Cristina Frei and Christine Kaufmann

Content Moderation in Social Media – Artificial Intelligence to the Rescue?

"Spirits that I've cited
My commands ignore."
(Johann Wolfgang von Goethe, The Sorcerer's Apprentice)

Abstract: This chapter addresses the question of how human rights apply to the moderation of religion-related content on social media platforms, examines the function of content moderation in balancing the various human rights interests involved and the implications for regulation. It provides an overview of the most recent and relevant legal developments regarding artificial intelligence and content moderation, including international non-binding standards and binding instruments from the OECD, the United Nations, UNESCO, European Union, and the Council of Europe. The often complex interactions between freedom of expression and freedom of religion become apparent when applied to users who publish or are affected by religion-related content, as outlined in this chapter. Algorithmic content moderation faces limitations in balancing the various interests protected by human rights, which requires human intervention. The chapter highlights the importance of standards and regulations that provide tools for balancing different interests through remedies, procedural rights and safeguards.

Dieses Kapitel befasst sich mit der Frage, wie Menschenrechte auf die Moderation religionsbezogener Inhalte auf Social Media Plattformen angewandt werden, beleuchtet die Rolle der Content Moderation bei der Abwägung der verschiedenen betroffenen Menschenrechtsinteressen und die Konsequenzen für die Regulierung. Es bietet einen Überblick über die neusten und wichtigsten rechtlichen Entwicklungen in Bezug auf künstliche Intelligenz und Content Moderation, einschliesslich internationaler rechtlich nicht verbindlicher Standards und rechtsverbindlicher Instrumente der OECD, der Vereinten Nationen, der UNESCO, der Europäischen Union sowie des Europarats. Das oft komplexe Verhältnis zwischen Meinungsfreiheit und Religionsfreiheit zeigt sich bei der Anwendung auf Personen, die religionsbezogene Inhalte veröffentlichen oder davon betroffen sind, wie in diesem Kapitel erläutert wird. Algorithmic Content Moderation stösst bei der Herstellung eines angemessenen Ausgleichs zwischen den verschiedenen durch Menschenrechte geschützte Interessen an ihre Grenzen und erfordert menschliches Eingreifen. Dieser Beitrag unterstreicht die Bedeutung von Standards und Regulierungen, die Instru-

Open Access. © 2025 the author(s), published by De Gruyter. BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
https://doi.org/10.1515/9783111721729-008

mente zum Ausgleich verschiedener Interessen durch Beschwerdemöglichkeiten, Verfahrensrechte und Schutzmassnahmen bieten.

1 Introduction

1.1 Situating the Project

With religion increasingly turning to the digital space, new legal challenges emerge: Information travels at unprecedented speed, regardless of its content. In addition, once information is in the digital space it remains there as it is impossible to permanently remove it. At best, measures can be taken for information not to be found anymore.

Why is this important from a legal perspective? Because as in the analogue world, people can get hurt, for instance in their religious feelings, and their rights violated (Kirchschläger 2021, 186–190). But when do such incidents amount to an infringement of human rights such as freedom of religion? To what extent are postings on social media protected by freedom of expression?

To add another layer of complexity, social media platform providers are private actors, which, in contrast to states, are legally speaking not bound by international human rights. Indeed, increasingly states have started regulating the digital space, but unlike in the analogue world the reach and boundaries of national legislation are not that clear. Moreover, states have different concepts, for example, of what activities constitute blasphemy or how far freedom of opinion and freedom of religion go and these concepts collide in a space for which we do not have clear rules of jurisdiction.

In addressing (some of) these challenges, our project looks at the legal protection of religions in cyber space and at the legal limits of these rights. For this purpose, we started with an analysis of the general protection of human rights in the context of digital religious practices, identifying specific examples of restrictions such as “digital authoritarianism”. We then moved on to elaborate on the regulatory framework that specifically governs the moderation of religious content on digital platforms which is at the core of this contribution. Our research benefits in particular from the research of the URPP subproject on argument structures in the automatic detection of intolerance and extremism (see Schneider et al. this volume), which reflects the current state of technological developments in content moderation and provides important elements for the discussion of the legal issues.

Finally, an in-dept analysis of the regulatory framework governing the moderation of religion-related content on social media platforms against the

background of international human rights law is the topic of Cristina Frei's PhD project as part of the URPP. It addresses the tensions that arise between global, regional and national approaches as well as public and private regulations and suggests key elements for regulation.

1.2 What's Artificial Intelligence Got To Do with It?

Artificial intelligence (AI) is relevant in the context of protecting freedom of religion in the digital space in three regards: First, it plays a critical role in defining the content that users of social media will see. This may lead to situations where users are confronted with content that affects their religious feelings or human rights.

Second, artificial intelligence is a double-edged sword: It facilitates the speedy spread of news allowing for broadly available instant information. However, the spread of potentially harmful information can pose significant risks, including human rights infringements and abuses. On X (formerly Twitter), researchers observed a significant increase of harmful content including anti-semitic posts since Elon Musk took over the social media platform (Lima-Strong 2023). This surge occurred despite his claim that the "new Twitter policy is freedom of speech, but not freedom of reach" and that hateful content would be "deboosted" to the greatest extent possible (Musk 2022).

Third, effective regulation will require social media platform providers to monitor posted content, in some instances remove it or block corresponding accounts. Facebook, one of the most widely used social media platforms, has currently more than 3 billion active users (Dixon 2024). In such circumstances, content is monitored with the support of AI (Meta 2023, 9). Other platforms apply different approaches: While TikTok reported a high number of content moderation decisions based on its implementation of the relevant EU law, the Digital Services Act and stated that it moderates content almost entirely automatically, X declared that it rarely moderates harmful content and only uses non-automated methods (Drolsbach and Pröllochs 2024, 940–941). Accordingly, the EU Commission launched a procedure against X for violation of the DSA on various grounds including content moderation and requested more information from X on its content moderation activities and risk management (European Commission 2024). The procedure is still ongoing at the time of this writing.

