Cambridge
Forum

Platforms on the hook? EU and human rights requirements for human involvement in content moderation

Emmanuel Vargas Penagos (iD)

School of Behavioural, Social and Legal Sciences, Örebro University, Örebro, Sweden
Email: Emmanuel.vargas-penagos@oru.se

Abstract

This article explores the human rights standards relevant to ensuring human involvement requirements in EU legislation related to automated content moderation. The opinions given by different experts and human rights bodies emphasise the human rights relevance of the way in which platforms distribute automated and human moderators in their services. EU secondary legislation establishes basic requirements for these structures that are called to be read under a human rights perspective. This article examines the justifications given for incorporating human involvement in content moderation, the different types of human involvement in content moderation, and the specific requirements for such involvement under EU secondary law. Additionally, it analyses the human rights principles concerning procedural safeguards for freedom of expression within this legal framework.

Keywords: Human-in-the-loop; content moderation; Digital Services Act; Social media

1. Introduction

In May 2023, during a major political crisis in Pakistan, a politician addressed parliament calling for the 'sacrifice' and 'hanging' of public officials, including himself, for their role in the situation. A news outlet shared a video of his speech on Facebook, reaching around 20,000 shares and 40,000 user reactions. Shortly after, Meta, Facebook's owner, activated a social media system involving human and automated tools to assess whether the video incited violence and should be removed (Reporting on Pakistani Parliament Speech, 2024).

The video had been flagged 45 times by Meta's automated systems as potentially violating the company's Violence and Incitement Community Standards. Afterwards, two human reviewers reached opposite decisions on whether the content actually violated those rules, one saying that there was no violation, the other one saying there was. An additional level of review by 'policy and subject matter experts' was given and concluded that the video was not in violation of the rules. Meta then referred the case to its Oversight Board, which upheld the decision after evaluating the rules, past decisions, and consulting experts (Reporting on Pakistani Parliament Speech, 2024).

While Meta's January 2025 announcement to change their moderation policies – starting in the US – (Meta, 2025) will surely change the way in which resources are deployed, this case exemplifies how online content moderation can sometimes involve complex structures in which human

and automated solutions are combined in different degrees. The way in which these structures are designed and deployed is critical from a human rights perspective. The European Court of Human Rights (ECtHR) has noted that the internet's capacity 'to store and communicate vast amounts of information' enhances the right to access and share information, but also increases risks to human rights (Delfi as v. Estonia, 2015, para. 133). Within that context, the ECtHR has also considered that social networks 'necessarily have certain obligations' in relation to content posted by their users (Sanchez v. France 2023, para. 185). Moreover, bodies like the Council of Europe and the United Nations Special Rapporteur on freedom of opinion and expression (the UN Foe Rapporteur) have stressed the need for social media companies to ensure the possibility of human review of online moderation decisions, instead of merely automated processes, as a human rights concern (Recommendation of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, 2018; UNGA, 2022).

The EU legislature has also recognised the significance of involving humans in online moderation systems. As is explained in this article, recent legislative developments on content moderation at EU level have referred to this issue in different ways. Particularly in the context of the Digital Services Act (DSA),[1] human involvement is implemented as part of a set of procedural safeguards (Ortolani, 2023). In the human rights context, procedural safeguards are fundamental for protecting substantive rights (UNCHR, 2007, para. 58) and the ECtHR has considered that the procedural limb of Article 10 of the European Convention on Human Rights (Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR), 1950) serves the broader purpose of safeguarding freedom of expression as a substantive right (Bulgakov v. Russia, 2020, para. 45).

It should be noted that, as this article shows, policy decisions on human involvement in content moderation are part of a broader debate about the role of automated or semi-automated decision-making in everyday life. Moreover, as explained in section 2 of this article, the justifications given for human involvement are related to practical issues, and sometimes to ethical ones.

At the same time, the European human rights framework, namely the EU Charter (Charter of Fundamental Rights of the European Union, 2012) and the ECHR as interpreted by the Court of Justice of the European Union (CJEU) and the ECtHR, are a baseline for interpreting the obligations on content moderation stemming from EU law(Frosio & Geiger, 2023). This is not only due to express reference in such legislation (DSA, Recital 47), but also because of the prevalence of the ECHR in the EU treaties (Consolidated Version of the Treaty on European Union, 2016, art. 6(3); Consolidated Version of the Treaty on the Functioning of the European Union, 2016, Protocol 8) and EU Charter (art. 53), as well as CJEU case law interpreting legislation related to freedom of expression (Case C–401/19, 2021, para. 44). Moreover, the United Nations Guiding Principles on Business and Human Rights are likely to play an additional guiding role, as suggested by Recital 47 of the DSA and by the work of the UN Foe Rapporteur (UNCHR, 2018).

The purpose of this article is to discuss the relevant EU secondary legislation related to human involvement in automated content moderation from a human rights perspective by addressing the following question: *what are the human rights principles applicable to the implementation of human involvement in online content moderation systems in the context of EU law?*

To better answer this question, the article will also consider these sub-questions: *i) What are the justifications given for human involvement in online content moderation?; ii) What are the different types of human involvement applicable for online content moderation? ii) What are the requirements for human involvement in EU law dealing with online content moderation? and iii) How can these requirements be read from a human rights perspective to enhance procedural safeguards for freedom of expression applicable in the context of such legislation?*

---

[1]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 (DSA) (Text with EEA Relevance), 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

This article is situated within two intersecting areas of legal research: first, the broader discussions on human involvement in automated decision-making processes across various domains, which examine how humans can complement or oversee automated systems; and second, the specific discussions on human involvement in automated content moderation within the field of content moderation itself. Within the first area, there is a vast body of legal research focusing on the right not to be subject to decisions based solely on automated processing, particularly profiling, as provided under Article 22 of the General Data Protection Regulation[23] and, more recently, there's legal research focused on explaining the requirements within the EU AI Act (Constantino, 2022; Enqvist, 2023). In addition to this, there is relevant research explaining and reflecting on different justifications given for human involvement in automation (Crootof, Kaminski & Price Ii, 2022; Goldenfein, 2024; Jones, 2015; Solove & Matsumi, 2024). The second area relates mostly to, beyond assumptions of human moderation as inherently superior or inferior, discussing the view that human involvement can operate as a safeguard against potential shortcomings of automated content moderation, as well as reflecting on the challenges that human moderators face while doing their job (Enarsson, Enqvist & Naarttijärvi, 2022; Gorwa, Binns & Katzenbach, 2020; Griffin, 2022; Roberts, 2019; Young, 2022). The legal literature on these discussions at EU level tends to focus on those concerns and examines whether the legislation mandates or allows automated moderation and the fact that human review is considered a safeguard (Coche, 2023; Romero Moreno, 2020; Senftleben, 2023). There is relevant literature focusing on the use of human rights standards to interpret the DSA (Enarsson, 2024; Frosio & Geiger, 2023), and the research by Douek (2022) centred on discussing the importance of platforms investing their time in creating adequate structures for moderation is key for the discussion on the value of human involvement in moderation, as well as the manner to do it.

The contribution of this article lies in the legal analysis of requirements under EU secondary law for human involvement in automation in general, as well as in content moderation in particular, from a human rights perspective. It aims to enrich legal literature by interpreting these requirements through the lens of human rights considerations.

