on the different ACM methods is not a straightforward method as the different categories overlap and the examples given in Annex III of the proposal are not very concrete.

The risk categorisation of **filtering** can be assigned to limited risk when **filtering** is applied *ex ante* and the content cannot be shown to potential users or consumers. Thus, transparency obligations need to be followed to communicate the flagging procedure as comprehensibly as possible. For *ex post* removing of content according to *notice and action procedure*, the reasons for the removal must be clearly stated and, therefore, it falls at least under the category of limited risk with transparency obligations. Nonetheless, removing content can also fall into the high-risk category when algorithms or humans make false removals and, thus, affect freedom of expression. In such scenarios removing content does fall into the high-risk category and needs to follow the obligations provided in *Chapter 3* of the

proposal. Similarly, **blocking** can be applied to different risk categories depending on the scenario. When users deploy ad blockers, it certainly does not contain any of the proposed risk categorisations. But when search engine filters prevent access to certain content, as with Google's rules against hate speech, **blocking** falls into the high-risk categorisation as it can negatively affect media pluralism and free speech. Thus, the blocked content need to be assessed carefully.

Algorithmic **(de)prioritisation** is key to user experiences that search engines and social media platforms offer, yet it is not straightforward in its implications for media pluralism and freedom of expression [24]. Hence, **(de)prioritisation** can be mapped to the limited risk category with transparency obligations. Nowadays, many social media platforms are used as a primary income source by bloggers and advertisers. For this reason, it can lead to unjustified substantial revenue loss (*demonetisation*) if the algorithm makes a mistake. It is vital to transparently communicate if the content (de)prioritisation has been initialised by a human or machine and the reasons behind the decisions, especially when negative effects on the users can be expected. More clarity can be achieved by the risk categorisation of **flagging**. **Flagging** can be assigned to the risk category limited risk with transparency obligations as it does not block or remove any content and, therefore, does not affect fundamental rights such as freedom of expression. Nevertheless, users should be able to easily find out why the content got flagged and if the message was flagged by a human or by a machine. The ACM method of **disabling** spreaders can be clearly assigned to the high-risk category which needs to follow the obligations stated in *Chapter 3* of the proposal. The countermeasure of **disabling** has significant implications on fundamental human rights such as freedom of expression, freedom of assembly [35] and the democratic process.

Overall, we argue that the ACM methods **(de)prioritisation** and **flagging** can be mapped clearly onto the limited risk category with transparency obligations. Also, for the case of **disabling** users to participate further on a platform, the high-risk category can be applied. For the ACM methods **blocking** and **filtering**, the risk categorisation is not so clear as it depends heavily on the scenario. Moreover, considering the current performance of automated systems, there is an obvious need for human oversight. This is addressed by the proposed regulation by the obligations listed in *Chapter 3* of the proposal which demands, among others, human oversight, transparency and provisioning of robustness and cybersecurity for high-risk AI systems. According to *Article 54* [23], regulatory sandboxes should provide a controlled environment facilitating the development, testing and validation of innovative AI systems for a limited time. Regulatory sandboxes can be useful to test further AI applications for different ACM methods. However, when affecting fundamental rights and being prone to biased results, they need to be tested under direct supervision and guidance by competent authorities defined in *Chapter 4* of the proposal before they can be applied in a broader scope.

4. Discussion

Risk management, human oversight and *ex ante* testing should facilitate the respect of fundamental rights by minimising the risk of erroneous or biased AI-assisted decisions in critical areas such as education and important services. In case of the infringement of fundamental rights, the proposal on harmonised rules for AI mentions effective redress for affected persons made possible by ensuring transparency and traceability of AI systems coupled with strong *ex post* controls. Yet, how these strong *ex post* controls can look like for different areas of application is not clear. Moreover, in *Annex III* of the proposal, social media is not considered under the high-risk category. Hereby, it is not clear whether ACM would also fall under the definition of social media or if it can be analysed as an independent component that can be applied in various forms. Moreover, the regulation aims at protecting users but does not consider *humans-in-the-loop* within these AI systems. Current working conditions of human content moderation labourers are dangerous for their mental well-being and health as found by [36]. Thus, a broader perspective on users of such AI systems needs to be considered.