An overview of recent legal developments with regard to artificial intelligence and content moderation of social media will set the stage for a discussion of the manifold interactions between AI and freedom of religion and freedom of expression and their impact on content moderation.

2 Legal Framework and Regulatory Approaches

Standards and regulations for AI have increasingly been developed to support the innovation potential and the related manifold societal benefits of AI while at the same time manage the associated risks. They take different forms and follow different regulatory approaches. Some are legally binding, while others are based on voluntary commitments (Figure 1). Regardless of their formal legal nature, they vary widely in both their geographical as well as substantive scope.

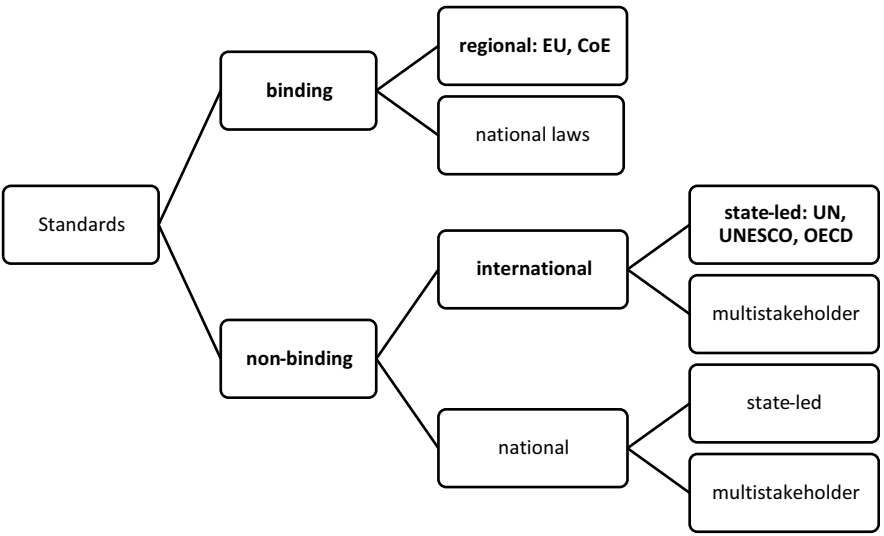

Figure 1: Legal nature of standards.

Generally, there are three different approaches to regulation: (1) Standards which contain requirements for the *actors*, i.e. platforms and users; (2) standards which focus on a specific *technology* and its impacts on human rights; and (3) regulations which serve as a comprehensive legal *framework* with principles for states to elaborate on in national law.

The following sections analyse selected instruments (highlighted in bold in Figures 1 and 2) with a view to their relevance for content moderation in the context of freedom of religion.

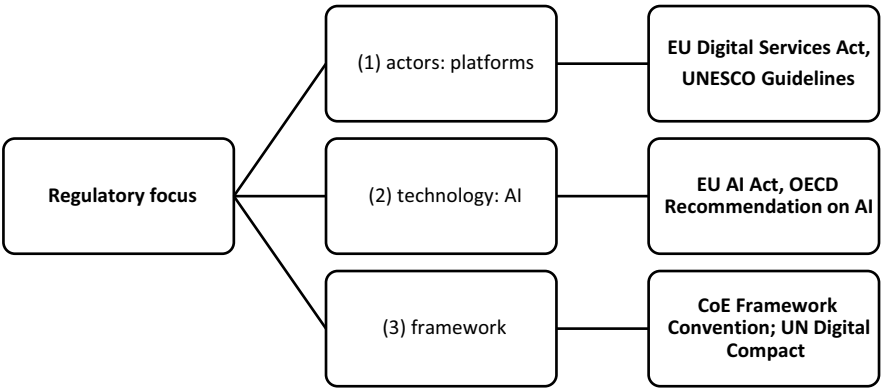

Figure 2: Regulatory focus.

2.1 International Non-Binding Standards

At the time of this writing in December 2024, no international legally binding standards on AI exists from organisations such as the UN and the OECD. However, several instruments developed within the framework of the UN and the OECD are relevant and have partially informed binding legislation (above Figure 1).

2.1.1 Organisation for Economic Cooperation and Development

The OECD adopted a *technology* specific instrument (Figure 2), the Recommendation on AI, with a set of principles for trustworthy AI in 2019 and updated it in May 2024 to reflect new technological and policy developments (OECD Recommendation on AI 2019; see Kaufmann 2024).

The Recommendation was the first intergovernmental standard on AI and developed in a multi-stakeholder process. It paved the ground for other regulations discussed in this section. The Recommendation is people-centered. Its ten principles should help governments, organisations and individuals to design and operate AI systems in a way that puts the interests of people first. They should also ensure that the developers and operators of AI-based systems and products are held accountable for their proper functioning. In sum, governments should ensure that AI systems are developed in alignment with values and laws, so that people can trust that their safety and privacy are put first (Gurría 2019).

This focus was further strengthened in the update in 2024 with its increased emphasis on people and the fight against abuse of AI by addressing the following

themes: (1) Inclusive growth, sustainable development and well-being, (2) respect for the rule of law, human rights and democratic values, including fairness and privacy, (3) transparency and explainability, (4) robustness, security and safety and (5) accountability.