This article is divided into four parts. First, it delves into the discussion of the reasons for human involvement in online content moderation. Second, it refers to the different ways in which humans are involved in online content moderation by looking into previous literature examining the issue. Third, it systematises the requirements for human involvement in EU law by looking both into the applicable legislation and the relevant case law at the CJEU on the subject matter. Fourth, it makes a link between those requirements and human rights law, with an emphasis in procedural safeguards.

Human involvement can happen at different stages and ways and the discussion tends to refer to non-interchangeable varied terms, such as 'human-in-the-loop,' 'human oversight,' among others. For the purposes of this article, the term to be applied most of the time is 'human involvement' for being the broadest one. Nevertheless, the article will also make use of the specific terms applied as they are referenced by the sources used.

2. Why is a human involved?

An adequate understanding of the justifications that are provided for the involvement of humans in content moderation requires first examining why content moderation processes rely on automation, which is what this section begins with. It then explains the ethical and human rights reasons that have

---

[2]See for instance: (Binns & Veale, 2021; Brkan, 2019; Enarsson et al., 2022; Malgieri & Comandé, 2017; Mendoza & Bygrave, 2017; Roig, 2017).

[3]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj

been given by different stakeholders for human intervention, followed by how humans are actually involved in practice.

2.1. Why is moderation automated?

Discussions about the way of implementing human or automated moderators resurface worries as old as the times of John Milton, father of modern freedom of expression philosophy. During the Licensing Act's censorship regime in 17th century England, Milton warned that those deciding on 'the birth, or death' of books must be 'studious, learned, and judicious' but would face a 'tedious and unpleasing' task (Milton, 1918). In other words, Milton was concerned that proper censorship required highly qualified individuals burdened by the responsibility of reviewing every book intended for publication.

Centuries later, many factors call for automation of online content moderation, including the overwhelming amount of content that has to be reviewed in real time, making it difficult if not impossible for humans to meet the moderation needs by themselves (Link, Hellingrath & Ling, 2016). The massive scale of content that requires moderation, as well as the complexity of its assessment is likely to be heightened by the exponential surge of generative AI (Jaidka et al., 2024; Shoaib, Wang, Ahvanooey & Zhao, 2023). Moreover, humans in charge of moderation are often faced with the task of reviewing content of violent, sexual or degrading nature, making their job an 'arduous and trauma-inducing' one (Langvardt, 2017). Consequently, hiring and maintaining a sufficiently large team of human moderators would be highly costly in economic and human toil terms (Crootof et al., 2022). These complexities are likely to be heightened if the Metaverse becomes mainstream as moderation within it is not limited to audio, video or text but rather to 'conduct' within a tridimensional space (Bovenzi, 2024).

Referencing research exposing the draining work they have to undertake (Roberts, 2019), the UN FoE Rapporteur has noted that human moderators face several challenges. They can make mistakes in different tasks, such as 'enforcing internal policies, training artificial intelligence systems and actively screening and removing offensive material' (UNGA, 2021, para. 82). This, says the Rapporteur, carries at the same time an 'emotional toll' for moderators due to the type of content that they have to see (UNGA, 2021, para 82).

Furthermore, a common argument for automation relates to a belief that human decision-making is flawed, seeing humans as slow, inefficient, emotional, irrational and impulsive, with different biases and heuristics that makes them unreliable. In contrast, automated decisions are viewed as superior because of their efficient, fast, objective, and data-driven nature. However, as Solove and Matsumi (2024) point out: machine and human decisions differ, with machines relying on quantifiable judgments and humans using emotional and non-rational elements that can be valuable in certain situations.

2.2. The justifications given for human intervention

Reliance on solely automated means for content moderation decisions has been criticised from a human rights perspective. In that sense, the UN FoE Rapporteur has noted a 'challenge of assessing context and taking into account widespread variation of language cues, meaning and linguistic and cultural particularities' (UNGA, 2018, para. 15). Based on technical research finding tendencies from machine learning to inherit human-like prejudices(Barocas & Selbst, 2016; Caliskan, Bryson & Narayanan, 2017), the Rapporteur has claimed that automated moderation systems may cause risks of discrimination against vulnerable groups due to the tendency of automated tools to be based on datasets with discriminatory assumptions. According to the Rapporteur, that is a factor that, in scenarios in which over-moderation costs are low, makes these systems prone to remove non-problematic content or suspend accounts based on biased criteria (UNGA, 2018, para. 15). To

illustrate this point, the Rapporteur refers to news coverage on the use of word-embedding algorithms by Instagram showing that a scientist, not related to Instagram or its parent company Meta, had built a word-embeddings algorithm to understand the underlying sentiment on restaurant reviews, which tended to give bad scores to the word 'Mexican' regardless of not being used with a negative or discriminatory tone but rather because online content, used for training of these models, tends to associate that word with illegality (Thompson, 2017).

From a technical perspective, reliance on solely automated means for moderation is problematic due to the issue of 'distribution shift.' This occurs when automated tools face content that is significantly different from the data they were trained on, resulting in a drop in their performance (Lai et al., 2022). Although automated models improve through feedback and self-learning processes, they have not yet reached a 'point of no return' – a theoretical stage where their performance is so advanced that human involvement would actually diminish their effectiveness(De Lemos, 2020). Even the most advanced models face risks such as errors in understanding nuanced content, bias, and under or over enforcement, underscoring the relevance of human oversight (Vargas Penagos, 2024).

Crootof et al. (2022) point that the inclusion of humans in decision-making systems could be aimed at different interests: i) correcting issues with the system's performance, such as on issues of error and bias; ii) having someone to act when the system is on 'failure mode' or when it is necessary to stop the system under emergencies; iii) increasing the system's legitimacy by providing reasoning to the system's decisions; iv) protecting the dignity of humans involved in the decision; v) establishing accountability; vi) creating 'stand-in' roles for regulatory compliance or compliance with human values; vii) creating friction roles to slow the pace of the system; and vi) creating intermediary roles between users and systems. Griffin and Stallman (2024) argue that, in the context of the DSA, the most relevant goals would be those related to the improvement of the systems and those aiming at having decisions properly justified and comprehensible to human users. They also argue that aims like having dignitary roles or adding a 'warm body' to the process, namely including a human for the sake of protecting jobs (Crootof et al., 2022), would be questionable due to the poor labour conditions faced by moderators. Moreover, having a human just for accountability reasons would not be necessary as the DSA already establishes rules on intermediary liability and regulatory oversight (Griffin & Stallman, 2024).

Some of these considerations seem aligned with the position taken by different human rights bodies. The Council of Europe has said that a human-rights based approach to moderation must consider the labour rights and mental health of workers involved in manual content review.(CDMSI, 2021) On its side, the UN FoE Rapporteur has taken several stances for including humans in moderation processes, which include: mitigating problems related to identifying underlying issues of nuance and context related with disinformation, particularly in cases with a risk of real-world injury or violence (UNCHR, 2021, para. 71) or in critical contexts like armed conflicts (UNGA, 2022, para. 126); in the context of internal remedial processes, providing appropriate checks on systems and guaranteeing accountability (UNGA, 2018, para. 60); as well as providing remedy for adverse human rights impacts of the systems (UNGA, 2018, para 70). Moreover, the Rapporteur argues that the inclusion of humans in the moderation process addresses the need of ensuring accurate, context-sensitive content moderation by professionalising human evaluators, protecting their labour rights, involving cultural and linguistic experts in each market and diversifying leadership and policy teams aiming to apply local expertise to content issues (UNCHR, 2018, paras. 56–57).