Another source of concern is the *ex ante* risk self-assessment by providers of AI systems themselves and *ex post* enforcement for high-risk AI. Considering the level of generalisation of the definitions of the risk categories, they remain very open for interpretation. This could lead to the situation that most providers of AI systems aim to classify their applications and systems into the limited risk or minimal risk category even though fundamental rights might be affected. Additionally, bridging the gap between legal principles and technical implementation is a major barrier to develop ethically aligned AI systems. Major difficulties can be observed when the General Data Protection Directive was enforced, and yet, privacy infringements can still be detected [37]. This is not necessarily caused by bad intentions of developers but also results from the difficulty of fuzzy regulations where no clear guidance can be inferred in concrete scenarios. For example, "training, validation and testing data sets should be sufficiently relevant, representative, free of errors and complete in view of the intended purpose of the system" (*Recital 44*), which is very difficult to achieve. Basically, all the ACM methods and AI systems trained on user data do not comply with this requirement.

Moreover, the very broad definition of AI, which classifies most of the existing and also future software as AI the proposed regulation would cover, might hinder future development of AI systems that fall under the high-risk category and result in over-regulation [34]. The new legislation might not cause an improvement of risky AI systems but pushing the development of critical systems outside of EU borders and also its related ethical and legal problems [38]. Companies might be more willing to develop their products and services in other countries where legal constraints are less stringent or even absent. Technically, the proposed regulations will be challenging to translate into concrete guidelines. Furthermore it is expensive for service providers and sometimes even internationally problematic. In contrary, the challenge nowadays is no longer digital innovation but the governance of the digital sphere and shaping of digital sovereignty. These normative challenges have not been tackled so far and the EU is not simply ahead; it has no competition [38].

5. Conclusion

The analysis of the risk-based approach with respect to ACM methods shows that the proposed regulations suffer from major limitations. The regulation lacks clarity when fundamental rights are affected and in which risk category different applications fall which will have a major impact on the *ex ante* risk self-assessment of providers of AI. Nevertheless, the proposed regulations are a significant step towards a *digital constitutionalism* [39] where an *infosphere* [40] can create a space where citizens may live and work better and more sustainable.

6. References

[1] J. Suler, "The online disinhibition effect," *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, vol. 7, pp. 321–6, 07 2004.

[2] E. Barendt, "What is the harm of hate speech?" *Ethical Theory and Moral Practice*, vol. 22, no. 3, pp. 539–553, 2019.

[3] U. M. Ananthakrishnan and C. E. Tucker, "The drivers and virality of hate speech online," *Available at SSRN 3793801*, 2021.

[4] J. A. Braun and J. L. Eklund, "Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism," *Digital Journalism*, vol. 7, no. 1, pp. 1–21, 2019.

[5] I. S. Florence Davey-Attlee, "The fake news machine," 2020. [Online]. Available: https://money.cnn.com/interactive/media/the-macedonia-story/

[6] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

[7] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts, "Evaluating the fake news problem at the scale of the information ecosystem," *Science Advances*, vol. 6, no. 14, p. eaay3539, 2020.

[8] W. Soral, M. Bilewicz, and M. Winiewski, "Exposure to hate speech increases prejudice through desensitization," *Aggressive Behavior*, vol. 44, p. 136–146, 2018.

[9] L. Patterson, A. Allan, and D. Cross, "Adolescent bystander behavior in the school and online environments and the implications for interventions targeting cyberbullying," *Journal of School Violence*, vol. 16, no. 4, pp. 361–375, Oct. 2017.

[10] D. Gadd, "Aggravating racism and elusive motivation," *British Journal of Criminology*, vol. 49, 11 2009.

[11] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[12] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," 2020.