Based on these principles, five governmental actions are recommended: (1) Investing in AI research and development, (2) fostering an inclusive AI-enabling ecosystem, (3) shaping an enabling interoperable governance and policy environment for AI, (4) building human capacity and preparing for labour market transformation and (5) international co-operation for trustworthy AI.

Currently, 46 states and the European Union signed the Recommendation.

2.1.2 United Nations Global Digital Compact

In September 2024, the Global Digital Compact was adopted by the UN General Assembly. It covers the impacts of the digital transition on society in general and aims at ensuring that the digital transition is in line with the Sustainable Development Goals. Its objectives include enhancing the international governance of artificial intelligence for the benefit of humanity (UN Global Digital Compact 2024, para. 7.5, para. 50), and co-operation for trustworthy technologies, including AI, to advance a responsible, accountable, transparent and human-centric approach to their life cycle (UN Global Digital Compact 2024, para. 8(i)). This is complemented by an "urgent call" on technology companies and developers to take measures against potential harms induced by AI-enabled content, including hate speech and discrimination (UN Global Digital Compact 2024, para. 35(c)). States acknowledge the need for international coordination of AI standards, but no specific measures apart from convening a high-level meeting for reviewing the Global Digital Compact have been agreed. With its broad approach the Digital Compact qualifies as a *framework* instrument referred to in Figure 2.

2.1.3 UNESCO Recommendation on the Ethics of AI

The UNESCO Recommendation on the Ethics of AI adopted by its then 193 member states in 2021 (UNESCO Recommendation on AI 2021) follows the human rights by design approach as it states ten core principles for a human-rights centred approach to the ethics of AI and translates them into key areas for policy action (Ramos, Squicciarini, and Lamm 2024, 33–34). Among the ten principles is the right to privacy, which is to be protected throughout the whole AI lifecycle. The recommendation also states that AI systems should only be used to the extent nec-

essary to achieve a legitimate aim (proportionality) and be subject to a risk assessment to prevent harms resulting from their use. States are called upon to ensure that AI does not "foster disruption of freedom of expression and access to information" (UNESCO Recommendation on AI 2021, para. 66). Strong emphasis lies on the ultimate human responsibility and accountability. In other words, responsibility for human rights risks cannot be delegated to technology but remains with humans and, ultimately, the state. In addition, the Recommendation provides two specific instruments for implementation addressing both member states and non-state actors involved in AI projects.

In sum, the Recommendation – somewhat implicitly – acknowledges potential conflicting human rights interests which need to be balanced. Its regulatory focus is mainly on the *technology* (Figure 2).

2.1.4 UNESCO Guidelines for the Governance of Digital Platforms

The UNESCO Guidelines for the Governance of Digital Platforms adopted in 2023 focus on the *actors* and on the role of digital platforms and providers in protecting freedom of expression (Figure 2) (UNESCO Guidelines 2023). The Guidelines specifically address the lack of corporate transparency, accountability and due diligence, and call for a human rights-centred approach in state regulation. An important feature of any policy action is a multi-stakeholder approach. Five principles call on platforms first to conduct human rights due diligence in accordance with the UN Guiding Principles on Business and Human Rights (UNESCO Guidelines 2023, principle 1, paras. 85–90), and second to adhere to international human rights standards (UNESCO Guidelines 2023, principle 2, para. 91). Due diligence includes platform design and content. For this purpose, the Guidelines elaborate not only on the design process (UNESCO Guidelines 2023, principle 2, paras. 92–93), but also on *content moderation and curation policies and practices* (UNESCO Guidelines 2023, principle 2, paras. 94–100), including human content moderation (UNESCO Guidelines 2023, principle 2, paras. 101–102), and the use of automated systems for content moderation and curation, i.e. AI systems (UNESCO Guidelines 2023, principle 2, paras. 103–109). Generally, platforms should act in a transparent manner (UNESCO Guidelines 2023, principle 3, paras. 111–118), including with regard to the implementation of content moderation, curation policies and practices, and users should be notified when and on what grounds their content is removed (UNESCO Guidelines 2023, principle 2, para. 110, principle 3, paras. 115.e-j). Moreover, the Guidelines aim to hold platforms accountable to relevant stakeholders, which includes user reports as well as appeal and redress mechanisms (UNESCO Guidelines 2023, principle 5, paras. 123–129). The Guidelines conclude with a set of context-specific provisions

related to vulnerable and marginalised individuals (UNESCO Guidelines 2023, context-specific provisions, para. 130), situations of armed conflict, crises and emergencies (UNESCO Guidelines 2023, context-specific provisions, paras. 142–144), as well as to the specific issue of integrity of elections (UNESCO Guidelines 2023, context-specific provisions, paras. 131–141).

While – and maybe because – the Guidelines are not binding, they contain fairly detailed principles on content moderation and curation. It remains to be seen to what extent they will be taken up in future legislation.

2.2 European Union: Legally Binding Instruments

The first comprehensive and legally binding framework for specifically regulating AI and thus a regulation with a focus on *technology* (Figure 2) is the European Union's *Artificial Intelligence Act (AIA)* (EU Artificial Intelligence Act 2024), which entered into force in 2024. It needs to be read together with the *Digital Services Act (DSA)* of 2022 (EU Digital Services Act 2022).

The DSA does not focus on a specific technology (such as AI) but on the *actors* (Figure 2), i.e. online platforms. On the one hand, it requires social media platforms to rapidly remove illegal content; on the other, it attempts to balance the protection against illegal content with users' freedom of expression. For this purpose, platforms are required to counter illegal content. Moreover, they must make transparent how their content moderation and their algorithmic recommender systems work. In addition, large platforms have to conduct a risk assessment for their products, including risks to human rights. In the interest of protecting freedom of expression, users have the right to challenge content moderation decisions under the DSA. The DSA does not contain specific provisions on freedom of religion or focus on AI only but applies a general, technology-neutral approach which includes all human rights.