In the broader context of AI systems in general, bodies like UNESCO and the Commissioner for Human Rights of the Council of Europe have referred to concepts of human 'oversight' or 'control,' which imply continuous monitoring and influence over a process, rather than 'review,' which refers to evaluating a decision or outcome after being made. From an ethical-oriented perspective, on its Recommendation on the Ethics of Artificial Intelligence, which aims to provide

guidance to its Member States,[4] UNESCO (2021) referred to a need for States to ensure that legal and ethical responsibility for any of the different stages of the life cycle of AI systems falls on physical persons or legal entities. Moreover, UNESCO (2021) takes the stance that 'an AI system can never replace ultimate human responsibility and accountability.' On its side, the Commissioner for Human Rights stresses that AI systems must 'always remain under human control,' even when they can make decisions independently without human intervention. In a similar way as UNESCO, the Commissioner links human control to accountability (Commissioner for Human Rights, 2019).

2.3. Human involvement in practice

Those policy approaches by UNESCO and the Commissioner for Human Rights are broad enough to be interpreted as seeing human involvement's relevance in deployment stages, or within the training of the models applied. In that sense, there is a long stream of technical research on finding the best ways to incorporate human domain knowledge into the models. This can include teams of humans merely labelling the data that is fed into the model's training, but can also include different ways of incorporating interaction between the model and human 'teachers,' which can sometimes be domain-experts (Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán & Fernández-Leal, 2023; Wu et al., 2022). Something noteworthy of these recent trends is the growing emphasis on improving interaction between humans and models, as well as on expert involvement to reduce the technical knowledge needed to introduce domain knowledge into the models (Mosqueira-Rey et al., 2023). In other words, by facilitating human interaction in the learning process, people who have relevant knowledge on the topic that the model is meant to learn, but who are not computer scientists or engineers, are able to make a meaningful and direct contribution to the learning process.

In practice, as summarized in Annex 11 to the EU Commission's impact assessment for the DSA (EU Commission, 2020), human involvement in content moderation has been considered crucial due to the limitations of automated systems in handling complex, nuanced decisions. The Annex notes that, while machine learning tools are widely used to flag potentially harmful or illegal content, they often struggle with contextual judgments, requiring human reviewers to make final decisions. This includes cases in which algorithms are used to prioritize flagged content, with human moderators deployed for cases where the automated system's confidence is low. The Annex reflects on this relevance by pointing at the context of the COVID-19 pandemic, where a reliance on automation due to reduced human moderation led to an increase in erroneous content removals and user appeals. According to this document, this highlights the importance of human intervention to ensure accountability and reduce mistakes. Ultimately, the Annex posits that automated systems excel in triage and scalability, helping prioritize content for review, but human oversight is necessary to interpret context, handle sensitive cases, and refine system performance, which would imply that human and automated processes can be seen form a complementary approach that balances efficiency with accuracy and adaptability.

3. How are humans involved?

3.1. Human involvement in automated systems

Automated systems with human involvement are known as 'semi-automated,' or 'hybrid' (Enarsson et al., 2022). This may include systems where humans retain full decision-making autonomy with algorithmic assistance and recommendation systems, as well as those where humans act mainly as rubber-stampers with nominal control, also referred to as 'quasi-automation' (Enarsson et al., 2022). The EU High-Level Expert Group on Artificial Intelligence refers to the importance of human oversight as crucial for preventing AI systems from undermining human autonomy or causing

---

[4]To the date, UNESCO has 194 Member States (see Member States | UNESCO, n.d.).

adverse effects (Independent High-Level Expert Group on Artificial Intelligence, 2019). The Expert Group refers to three methodologies for that purpose: human-in-the-loop (HIL), human-on-the-loop (HOL), and human-in-command (HIC). HIL allows human intervention in every decision cycle of the system, for instance automated weapons that search or assess threats with humans deciding which targets are selected and engaged (Nahavandi, 2017); HOL involves intervention during design and monitoring the operation of the system, for example automated weapons acting independently to select and engage targets, but with supervision of humans who are in capacity of stopping operations when needed (Nahavandi, 2017); and HIC,[5] which has a more managerial approach, implies human capability to oversee the overall activity of the system, as well as its economic, social, legal and ethical impacts, and ability to decide the moments and the way in which the system is used, and even to override it. According to the expert group, the application of either category would vary from system to system and depending on factors like its level of risk and area of application (Independent High-Level Expert Group on Artificial Intelligence, 2019). In addition to this, other factors like regulatory requirements from one jurisdiction to another or policy decisions by deployers, among many others, are very likely to influence the way in which each category is applied (Veale, Matus & Gorwa, 2023).

3.2. *Human involvement in automated content moderation*

Online content moderation is one among many industrial processes that has increasingly resorted to automation in current times (Endsley, 2023). In addition to this, content moderation is an industrial process with distinct democratic challenges, as it has the potential of influencing public debate and democratic processes, as well as impacting the rights of individuals and communities (Sander, 2021). Moreover, content moderation is a process that has been increasingly regulated in recent times at EU level, with the DSA as a significant milestone. Moderation structures are varied and their shapes may change at any moment under different considerations, as exemplified by Meta's January 2025 announcement of changing their moderation policies and practices in the US, which includes reducing its reliance on automated tools and implementing more community and user based moderation (Meta, 2025). Furthermore, as Singhal et al. (2023) explain, 'there is no unified method for content moderation among the different social media platforms.' From an abstract point of view, moderation structures across companies can involve moderation solely performed by either humans or by automated tools, or moderation done by the former with the aid of the latter (Douek, 2022). In addition to this are layers of subsequent review in appeal procedures (Douek, 2022). Enarsson et al. (2022) condense this into three categories: human to human moderation, when one person reports content to be reviewed by a human moderator; fully automated moderation and semi-automated moderation, /involving flagging by automated tools with a contextual-based review by a human. In addition to this, there can be instances in which a human 'reports' or 'flags' and it is afterwards evaluated by an automated tool (Kou & Gui, 2021). A non-exhaustive illustration is shown in Fig. 1.

In practice, semi-automated moderation would generally consist of the use of automated tools with a set of pre-established rules to detect problematic content in order to take subsequent actions, such as deletion or further review by a human (Lai et al., 2022). On the opposite side, non-flagged content would not be subject to further actions (Lai et al., 2022). This can include several variations. For instance, Lai et al. (2022) have proposed a structure of 'conditional delegation,' consisting of a preliminary collaboration between the automated models and humans to determine the trustworthy cases of the model before deployment (e.g., identifying a specific word). After deployment, the model only affects decisions on those trustworthy cases while the rest would require further actions like human review or using a different model.