[13] S. Frenda, B. Ghanem, M. Montes, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on twitter," *Journal of Intelligent & Fuzzy Systems*, vol. 36, pp. 4743–4752, 05 2019.

[14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, and N. Farra, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," 03 2019.

[15] T. Scheffler, V. Solopova, and M. Popa-Wyatt, "The telegram chronicles of online harm," 03 2021.

[16] "Ai incident database," 2021. [Online]. Available: https://incidentdatabase.ai/

[17] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019.

[18] Z. Zhou, H. Guan, M. M. Bhat, and J. Hsu, "Fake news detection via nlp is vulnerable to adversarial attacks," *arXiv preprint arXiv:1901.09657*, 2019.

[19] T. Le, S. Wang, and D. Lee, "Malcom: Generating malicious comments to attack neural fake news detection models," *arXiv preprint arXiv:2009.01048*, 2020.

[20] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)," *Information Systems*, p. 101584, 2020.

[21] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320.

[22] C. Goujard, "Hate speech & hate crime – inclusion on list of eu crimes," 2021. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12872-Hate-speech-hate-crime-inclusion-on-list-of-EU-crimes_en

[23] "Proposal regulation: laying down harmonised rules artificial intelligence," 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[24] C. Marsden and T. Meyer, *Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism*. European Parliament, 2019.

[25] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe report*, vol. 27, pp. 1–107, 2017.

[26] H. L. E. G. on Fake News and O. Disinformation, "Report to the european commission on a multi-dimensional approach to disinformation," 2018. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

[27] Council of European Union, "Council framework decision (EU) no 913/jha," 2008. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32008F0913

[28] "Tackling disinformation and online hate speech: Eu and member state approaches, so far," *Democracy Reporting International*, 2020. [Online]. Available: https://democracy-reporting.org/wp-content/uploads/2021/01/Tackling-Disinformation-and-Online-Hate-Speech-DRI.pdf

[29] T. E. Parliament, "Consolidated version of the treaty on the functioning of the european union,part three: Union policies and internal actions, title v: Area of freedom, security and justice, chapter 4: Judicial cooperation in criminal matters, article 83," 2008. [Online]. Available: http://data.europa.eu/eli/treaty/tfeu_2008/art_83/oj

[30] Bundesministerium des Justiz und für Verbraucherschutz, "Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzwerkdurchsetzungsgesetz - netzdg)," 2017.

[31] L. C. supérieur de l'audiovisuel, "Décision n° 2020-435 du 8 juillet 2020 relative à la composition et aux missions de l'observatoire de la haine en ligne," 07 2020.

[32] Bundesministerium für Digitalisierung und Wirtschaftsstandort, "Bundesgesetz über maßnahmen zum schutz der nutzer auf kommunikationsplattformen," 10 2020.

[33] R. Barnes, A. Cooper, O. Kolkman, D. Thaler, and E. Nordmark, "Technical considerations for internet service blocking and filtering," *Request for Comments (RFC)*, vol. 7754, 2016. [Online]. Available: https://www.rfc-editor.org/info/rfc7754

[34] P. Glauner, "An assessment of the ai regulation proposed by the european commission," *arXiv preprint arXiv:2105.15133*, 2021.

[35] I. Siatitsa, "Freedom of assembly under attack: General and indiscriminate surveillance and interference with internet communications," *International Review of the Red Cross*, vol. 102, no. 913, pp. 181–198, 2020.

[36] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease, "The psychological well-being of content moderators," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI*, vol. 21, 2021.

[37] M. Hatamian, "Engineering privacy in smartphone apps: A technical guideline catalog for app developers," *IEEE Access*, vol. 8, pp. 35 429–35 445, 2020.

[38] L. Floridi, "The european legislation on ai: a brief analysis of its philosophical approach," *Philosophy & Technology*, pp. 1–8, 2021.

[39] G. De Gregorio, "The rise of digital constitutionalism in the european union," *International Journal of Constitutional Law*, vol. 19, no. 1, pp. 41–70, 2021.

[40] L. Floridi, *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford, 2014.