In contrast, the AIA addresses the *specific* features and the associated risks of artificial intelligence. It applies a "human rights by design" approach by requiring developers and deployers to take potential risks related to specific uses of AI into account. The goal is to foster responsible AI systems by defining clear legal requirements for AI systems according to the level of risk. Accordingly, the AIA bans AI systems with *unacceptable risks* such as threats to the safety, livelihood and rights of people outright, and strictly regulates those with *high risk*. AI systems with *limited risks* are subject to transparency requirements; and the use of *no or minimal risk* AI systems is free.

The regulatory approach of the European Union is particularly interesting because it aims at addressing the tension between protecting society against unac-

ceptable AI-related risks and safeguarding freedom of expression and the innovative potential of AI technology. When discussing algorithmic content moderation in the context of religion-related content (section 3), we therefore need to combine the approaches of the DSA and the AIA which in sum require platform providers to conduct *specific AI risk analysis* and at the same time give platform users rights to protect their human rights.

2.3 Council of Europe: Framework Convention on Artificial Intelligence

With its Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law (CoE Framework Convention on AI 2024), the Council of Europe pursues a different strategy. It is the first international legally binding treaty in the field of AI and qualifies as a framework instrument (Figure 2). It requires states to take measures to ensure that activities within the lifecycle of AI systems comply with the fundamental principles of human dignity and individual autonomy, equality and non-discrimination, respect for privacy and personal data protection, transparency and oversight, accountability and responsibility, reliability and safe innovation. Respecting these principles requires balancing different interests. The Convention does not elaborate on detailed criteria for such a balancing exercise but provides a set of instruments in the form of remedies, procedural rights and safeguards that states need to establish. It also defines general requirements for the risk and impact management. While states are signatories to the Convention and thus the primary duty holders, the Convention covers the use of AI systems by both public authorities and private actors. According to the nature of the Convention as a framework states define the specific application to private actors (CoE Framework Convention on AI 2024, Article 3(1)).

It is important to note that the Framework Convention builds on existing international human rights law and on the Council of Europe's standards on democracy and the rule of law. The purpose of the Convention is to complement existing instruments to ensure that they also apply for the use of AI systems and thus beyond the analogue world. What this application entails with regard to the moderation of religion-related content will be elaborated in the next section. This includes an analysis of the related risks to the protection of human rights and the role of specific regulatory requirements.

3 Interactions of AI with Religion-Related Content – Selected Aspects

3.1 Moderation of Religion-Related Content and the Role of AI

Religion-related content may include a recording of a church mass on YouTube, a young woman sharing knowledge about her religion in a live-stream on Instagram (Fazel 2023), or a Salafi online influencer discussing theological issues in short videos on TikTok (Klapp 2023, 11). The platform's algorithms define the display of such content in newsfeeds, to whom it is suggested and how easily and widely it is accessible. An example are the high view and engagement numbers on Facebook for AI-generated images of Jesus rendered as a crab, which according to research are at least partially attributable to the recommendation algorithm (DiResta and Goldstein 2024).

Platforms collect user data by AI systems, enabling, *inter alia*, customised content on the user interfaces, which leads to higher user engagement and higher revenues for the platforms (Narayanan 2023). The resulting personalised online experiences may facilitate access to relevant content, including in terms of religious practice.

Religion-related content may be harmful or misleading and spread at high speed and widely on social media platforms. A striking example is the high number of Facebook posts inciting violence and discrimination against Muslims, especially Rohingyas, in Myanmar in 2017 (see UN Human Rights Council 2018, para. 1352). Accordingly, social media platforms have been called upon to adopt new or enhance existing policies and moderate content accordingly. Given the high number of accounts and users, content moderation by humans only will likely not suffice. As a result, social media platforms increasingly rely on algorithms for identifying religion-related content that contradicts the content policies of the platforms and/or national or international law. Such content may then be taken down, demoted or a warning note be given to the respective user.

An example for the important role which AI can play in content moderation is a video on Instagram accusing a Pakistani political candidate of having "crossed all limits of kufr" which was identified by Meta's AI-based High Risk Early Review Operations (HERO) system as a potential risk and forwarded to Meta's (human) policy experts (Meta Oversight Board 2024b). The experts considered the terminology as an accusation of blasphemy under Pakistani law based on its insinuation that the political candidate would believe in more than one God or equated someone with God. Due to the risk of offline harm that such an accusation could provoke in Pakistan, the video was taken down.

3.2 Challenges and Limitations of Algorithmic Content Moderation

However, algorithmic content moderation is not a panacea as it comes with its own challenges and limitations. Known issues include the inability of AI technology to recognise nuanced elements in some cases, the constraints with less widely used languages and contextual word interpretation, the difficulty of identifying prohibited content in certain forms of expression (such as voices, images or symbols), or the problem of coping with users' evasion tactics (for further details, see Hatano 2023, 149–150) and coded language that makes it difficult to detect whether someone is intentionally inciting violence on the basis of religion (Sanchez v. France 2021, para. 68). The use of implicit hate speech poses a problem particularly for the identification of antisemitism (Becker and Troschke 2023).

The UNESCO Guidelines (above 2.1.4) primarily acknowledge the challenges related to linguistic and cultural particularities of content (UNESCO Guidelines 2023, principle 2, paras. 94–95) and recommend regular external audits with binding follow-up steps of the automated and human tools used, including reviewing their precision, accuracy and linguistic capacity (UNESCO Guidelines 2023, principle 2, para. 103). In addition, the DSA (above 2.2) requires the providers of very large online platforms to include linguistic aspects in their risk assessment (EU Digital Services Act 2022, Article 34).