Jhaver, Birman, Gilbert and Bruckman (2019) examined Reddit's structure, which is based on the deployment of the 'Automoderator,' a relatively simple rules-based model used to flag content for

---

[5]To see examples of HOC applied as systems in which the user can 'independently train an ML model and in an iterative fashion, interact with it and interpret and understand its decisions' see Holmberg (2021).
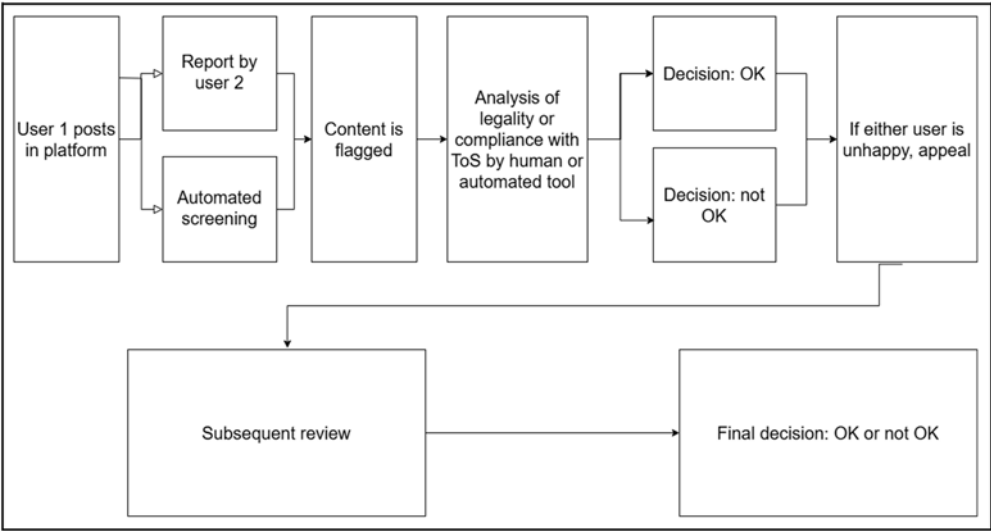
Figure 1. Non-exhaustive illustration of the moderation cycle.[6]

voluntary moderators on the platform. They found that, although the use of the tool may improve the efficiency in moderation processes, it creates direct, secondary effects on different stakeholders, such as the human moderators themselves. According to these researchers, the automated tools create additional work to the initial moderation because moderators end up having to correct mistakes of the model, as well as to respond to complaints due to an increase of mistakes. They also note that the use of the automated tool increases the need for higher technical expertise of the moderators. This example suggests that, like in other automated industrial processes, semi-automated moderation can reduce human workload while making oversight more crucial and complex (Bainbridge, 1983).

The mandatory reports that platforms have to publish as part of their obligations set in the DSA (art. 15(1)(e)) give some insights into how they involve humans in moderation. These reports, published for the first time in late 2023 and early 2024, expose some of the complex structures behind platform moderation. For example, Meta's transparency reports concerning Instagram and Facebook show that most of their moderation is done automatically, but they also include different types of semi-automation. This includes the automated detection of content, as well as the prioritisation of the most critical content on the basis of severity, virality, likelihood of a violation, of offline harm or of likely spread. Meta also explains that the outcome of human review turns into feedback for improving the automated model, meaning that if a content is found as violating a specific policy by a human reviewer, this person applies a label to the content, which will afterwards form part of the data for training and refining (Meta, 2023a, 2023b).

Meta's reports highlight that their human moderators 'receive in-depth training and often specialise in certain policy areas and regions' (Meta, 2023a, 2023b), but very little detail is given on how this is structured. Meta states that users can appeal decisions, but it is unclear whether appeals are handled automatically or by humans. They also mention the option to make a final appeal to their Oversight Board, a semi-external expert body that selects cases for review in a similar way as a Common law Supreme Court (Muniz Da Conceição, 2024), meaning not all appeals receive a final decision from this board (Douek, 2020).

A more detailed review of the reports by the rest of platforms classified as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to have a more comprehensive view and classification of the different structures applied would require separate, detailed and lengthy research. Nevertheless, being Meta a company with over three billion users in the world and 259 million users within the EU, which makes it one of the largest players globally and among those designated as VLOPs in the Union (EU Commission, n.d.). it can be seen as an influential powerhouse amongst social media platforms. This is emphasised by the fact that, at least until January 2025, Meta has been considered as one of the companies investing more efforts to implement human rights considerations within their moderation processes (Nourooz Pour, 2024). However, other types of structure may deserve attention, such as Reddit's mentioned above, based mostly on voluntary moderators assisted by automated tools.

4. The requirements for human involvement under EU secondary law

The EU legal regime refers to human involvement in automated decision making in different legislative acts. For content moderation, Barral Martínez (2023) argues that the current framework has different standards 'depending on the type of content at stake and on the subject requesting the removal.' This can also be seen as imposing an obligation to involve humans in a more complex structure than the mere adoption of rubber-stampers at the end of the process. As seen in the previous chapter, there is no unique model of human involvement. Moreover, the desired impact of human intervention in these systems is likely to change depending on 'what' is intervened, 'when,' and 'by whom' (Enqvist, 2023). The AI Act[7] provides standards for human involvement in high-risk systems, but these do not cover content moderation. During the legislative process, the Parliament sought to expand the high-risk list to include social media recommender systems (P9_TA(2023)0236, 2023), which are different from content moderation yet highly interconnected to it, but this was withdrawn from the final agreement (Bertuzzi, 2023). The final stance on this issue is reflected in Recital 118 of the AI Act, establishing that AI models deployed by VLOPs or VLOSEs are subject 'to the risk-management framework provided for' in the DSA, unless 'significant systemic risks' not covered by it are identified in such models. According to Hacker (2024), this interaction relates to the deployment of General Purpose AI (GPAI) by Social Media Companies, which can include AI models used for moderation. Neither Article 53, related to GPAI models in general, nor Article 55 of the AI Act, related to additional obligations for GPAI models with systemic risks, provide rules related to human involvement. Human involvement is only mandated to high-risk AI systems in Article 14 of the Act. For that reason, this section does not make emphasis on that Regulation. In any case, it should be noted that Article 7 AI Act allows the possibility for the EU Commission to expand the list of AI systems deemed as high-risk. This means that AI used for moderation could fall under the obligations on human involvement of the AI Act in the eventuality in which the Commission included them in the list of high-risk AI systems.[8]

This section outlines the requirements for human involvement in online content moderation under EU law. It begins by explaining the general rules for automated decision-making under the GDPR, followed by rules for automated content moderation, and concludes by looking at *lex specialis* for moderating certain types of content.

---

[7]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 (AI Act) (Text with EEA Relevance) (2024b). http://data.europa.eu/eli/reg/2024/1689/oj/eng

[8]Bayer (2024) has argued for classifying AI systems used for content moderation as high-risk due to their 'profound formative effect on the public discourse as they affect the fundamental rights of freedom of expression, freedom of information, the right to vote, and thereby democratic participation.'

4.1. The general rules for human involvement (Article 22 GDPR)

The first salient provision in relation to human involvement is Article 22(1) of the GDPR, which provides a right of data subjects 'not to be subject to a decision based solely on automated processing (..) which produces legal effects concerning him or her or similarly significantly affects him or her.' This provision is relevant in the context of content moderation because the GDPR is meant to cover a wide range of 'processing,' which the CJEU has interpreted to apply broadly (Case C-740/22, 2024), and which is understood under that Regulation as 'any operation or set of operations which is performed on personal data or on sets of personal data' (GDPR, art. 4(2)), which is at the same time understood as 'any information relating to an identified or identifiable natural person' (GDPR, art. 4(3)). The GDPR and its interpretation by the CJEU has given a broad scope to the meaning of 'identifiable,' which includes an assessment of 'all the means reasonably likely to be used, such as singling out, either by the controller or by another person' to directly or indirectly identify an individual, without requiring that 'all the information enabling the identification of the data subject must be in the hands of one person' (Case C–604/22, 2024). This can include, *inter alia*, anonymous information in power of a controller in combination with 'the IP address of a user's device or with other identifiers' (*Case C–604/22*, 2024). Such a definition is broad enough to encompass users' content in social media, as it constitutes information or opinions by them, which can be combined with their IP address or other several identifiers. The DSA establishes several rules for platforms' content moderation, which encompasses several 'activities' performed by them for 'detecting, identifying and addressing illegal content or information incompatible with their terms and conditions' (DSA, art. 3(t)). Given the broad definition of data 'processing,' those activities fall under the GDPR, something that is also underpinned expressly by the DSA, which provides that it applies without prejudice to the former (DSA, art. 2(4)(g)).