Given these limitations of AI systems, the incomplete removal of prohibited content remains a challenge. Members of religious minorities are often disproportionately affected by false negatives, i.e. content that is not identified by the platforms' moderation systems and therefore remains visible. For instance, the lack of Hindi and Bengali hate speech classifiers resulted in harmful content targeting Muslims in India not being removed on Facebook (Saaliq and Pathi 2021). Moreover, reports suggest that Facebook's algorithmic content moderation may have amplified the distribution of content inciting violence and discrimination against Rohingya in Myanmar by recommending it to more users (Amnesty International 2022, 45–48). Such content can provoke further hateful reactions in the form of likes, comments and shares. From a platform's perspective higher user engagement is desirable because it contributes to higher revenues. With regard to extremist content, recent studies concerning YouTube show that the adjustment of the recommendation system in 2019 was successful with regard to people who were not subscribers of extremist channels; yet the risk of algorithmically influenced radicalisation cannot be completely prevented (Chen et al. 2023, 8).

AI systems may also "overperform" and excessively remove or demote religion-related content or wrongly treat it as prohibited content (see Ashraf 2022, 773). Reports concerning the removal of content with the hashtag #AlAqsa on In-

stagram due to the confusion between the mosque revered by Muslims and a sanctioned organisation indicate that over-removal may also be a concern when it comes to religious content and would warrant further research (Mac 2021).

Finally, the risk of bias, i.e. inaccurate or unfair results due to the data that algorithms are trained on or use, is of particular relevance in the context of religion-related content. Bias in algorithmic systems can lead to discriminatory decisions, including in content moderation. Members of religious minorities or non-Christians are at higher risk that content which they publish on social media platforms is falsely classified by algorithms as prohibited, and thus deleted or demoted. A report by the European Union Agency for Fundamental Rights (FRA) demonstrates how the sentences "I am Muslim" and "I am Jew" are much more likely to be categorised as offensive than "I am Christian" (European Union Agency for Fundamental Rights 2022, 11). Accordingly, the UNESCO Guidelines (above 2.1.4) stipulate that the external audits mentioned above should check content moderation tools for possible bias or discrimination (UNESCO Guidelines 2023, principle 2, para. 103). Potential discrimination is also part of the risk assessment required by the DSA (above 2.2) (EU Digital Services Act 2022, Article 34).

In sum, algorithms are needed to moderate content on large social media platforms to effectively remove prohibited content, including hate speech and extremism, both from and against religious individuals. However, human rights compliant content moderation cannot be delegated entirely to AI. The challenge lies at the interface between humans and AI (addressed in the OECD Recommendation above 2.1.1): How can bias induced by humans be prevented? And how can the different human rights affected by AI supported content moderation be balanced?

4 Freedom of Expression, Religious Freedom and the Need for Protection – It's Complicated!

4.1 Social Media Platforms' Human Rights Responsibilities

Decisions about religion-related content by social media platforms, including by means of algorithmic content moderation, have an impact on human rights. While these platforms are not directly bound by international human rights treaties, which primarily address states, international instruments, such as the UN Guiding Principles on Business and Human Rights and the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct, expect businesses to respect international human rights and recall that states have a legal obligation

to ensure corporate respect of human rights. This accepted corporate human rights responsibility forms the starting point for specific content moderation rules.

4.2 Protecting Freedom of Expression and Religious Freedom, But Not Religion as Such

As a result, content moderation applied to social media platforms has to be aligned with human rights. Particularly, the right to freedom of expression and the right to freedom of religion are affected in the context of religion-related content. These rights apply to both the person producing and/or distributing the content as well as the recipients.

When users publish or access content on a social media platform, they generally exercise their right to freedom of expression. Article 19 of the International Covenant on Civil and Political Rights (ICCPR) protects the right to freedom of opinion and expression, including the freedom to seek, receive and impart information and ideas of all kinds – religious content included – regardless of frontiers. The drafters of the Covenant made it clear that the type of media is irrelevant, thus the protection extends to social media.

Anyone who disseminates or consumes the content as part of her or his religion or belief (theistic, non-theistic or atheistic) is also protected by freedom of religion. Article 18 of the ICCPR guarantees the right to freedom of thought, conscience and religion. It encompasses the freedom to manifest one's religion or belief in worship, observance, practice and teaching, whether individually or collectively in both public and private settings. As the scope of freedom of expression includes religious content, public freedom of religion and belief typically constitutes a subset of freedom of expression (Schabas 2019, 518).

Social media platforms that moderate, respectively restrict such activities will inevitably interfere with users' rights. Conversely, content moderation can be essential for protecting the rights of other users and affected third parties. An infringement of religious freedom may be assumed in situations where harmful content directed at members of a religious community is fostering a climate of public hostility to the extent that these members can no longer safely practice their religion publicly (Bielefeldt, Ghanea and Wiener 2016, 492). In addition, religion-related content can affect other peoples' rights, for instance by inciting violence, undermining gender equality or conflicting with protection from discrimination on the ground of sexual orientation.

The rights to freedom of expression and freedom of religion find their limits not least in the protection of the fundamental rights of others. The justification

for an interference with freedom of expression often implies considering the freedom of religion of others and *vice versa*. Publishing content which denigrates an object of religious veneration and thus may offend the religious feelings of others, is in principle protected by freedom of expression. This raises the question, whether the protection of the religious freedom of those who feel offended justifies a restriction on freedom of expression and, consequently, the removal of content.

Efforts to protect religious feelings from "offensive speech", including initiatives to establish a concept of "defamation of religions" at the international level (Langer 2014), contributed to the view that the right to freedom of expression and the right to freedom of religion are contradictory human rights concerns (Bielefeldt, Ghanea, and Wiener 2016, 483). Moreover, it was argued that freedom of religion would entail a right to respect for one's religious feelings or the protection of one's own religion from criticism, ridicule or insult, whereas freedom of expression would require the unlimited possibility to express oneself freely (Temperman 2008, 527). Yet, the international human rights framework does not assume an inherent conflict between these two rights.

While the interplay between these rights remains complex, case law and academic literature provided important guidance in defining their limits. Article 18 as well as further international human rights norms do not protect religion as such (Langer 2014, 123–142). Furthermore, the UN Human Rights Committee made clear that the scope of Article 19 extends to deeply offensive views (UN Human Rights Committee 2011, para. 11). Based on the decisions Ross v. Canada (Ross v. Canada 2000, para. 11.5) and Faurisson v. France (Faurisson v. France 1996, para. 9.6), it can be said that the UN Human Rights Committee, in contrast to the European Court of Human Rights, only accepts offenses against religious feelings as grounds for a restriction of freedom of expression to the extent that they meet the severity threshold of Article 20 Paragraph 2 ICCPR (Petzhold 2015, 216–217). This provision prohibits hate speech that amounts to the advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. In the same vein, General Comment No. 34 states that prohibitions of displays of lack of respect for a religion, including blasphemy laws, are not compatible with the ICCPR unless the particular conditions of Article 20 Paragraph 2 are met (UN Human Rights Committee 2011, para. 48).

4.3 Adequate Balancing of Human Rights by Algorithms?

These human rights-based ground rules for the moderation of religion-related content set the standard for balancing different human rights interests. When

freedom of expression and freedom of religion are implicated, the proportionality of an interference and the balancing of rights are central. The scope of human rights protection therefore needs to be translated into instructions for the algorithms which are needed for effective content moderation on large social media platforms. Determining which interest should be given more weight in individual cases is already a complex task for humans and even more so for AI systems which build on data provided by humans (see Prem and Krenn 2024, 485). In particular, the proportionality test, which includes assessing whether a restriction is necessary for and proportionate to the aim pursued, cannot be encoded in algorithms easily (Lennartz and Kraetzig 2022). For the time being, algorithmic decision making can therefore not guarantee the adequate balancing of rights (Peukert 2021).

Assessing the severity of a certain expression and accordingly the decision whether it should be prohibited in accordance with Article 20 Paragraph 2 ICCPR is challenging. The so-called *Rabat Plan of Action* summarizes the results of a series of expert workshops on the prohibition of incitement to national, racial or religious hatred, organized by the UN Office of the High Commissioner for Human Rights (OHCHR) (UN High Commissioner for Human Rights 2013). While it is not legally binding, it provides guidance to which international courts, bodies and experts regularly refer. It recommends using a six-part threshold test to consider the elements of context, speaker, intent, content and form, extent of the speech act und likelihood of harm. This requires an assessment of the *context* which includes the social and political circumstances. Furthermore, the *speaker*'s position or status in society, such as for instance the influence of religious leaders, needs to be duly taken into account. Importantly, *intent* is required for advocacy and incitement, negligence or recklessness is not enough to be considered prohibited behaviour. *Content* needs to be analysed with a view to the degree of provocation and directness, its form, style or the nature of arguments. With a view to human rights risks, the Rabat Plan emphazises the importance of analysing the *extent* of a published content, i.e. its reach, public nature, the applied means of dissemination, as well as the frequency and quantity. The *harm* must be likely in terms of reasonably probable and not only hypothetical. In practice, determining the elements of intent and likelihood of harm are highly complex, especially in the context of content moderation, partly because not only the original authors need to be included in the analysis but also people who share the content (Benesch 2020, 110).

The Meta Oversight Board already applied the Rabat Plan of Action test in various decisions (Meta Oversight Board 2024a). In algorithmic content moderation, however, the current limitations of AI technology in recognising nuanced elements in some cases may prevent the reliable implementation of the six factors

(Hatano 2023, 149). This may also lead to situations where content which does not meet the threshold criteria is deleted or demoted, and thus an over-removal. Often, algorithmic content moderation does not give a clear answer on where the line is drawn or how diverging interests are balanced.

5 Conclusion

Our research can be summarized in three main conclusions which relate to (1) the applicability of human rights to religion-related content on social media, (2) the balancing of different human rights interests involved and (3) to the consequences for regulation and eventually the role of the state.

(1) *Human rights apply to the digital space* as well as to the analogue world given the impact of digital technology on human rights. In particular, social media platforms can support human rights by allowing for fast and wide publication and dissemination of content, including religion-related. This potential comes with the *platforms' responsibility to respect* users' human rights and address potential risks. Human rights risks are particularly prevalent in the context of religion-related content.

(2) Addressing human rights risks associated with religion-related content requires a careful *balancing of different human rights interests*. An important instrument for this purpose is content moderation.

The applicable laws and the platforms' content policies are often of a rather general nature and leave broad room for discretion and interpretation. For religion-related content, such a balancing test is particularly challenging and depends on the interpretation of the right to freedom of expression and freedom of religion. Given the size of large social media platforms algorithms are needed for effective content moderation. The balancing of conflicting interests and ensuring the proportionality of restrictions is a key aspect of human rights law. Current AI based systems which identify, match and predict content, or classify content into one of several categories (see Gorwa, Binns and Katzenbach 2020, 4–5) are not equipped to conduct proportionality tests. Therefore, the balancing of interests cannot be delegated to AI but requires human intervention.