The SCHUFA decision by the CJEU (Case C–634/21 OQ v Land Hessen, 2023) provides detailed guidance on interpreting Article 22(1). This interpretation, complemented by guidance issued by the Article 29 Working Party,[9] suggests that Article 22(1) would be applicable in the context of content moderation.

According to the CJEU, Article 22(1) provides three cumulative conditions for its application: i) a decision must be made; ii) it is based solely on automated processing, which includes profiling; iii) it produces legal effects or similarly significantly affects the individual (Case C–634/21 OQ v Land Hessen, 2023, para. 43).

The Court interprets 'decision' in broad terms to include 'a number of acts which may affect the data subject in many ways' (Case C–634/21 OQ v Land Hessen, 2023, para. 46), even without legal effects (para. 43). Before this judgment, it had been argued that 'decision' entails the 'outcome' of processing (Tosoni, 2021) which has actual effects on the individual concerned (Binns & Veale, 2021). On its side, the CJEU considered that automated parts of the process that are 'preparatory' or performed by third parties are included in the concept of 'decision' (paras. 61–62). While the Court focused on defining 'profiling' – which is beyond the scope of this article – the Article 29 Working Party has defined 'automated decision-making' as 'the ability to make decisions by technological means' (Article 29 Data Protection Working Party, 2016). Instead of focusing on the word 'solely,' the Court stated that the condition of producing legal effects is met if the automated tool's output 'strongly' affects the final outcome (para. 48) or 'plays a determining role in it' (para. 50). The Article 29 Working Party has emphasised that 'based solely' means 'there is no human involvement in the decision process' (Article 29 Data Protection Working Party, 2016, p. 20). These positions seem aimed at preventing humans acting as mere rubber-stampers without real authority to overturn decisions (Binns & Veale, 2021).

---

[9] The Article 29 Working Party was a body created by the predecessor of the GDPR, the Data Protection Directive, with the role of providing guidance on the interpretation of that Directive. With the GDPR, this body was replaced by the European Data Protection Board (EDPB), which has similar functions. The EDPB has endorsed several guidelines by the Working Party, including the ones cited in this article (see Endorsed WP29 Guidelines | European Data Protection Board, n.d.).

Furthermore, the Working Party considered 'legal effects' broad enough to encompass both statutory and contractual effects, and interpreted 'similarly significantly' to mean effects that are 'more than trivial' and 'sufficiently great or important to be worthy of attention' (Article 29 Data Protection Working Party, 2016, p. 29). This implies the decision must have the potential to significantly influence an individual's 'circumstances, behaviour or choices,' whether the effects are negative or positive (p. 21).

The Court and the Article 29 Working Party's broad interpretation of Article 22(1) makes it applicable to several decision-making instances, such as banks' decisions to refuse or grant loans on the basis of automated credit scoring (Case C–634/21 OQ v Land Hessen, 2023), automated evaluations of workers (Aza, 2024), and content moderation. Content moderation decisions fall under this article because they are different acts emerging from the outcome of the 'processing' of users' personal data (i.e. Content published by them), with impact on the 'availability, visibility, and accessibility' of such content or the potential suspension or termination of accounts (DSA, art. 3(t)). Those effects would be of a legal nature, given that the relationship between users and platforms is a contractual one governed by the terms and conditions (DSA, art. 3(u)). In other words, moderation decisions would determine if a user is in breach of the T&C as a contractual relationship, or in breach of legal prohibitions, and would activate measures that ought to be determined in the T&C, like suspensions, terminations of accounts, or any limitation to availability, visibility and accessibility of content. Moreover, it would also mean the application of rights of users filing notices under those T&C. Outsourcing moderation practices, like those used by Facebook (Laux, Wachter & Mittelstadt, 2021), would also fall under this article's scope.

The CJEU considered that the right 'not to be the subject of a decision solely based on automated processing' laid down in Article 22(1) GDPR is a prohibition in principle and the persons concerned do not need to invoke its infringement individually (*Case C–634/21*, 2023, para. 52). However, Article 22(2)(b) provides that the prohibition laid down in Article 22(1) shall not apply if the decision is authorised under EU or Member State law, as long as it 'lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests.' Article 22(4) establishes that these types of processing cannot be based on the special categories of personal data enlisted in Article 9(1),[10] unless the processing is done with consent of the data subject or is 'necessary for reasons of substantial public interest, on the basis of Union or Member State law' and done in a proportionate manner, with suitable measures to safeguard users' rights.

In this case, and as is explained below, the DSA provides a legal basis under EU law for automated moderation while, at the same time, establishes a set of requirements and safeguards for it. These safeguards are complemented by additional *lex specialis* requirements for the moderation of specific types of content or activity online. As will be seen, while compliance with the EU framework is not necessarily fulfilled by implementing rubber stampers for moderation, the legislation does not establish strict or rigid requirements for human involvement in that context. This is regulated by a set of minimum standards with dedicated rules for those scenarios contemplated by *lex specialis*.

4.2. *The general rule for human involvement in moderation decisions*

As such, the applicability of Article 22 GDPR to content moderation does not entail that all moderation decisions should be subject to human involvement, but instead that it should be subject to 'suitable measures to safeguard' the users' rights. Interpreting this as a right to have every moderation decision reviewed by a human would have implications in terms of cost and workforce that could become detrimental to the actual purposes of automating content moderation (Barral Martínez,

---

[10]According to Article 9(1) GDPR, this encompasses 'personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.'

2023). Given the massive scale of content that is published in social media, automated tools should not be a mere triage for human decision-makers, particularly given the possibilities for incorporating automated tools in cases that are uncontroversial or in which they are able to perform at a desirable level.

In that vein, it is key to note that the CJEU has been, to some extent, open to interpret the GDPR in consideration of its impact on key services of the internet. The Court has considered that data controllers' 'responsibility and obligations' are dependent on 'the specific features of the processing (..) in connection with the activity' (Case C-136/17, 2019, paras. 48–50). The DSA appears to be aligned with that vision, given the fact that it tailors obligations based on the type of service. For 'mere conduit' and 'caching' services, which only transmit or temporarily store information, the rules for liability for third-party content is minimal and technical.[11] 'Hosting' services, including platforms, have more responsibilities under the DSA. In any case, the DSA seems to provide some leeway on when and how to implement human involvement in the moderation cycle within platforms' services, which has at least four starting points: the enforcement of terms of service (art. 14 read jointly with art. 2(t)), responding to notices by users or by trusted flaggers on illegal content (arts. 16 and 22), the implementation of 'voluntary own-initiative investigations' or 'other measures' to address illegal content or necessary to comply with EU or national law (art. 7), as well as those adopted to enforce terms and conditions (art. 14(4)). The DSA does not prohibit automated moderation within any of these pathways, but rather establishes safeguards for it. In other words, the DSA is not written in a sense that requires all moderation decisions to be subject to human review. The DSA includes a series of obligations for transparency of platforms about the use of automated decision making in their terms and conditions (art. 14(1)), in their annual reports (art. 15(1)), in notifications for decisions on the basis of notice and action mechanisms (art. 16(6)), and in statements of reasons for restrictions imposed on the users on the basis of findings that their content was illegal or contrary to the terms and conditions (art. 17(3)(c)).