(3) The challenges of algorithmic content moderation are reflected in *standards and regulations for AI*. The binding and non-binding instruments, such as the EU's DSA and AIA or the UNESCO Guidelines for the Governance of Digital Platforms, elaborate on the responsibility of social media platforms to respect

human rights. This includes conducting human rights due diligence to identify, prevent, mitigate these risks and account for how they are addressed. In addition, appropriate measures are required to assess and mitigate the risks that the use of algorithmic content moderation entails. With regard to religion-related content, for example, social media platforms should analyse how algorithmic content moderation may automatically remove certain religion-related content, potentially resulting in the over-policing of certain religious minorities (Ashraf 2022, 773). It will be key that such risk assessments draw on the expertise of social scientists, including in the field of digital religion. The DSA requires platforms to grant researchers access to relevant data. The interdisciplinary URPP Digital Religion(s), which includes researchers from the disciplines of computational linguistics, religious studies and media and communication studies, can serve as a forum for further clarifying the notion of religion and offering different perspectives and insights on the protection of religious minorities which would benefit new regulation.

New standards and regulations for AI and digital platforms further define social media platforms' responsibility to respect human rights, including freedom of expression and freedom of religion but do not specifically stipulate how the complex content moderation decisions should be taken. Instead, they provide important instruments for balancing different interests through remedies, procedural rights and safeguards. Research and academia will play an important role in supporting states and platforms by providing more data on the human rights impacts in the context of religion-related content and in translating abstract human rights concepts into criteria that can be operationalized for content moderation in practice.

Our research is a first step, as it identifies key elements for addressing potential risks and paves the ground for further human rights-oriented interdisciplinary work.

Bibliography

Amnesty International, *The Social Atrocity, Meta and the Right to Remedy for the Rohingya*. 29.09.2022. https://amnesty.org/en/documents/asa16/5933/2022/en/.

Ashraf, Cameran. 2022. „Exploring the Impacts of Artificial Intelligence on Freedom of Religion or Belief Online." *The International Journal of Human Rights* 26, 757–791.

Becker, Matthias J., and Hagen Troschke. 2023. „Decoding Implicit Hate Speech, The Example of Antisemitism." In *Challenges and Perspectives of Hate Speech Research*, edited by Christian Strippel, Sünje Paasch-Colberg, Martin Emmer, and Joachim Trebbe, 335–352. Berlin: Böhland & Schremmer Verlag.

Benesch, Susan. 2020. „But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies." *Yale Journal on Regulation Bulletin* 38, 86–111.

Bielefeldt, Heiner, Nazila Ghanea, and Michael Wiener. 2016. *Freedom of Religion or Belief, An International Law Commentary*. New York: Oxford University Press.

Chen, Annie Y., Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. 2023. „Subscriptions and External Links Help Drive Resentful Users to Alternative and Extremist YouTube Channels." *Science Advances* 9, 1–13.

DiResta, Renee, and Josh A. Goldstein. "How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth." *Stanford University, Stanford Internet Observatory*, 18.03.2024. https://cyber.fsi.stanford.edu/io/publication/how-spammers-scammers-and-creators-leverage-ai-generated-images-facebook-audience.

Dixon, Stacy Jo. „Facebook – Statistics & Facts." *statista*, 20.03.2024. https://www.statista.com/topics/751/facebook.

Drolsbach, Chiara Patricia and Nicolas Pröllochs. 2024. "Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database." *WWW'24: Companion Proceedings of the ACM Web Conference 2024*, 939–942.

European Commission. "Commission Requests Information from X on Decreasing Content Moderation Resources under the Digital Services Act." *Press Corner*, 08.05.2024. https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-x-decreasing-content-moderation-resources-under-digital-services.

European Union Agency for Fundamental Rights. "Bias in Algorithms – Artificial Intelligence and Discrimination." *Publications Office of the European Union*, 08.12.2022. https://fra.europa.eu/en/publication/2022/bias-algorithm.

Fazel, Virginie. „Religious Quiz on Instagram: Exploring 'Horizontal' Religion." *University of Zurich, Digital Religion(s): Der Blog*, 28.03.2023. https://www.uzh.ch/blog/digitalreligions/2023/03/28/religious-quiz-on-instagram-exploring-horizontal-religion/.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7, 1–15.

Gurría, Angel. "Launch Ceremony for the Adoption of the OECD Recommendation on Artificial Intelligence," *OECD Web archive*, 24.05.2019.https://web-archive.oecd.org/2019-05-24/520712-launch-ceremony-for-adoption-of-oecd-recommendation-on-ai-paris-may-2019.htm.

Hatano, Ayako. 2023. "Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation." In *The Australian Year Book of International Law*, edited by Esmé Shirlow and Donald R. Rothwell, 127–156. Leiden: Brill.

Kaufmann, Christine. 2024. "Neue OECD-Instrumente zu künstlicher Intelligenz, Wege zu vertrauenswürdiger künstlicher Intelligenz." *Jusletter IT*, 1–11.

Kirchschläger, Peter G. 2021. *Digital Transformation and Ethics, Ethical Considerations on the Robotization and Automation of Society and the Economy and the Use of Artificial Intelligence.* Baden-Baden: Nomos.

Klapp, Marcel. 2023. "'That's Where I Get Reach!' Marketing Strategies of a Salafi Influencer on YouTube and TikTok." *Journal of Muslims in Europe* 13, 3–25.

Langer, Lorenz. 2014. *Religious Offence and Human Rights, The Implications of Defamation of Religions*. Cambridge: Cambridge University Press.

Lennartz, Jannis and Viktoria Kraetzig. "Filtering fundamental rights." *Verfassungsblog*, 05.10.2022. https://verfassungsblog.de/filtering-fundamental-rights/.

Lima-Strong, Cristiano. “Antisemitic Tweets Soared on Twitter after Musk Took Over, Study Finds.“ *The Washington Post*, 20.03.2023. https://www.washingtonpost.com/politics/2023/03/20/antisemitic-tweets-soared-twitter-after-musk-took-over-study-finds/.

Mac, Ryan. “Instagram Censored Posts About One Of Islam’s Holiest Mosques, Drawing Employee Ire.“ *BuzzFeed*, *BuzzFeed News*, 12.05.2021. https://buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque.

Meta. “Human Rights Report, Insights and Actions.“ *Meta’s Annual Human Rights Report*, 2023. https://humanrights.fb.com/wp-content/uploads/2024/09/2023-Meta-Human-Rights-Report.pdf.

Meta Oversight Board. *Posts That Include ‘From the River to the Sea’*. 04.09.2024a. https://oversightboard.com/decision/bun-86tj0rk5/.

Meta Oversight Board. *Pakistan Political Candidate Accused of Blasphemy*. 19.09.2024b. https://oversightboard.com/decision/ig-wxhs8uei/.

Musk, Elon (@elonmusk). „New Twitter Policy is Freedom of Speech, but not Freedom of Reach.“ *X*, 18.11.2022. https://x.com/elonmusk/status/1593673339826212864.

Narayanan, Arvind. “Understanding Social Media Recommendation Algorithms.“ *Columbia University, Knight First Amendment Institute*, 09.03.2023. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

Petzhold, Bianca. 2015. *Die “Auffassungen” des UN-Menschenrechtsausschusses zum Schutze der Religionsfreiheit*. Tübingen: Mohr Siebeck.

Peukert, Alexander. “Five Reasons to be Skeptical About the DSA.“ *Verfassungsblog*, 31.08.2021. https://verfassungsblog.de/power-dsa-dma-04/.

Prem, Erich, and Brigitte Krenn. 2024. “On Algorithmic Content Moderation.“ In *Introduction to Digital Humanism*, edited by Hannes Werthner, Carlo Ghezzi, Jeff Kramer, Julian Nida-Rümelin, Bashar Nuseibeh, Erich Prem, and Allison Stanger, 481–493. Cham: Springer.

Ramos, Gabriela, Mariagrazia Squicciarini and Eleonora Lamm. 2024. “Making AI Ethical by Design: The UNESCO Perspective.“ *Computer* 57, 33–43.

Saaliq, Sheikh, and Krutika Pathi. “Facebook Dithered in Curbing Divisive User Content in India.“ *The Associated Press*, 24.10.2021. https://apnews.com/article/coronavirus-pandemic-technology-business-media-religion-74175aa6f2cb50fc6fb1aedda11b2c6c.

Schabas, William A. 2019. *U.N. International Covenant on Civil and Political Rights, Nowak’s CCPR Commentary*. Kehl: N.P. Engel.

Temperman, Jeroen. 2008. “Blasphemy, Defamation of Religions and Human Rights Law.“ *Netherlands Quarterly of Human Rights* 26, 517–545.

UN High Commissioner for Human Rights. *Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred That Constitutes Incitement to Discrimination, Hostility or Violence*. Appendix, 11.01.2013. http://undocs.org/en/A/HRC/22/17/Add.4.

UN Human Rights Committee. *General Comment No. 34*. 12.09.2011. http://undocs.org/en/CCPR/C/GC/34.

UN Human Rights Council. *Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar*. 17.09.2018. https://undocs.org/en/A/HRC/39/CRP.2.

Standards and Regulations

Council of Europe (CoE). "Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law." *Council of Europe Treaty Series* (CETS) 225 ("CoE Framework Convention on AI"), 05.09.2024. https://rm.coe.int/1680afae3c.

European Union (EU). "Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC." *Official Journal* L 277/1 ("EU Digital Services Act"), 27.10.2022. https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

European Union (EU). "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828." *Official Journal* L 2024/1689 ("EU Artificial Intelligence Act"), 12.07.2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Organisation for Economic Co-Operation and Development (OECD). *Recommendation of the Council on Artificial Intelligence*. ("OECD Recommendation on AI"), 22.05.2019, amended 03.05.2024. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

United Nations Educational, Scientific and Cultural Organization (UNESCO). *Guidelines for the Governance of Digital Platforms, Safeguarding Freedom of Expression and Access to Information Through a Multistakeholder Approach*. ("UNESCO Guidelines"), 2023. https://unesdoc.unesco.org/ark:/48223/pf0000387339.

United Nations Educational, Scientific and Cultural Organization (UNESCO). *Recommendation on the Ethics of Artificial Intelligence*. ("UNESCO Recommendation on AI"), 23.11.2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

United Nations General Assembly. *Draft Resolution Submitted by the President of the General Assembly, The Pact for the Future*. Annex I ("UN Global Digital Compact"), 20.09.2024. http://undocs.org/en/A/79/L.2.

Jurisprudence

Faurisson v. France, Communication No. 550/1993, UN Human Rights Committee (1996). http://undocs.org/en/CCPR/C/58/D/550/1993.

Ross v. Canada, Communication No. 736/1997, UN Human Rights Committee (2000). http://undocs.org/en/CCPR/C/70/D/736/1997.

Sanchez v. France, Application number 45581/15, European Court of Human Rights (2021). https://hudoc.echr.coe.int/eng?i=001-211777.