Moreover, human involvement is not mandated in any of the provisions related to the enforcement of terms and conditions or of the removal or disabling of illegal content, but it is instead included at an appeals stage. The DSA obliges platforms to incorporate internal complaint-handling mechanisms that should be available for six months and free of charge for users wanting to contest moderation decisions. Pursuant to Article 20(6), those mechanisms should work 'under the supervision of appropriately qualified staff, and not solely on the basis of automated means.' In that sense, the inclusion of humans in moderation under the DSA is part of the safeguards for appeal processes, something that can be seen as aligned with the positions by the UN FoE Rapporteur (UNGA, 2018) and UNESCO (2021) that were previously mentioned.

Article 20(6) is drafted in a sufficiently wide manner, allowing a wide range of possible structures for human involvement at the appeals stage, 'spanning from "appropriately qualified" human moderators to a highly judicialised body such as the Oversight Board' (Ortolani, 2023). However, Griffin and Stallman (2024) argue that this article should be seen as requiring systematic oversight by knowledgeable staff, enhancing communication and the training of automated systems instead of just having an army of human reviewers.

4.3. The lex specialis to human involvement in moderation

In addition to the general rules set in Article 22 of the GDPR and in the DSA, the EU legal framework has, at this moment, two legislative acts regulating specific types of human involvement in moderation. One is the Copyright Directive[12] and the other one is the Regulation on

---

[11]For a more detailed analysis of the regime of these services (see Schwemer, Mahler & Styri, 2021).

[12]Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC (Copyright Directive) (Text with EEA Relevance.), OJ L (2019). http://data.europa.eu/eli/dir/2019/790/oj/eng

addressing the dissemination of terrorist content online (TERREG).[13] Moreover, the European Media Freedom Act[14] establishes additional requirements for moderation of journalistic accounts that, although do not expressly refer to human involvement, may have an impact on it. Other norms within the EU legal framework impacting content moderation, such as the Child Sexual Abuse and Exploitation Directive[15], the Audiovisual Media Services Directive[16], and the Counter-racism Framework Decision (2008) do not create obligations related to human involvement.

1. The copyright directive

Article 17 of the Copyright Directive mandates that platforms obtain authorisation, such as a licensing agreement, from rights holders to make copyright-protected works available to the public. These authorizations would cover content published by users on the platform, making it not liable for that user-generated content. If platforms lack authorization, they are liable for unauthorised public communication of copyrighted content unless they can show efforts to secure authorization, prevent unauthorised content, and promptly remove or disable access to infringing content when notified by rights holders.

Article 17(9) of the Copyright Directive establishes several obligations to safeguard users' rights, particularly those covered by exceptions and limitations to copyright like quotation, criticism, review, parody and pastiche, and to minimise the risks of broad filtering and over-blocking that Article 17 may entail (Quintais et al., 2020). Those safeguards include, among others, that 'decisions to disable access to or remove uploaded content shall be subject to human review' (art. 17(9)). The CJEU says that this provision establishes a right to a complaint mechanism subject to human review, which forms part of 'procedural safeguards (...) which protect the right to freedom of expression and information of users of online content-sharing services in cases where (...) the providers of those services nonetheless erroneously or unjustifiably block lawful content' (Case C–401/19, 2021).

2. TERREG

Article 5(1) TERREG establishes obligations for platforms to include provisions in their terms and conditions to prevent the misuse of their services for the dissemination of terrorist content, defined as material that incites, solicits, or instructs on committing terrorist offences, glorifies such acts, or threatens to commit terrorism (art. 3(7)). Moreover, Article 5(2) provides that platforms have to adopt 'specific measures to protect its services against the dissemination to the public of terrorist content,' including among those implementing technical tools and staffing to quickly identify and remove terrorist content, providing user-friendly reporting systems for flagging such content, enhancing user moderation to increase awareness, or any other strategies the hosting service provider finds suitable to combat the presence of terrorist content on their platform. According to Article 5(3), if the measures adopted are 'technical,' they should be subject to 'appropriate and effective safeguards, in particular through human oversight and verification' to ensure accuracy and avoid removal of legitimate content. Recital 25 provides guidance on how to interpret these obligations by saying that automated

---

[13]Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (TERREG) (Text with EEA Relevance), 172 OJ L (2021). http://data.europa.eu/eli/reg/2021/784/oj/eng

[14]Regulation (EU) 2024/1083 of the European Parliament and of the Council of 11 April 2024 Establishing a Common Framework for Media Services in the Internal Market and Amending Directive 2010/13/EU (European Media Freedom Act) (2024a). http://data.europa.eu/eli/reg/2024/1083/oj/eng

[15]Directive (EU) 2011/93 of the European Parliament and of the Council of 13 December 2011 on Combating the Sexual Abuse and Sexual Exploitation of Children and Child Pornography, and Replacing Council Framework Decision 2004/68/JHA (2011). https://eur-lex.europa.eu/eli/dir/2011/93/oj

[16]Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 Amending Directive 2010/13/EU on the Coordination of Certain Provisions Laid down by Law, Regulation or Administrative Action in Member States Concerning the Provision of Audiovisual Media Services (Audiovisual Media Services Directive) in View of Changing Market Realities (2018). https://eur-lex.europa.eu/eli/dir/2011/93/oj

tools are not mandatory, but platforms are able to use them if they deem it 'appropriate and necessary to effectively address the misuse of their services for the dissemination of terrorist content.' These obligations complement the requirement in Article 3 that platforms, except in very exceptional circumstances, must remove terrorist content within one hour of receiving a notice from competent authorities.

TERREG has been subject to criticism as it is considered to have a high level of risk for the rule of law and freedom of expression, particularly because of its broad definition of terrorist content, as well as the one-hour obligation, which, paired with the possibility of using automated tools without judicial oversight, would incentivize using them. As such, the DSA is seen as an instrument with complementary safeguards for fundamental rights at stake (Coche, 2023).

3. The European Media Freedom Act

The European Media Freedom Act was created to establish rules for media services, namely news outlets, while safeguarding their independence and pluralism (European Media Freedom Act, art 1(1)). Regarding content moderation, the EU legislature recognized that VLOPs 'act for many users as a gateway for providing access to media content and media services' and 'play a key role in the distribution of information and in the exercise of the right to receive and impart information online' (Recital 50). For this reason, to protect media freedom and pluralism, the European Media Freedom Act sets additional rules beyond the DSA for enforcing terms of service and applying suspensions or restrictions to visibility of content published by media services in platforms.

In practice, this Regulation establishes a right of media services to declare to the platform that they provide such service, triggering specific safeguards in the course of content moderation. First, in the course of internal proceedings by platforms to suspend accounts or restrict the visibility of content published by media services, they are obliged to provide a statement of reasons before the decision is taken (European Media Freedom Act, art. 18(4)(a)), while, in cases of non-media services' content, this statement comes after the decision(DSA, art. 17). Moreover, platforms also have to afford media services with the opportunity to respond within 24 hours after receipt of the statement of reasons, or less in cases of crisis (European Media Freedom Act, art. 18(4)(b)), understood as situations where 'extraordinary circumstances lead to a serious threat to public security or public health in the Union or in significant parts of it' (DSA, art. 36(2)). This regulation also requires that if a media service disputes a decision using the complaint-handling mechanism, the complaint must be handled 'with priority and without undue delay' (European Media Freedom Act, art. 18(5)).

Although these requirements do not directly mention human involvement, they may still warrant attention when assessing if such a safeguard is in place, and they may even point to a preference towards human involvement. As such, the European Media Freedom Act emphasises the need to protect journalistic activities in a timely manner, which may sometimes demand for more detailed, careful and specialised review.

5. The human involvement in automated moderation under EU secondary legislation from a human rights perspective

The DSA was adopted with the purpose of creating rules for a 'safe, predictable and trusted online environment' (art. 1(1)) and to protect fundamental rights online, particularly 'freedom of expression and of information, the freedom to conduct a business, the right to non-discrimination and the attainment of a high level of consumer protection' (Recital 3). In doing so, the DSA makes a strong emphasis on establishing a framework for the responsible and diligent behaviour of online intermediary services (Recital 3). In that sense, the DSA can be interpreted as a legal instrument setting forth rules for social media companies' role within the ECHR's aim of creating a 'favourable environment for participation in public debate by all the persons concerned, enabling them to express their opinions and ideas without fear' (Khadija Ismayilova v. Azerbaijan,

2019, para. 158). The state is the main actor responsible for creating this environment, but there is a growing recognition that social media platforms play a key role in this context due to their capacity to facilitate or obstruct access to online forums of public debate (McGonagle, 2019).

The EU legal framework introduces human involvement in automated content moderation as a procedural safeguard for achieving those aims of protecting fundamental rights and having an online environment that is safe, predictable and trustworthy. In this context, the way in which human involvement is implemented is not limited to operative considerations. While human involvement is explicitly mandatory in very specific and limited situations, the legal framework provides requirements influencing the way in which social media companies allocate their resources in moderation structures. This is likely to affect the design, deployment, appeals and feedback process of the automated tools used for moderation in ways that would potentially be benefited by the introduction of human involvement.

Those requirements, enlisted in Fig. 2 have emerged as legal concepts within different legal fields, including among those human rights law. These are concepts that have been fleshed out by different human rights bodies like the ECtHR and other authorities at the Council of Europe and the Universal System, which at the same time have served to develop other notions, such as good governance principles (Addink, 2019). For that reason, given the impact that content moderation has on different fundamental and human rights, as well as the focus given by the EU legislator to the balancing of rights within the DSA (Recital 47; art 1(1)), the interpretation of these requirements is benefited, and required to be seen from, a human rights perspective. This is because they put in words the actual way in which moderators as human and automated enforcers of law and terms and conditions have to balance the rights and interests at stake. Against that background, the following paragraphs will provide a reading to the DSA requirements on the basis of relevant sources for human rights law.

Furthermore, as will be seen across this subsection, the implementation of human involvement within content moderation structures is not merely a formalistic or symbolic obligation. Instead,

	Art 14(4) DSA: Enforcement of restrictions in general	Art 16(6) DSA: Notice and action mechanisms	Art 22(1) DSA: Notice and action mechanisms activated by trusted flaggers	Art 20(6) DSA Complaint-handling mechanisms	Art 18(5) European Media Freedom Act: complaint-handling mechanisms activated by media services	Art 5(3)TERREG: Counter-terrorism measures	Art 17(4) Copyright Directive: Copyright enforcement
Diligent	✓	✓	✓	✓	✓	✓	
High industry standards of professional diligence							✓
Objective	✓	✓					
Non-discriminatory				✓	✓	✓	
Timely		✓	✓	✓	✓		
Handled with priority and without undue delay			✓		✓		
Non-arbitrary		✓	✓	✓	✓		
Proportionate	✓						

Figure 2. The requirements for the moderation cycle.[17]

it is a procedural safeguard that, for its implementation, requires the assessment and balancing of different factors from a human rights perspective. In practice, this has an impact on the resources allocated for the number of human moderators, their qualifications and their roles within the structure.

5.1. Diligence

The DSA establishes 'diligence' as a requirement for platforms' application and enforcement of restrictions based on their terms and conditions or considered to be illegal (art. 14(4)), as well as for taking decisions on notice and action mechanisms (art. 16(6)) and in their complaint-handling system (art. 20(4)). Diligence is also a requirement for the implementation of measures in the context of TERREG (art. 5(3)(d)). Within the Copyright Directive, companies are required to act 'in accordance with high industry standards of professional diligence' (art. 17(4)(b)). This requirement sets the manner in which different activities should be performed, and is part of the chapter on 'due diligence obligations for a transparent and safe online environment' of the DSA. These obligations can be interpreted within the understanding of 'due diligence' as a 'standard of conduct required to discharge an obligation' or, in other words, to 'take reasonable precaution to prevent, or to respond to, certain types of harm' (Bonnitcha & McCorquodale, 2017; Rouas, 2022). Looking at this requirement from these lenses is desirable to avoid a view merely from the perspective of cost and efficiency in the implementation of the moderation systems, although, as explained by Senftleben (2023), there's a high likelihood that commercial considerations will lead to that path.

In the human rights context, due diligence is contextual and does not imply a predefined list of measures that the addressee needs to take to discharge their obligation (Malaihollo, 2021). Moreover, the DSA seems to be drafted with recognition that diligence will vary on the basis of different factors like 'the type, size and nature of the intermediary service concerned' as well as a specific end of addressing 'public policy concerns, such as safeguarding the legitimate interests of the recipients of the service, addressing illegal practices and protecting the fundamental rights enshrined in the Charter' (Recital 41). At first glance, this means that 'hosting' providers have different obligations and regimes than those providing 'caching' and 'mere conduit' services, as well as that VLOPs and VLOSEs are subject to additional due diligence requirements (arts. 4–6). However, this can also mean that the general obligations would also vary depending on the mentioned factors. This can be seen in the case law of the ECtHR considering that 'the size of the entity and whether or not it is engaged in a profit-making activity' is a factor for assessing the duties and responsibilities of internet intermediaries for third-party comments (Sanchez, 2023, para. 165). This is fundamental to avoid the imposition of a specific predefined moderation structure, which, at the same time, can alter the way in which platforms offer services. For example, the moderation structure of Reddit mentioned in section 3.2 may not be suitable for a platform like Facebook, but it has been fundamental for that platform's success and appeal to users (Roose, 2024).

In that vein, the diligence on how to distribute and organise automation and human involvement would also vary on the basis of those factors. For instance, platforms with the capacity of designing, training and testing their own automated tools for moderation would probably have a heightened diligence requirement in those stages, which can potentially be discharged by implementing state-of-the-art methodologies like the aforementioned ones on improving training by facilitating interaction and involvement of domain-experts (Wu et al., 2022). Likewise, platforms with access to a wide portfolio of automated tools may be in better capacity for implementing 'conditional delegation' to those tools deemed trustworthy enough to be implemented without human involvement, and which ones should be used as triage for further automation or for human review (Lai et al., 2022).

Those factors would also have an impact on the level of agency and the training that platforms should provide to their human moderators intervening at different stages (Wagner, 2019). In that same line, the way in which the 'appropriately qualified staff' for internal complaint-handling systems is configured would also be affected by those factors. The requirement for 'high industry standards of professional diligence' in copyright enforcement may be seen as a heightened duty in order to prevent special risks from the potential use of upload filters for copyright enforcement (Senftleben, 2023). This could imply, for instance, a need for specific training or knowledge of human moderators in copyright exceptions.

In addition to the above, as mentioned, VLOPs have additional due diligence obligations. Article 34 DSA mandates those platforms to analyse and assess systemic risks emerging from, among others, the design or functioning of their service and related systems, 'including algorithmic systems,' as well as their moderation systems. Subsequently, Article 35 DSA mandates the design and adoption of mitigation measures that may include adapting moderation processes and testing and adapting their algorithmic systems, among others.

5.2. *Objectivity and non-discrimination*

The DSA mandates that the application and enforcement of restrictions by platforms, as well as the decision-making in notice and action mechanisms have to be 'objective' (arts. 14(4) and 16(6)), while decision-making in complaint-handling mechanisms needs to be 'non-discriminatory' (art. 20(4)). Moreover, being 'non-discriminatory' is a requirement for counter-terrorism measures under TERREG (art. 5(3)(d)).

While objectivity and being non-discriminatory may be related, they are not the same. In adjudication, objectivity can be seen as legal interpreters deciding on the basis of established rules and standards, not just personal opinions (Fiss, 1982). There is a traditional tendency of seeing automated tools for decision-making as more objective than human beings because they are based on statistics (Araujo, Helberger, Kruikemeier & De Vreese, 2020). However, 'an objective interpretation is not necessarily a correct one' (Fiss, 1982). Despite their mathematical foundations, automated tools can have discriminatory outcomes due to several factors that can even be unintended or unobservable at a first stage (Ajunwa, 2020). Referencing research on gender biases within social media companies (Gerrard & Thornham, 2020), the UN FoE rapporteur has referred to this issue in the context of content moderation by saying that neither the community guidelines nor the algorithmic moderation in social media is objective because 'it reflects the biases and worldviews of the rule-setters, who tend to be typically from the specific sociocultural context of Silicon Valley: racially monochromatic and economically elite' (UNGA, 2021, para. 84).

In that sense, the use of the terms 'objective' and 'non-discriminatory' could be interpreted as providing two different layers in which the aim is to prevent discriminatory measures in content moderation and in which the human factor would play a key role. This is so particularly because the value of human intervention is considered as realised when the human 'has the ability to contextualise and assess complex situations' (Frosio & Geiger, 2023).

One first layer would relate to the general application and enforcement of restrictions and decisions within notice and action mechanisms. Here, the general rule would be to set simple rules and standards for both human and automated moderators when making decisions. However, this first layer would require additional safeguards against discriminatory outcomes when moderation relates to content that may be deemed as terrorist. This would be aligned with worries expressed by the UN High Commissioner for Human Rights, who has observed that counter-terrorism legislation imposing moderation obligations in different parts of the world creates risks of overcompliance and has resulted in undue discriminatory restrictions of content (UNCHR, 2022). In the words of the High Commissioner, 'surveillance and moderation of online content can disproportionately affect individuals from specific religious, ethnic or other groups, this raises concerns of discrimination' (para. 28).

The second layer implies the establishment of a complaint-handling mechanism with non-discrimination as a guiding principle. This can be interpreted as making this last stage aimed at correcting discriminatory moderation decisions, meaning those that cause a 'difference in the treatment of persons in analogous, or relevantly similar, situations' or 'without an objective and reasonable justification, fail to treat differently persons whose situations are significantly different' (Eweida and Others v. the United Kingdom, 2013, para. 87). Moreover, this can also be interpreted as mandating that the complaint-handling decisions should not be discriminatory themselves.

Therefore, a non-discriminatory approach to human involvement at these two stages would prioritise qualified and skilled intervention. This could include ensuring that human moderators have expertise in non-discrimination, the authority to reverse biased decisions, and the ability to identify patterns or issues that can help improve automated models.

5.3. Timeliness

Under the DSA, 'timeliness' is a specific requirement for notice and action (art. 16(6)) and complaint handling mechanisms (art. 20(4)). If seen as private adjudication, these mechanisms would be bound by the human rights principle of 'administering justice without delays which might jeopardise its effectiveness and credibility' (Scordino v. Italy, 2006, para. 224). However, timeliness in adjudication is relative, and the 'reasonableness of the length of proceedings is to be assessed in the light of the particular circumstances of the case' (Adiletta and Others v. Italy, 1991, para. 17) This could mean, for instance, that complex cases can be reasonably slower than simple ones, meaning that further internal reviews by humans or additional models can be expected in order to achieve an adequate decision. This is emphasised by the requirement to handle notices by trusted flaggers (DSA. art. 22(1)), as well as complaints by media services within the complaint-handling mechanism, with 'priority' (European Media Freedom Act, art. (18(5))), which can be seen as a duty to act more promptly, but also to allocate more qualified resources. This is critical from a human rights perspective. On the one hand, trusted flaggers under the DSA are meant to be independent experts that have gained relevant credibility, which would mean that their notices have a heightened degree of reliability. On the other hand, handling a complaint against a restriction on journalistic accounts relates to the possibility of reversing decisions to withhold or limit the visibility of press publications, something critical in ECtHR terms because 'news is a perishable commodity and to delay its publication, even for a short period, may well deprive it of all its value and interest' (Observer and Guardian v. the United Kingdom, 1991, para. 60).

5.4. Non-arbitrariness

The DSA mandates that notice and action and complaint-handling procedures are 'non-arbitrary' (arts. 16(6) and 20(6)). The ECtHR views arbitrariness as a violation of the rule of law, making its avoidance a key principle that restricts discretion in adjudication (Al-Dulimi and Montana Management Inc. v. Switzerland, 2016, para. 145). The ECtHR has considered that arbitrariness is avoided when adjudicators conduct 'an analysis of the facts and applicable law,' taking into account 'a combination of the provisions of ordinary law and of the special law,' and providing an explanation 'with convincing reasons' (Lupeni Greek Catholic Parish and Others v. Romania, 2016, para. 95). This concept is closely related to objectivity, but the key distinction from it is that arbitrariness concerns the limitations placed on the decision maker's authority. The ECtHR has emphasised that legal frameworks must not grant decision makers 'unfettered power,' and that there must be 'sufficient clarity' regarding the scope of discretion and its exercise(Gillan and Quinton v. the United Kingdom, 2010, para. 77).

This requirement suggests improving automated systems to assess exceptions, nuance, and context, rather than just enforcing rules blindly. Including human intervention at the review stage could be

seen as a guardrail to achieve that purpose, but this would demand clear and constrained guidelines that allow reasonableness and clarity in the decision.

5.5. Proportionality

Article 14(4) DSA mandates that the application and enforcement of restrictions to users' rights has to be 'proportionate.' This can be interpreted as establishing that restrictions, be them based on terms and conditions or grounded on legal provisions, must be appropriate to attaining the legitimate objectives pursued by the DSA or the legislation grounding them and 'do not exceed the limits of what is appropriate and necessary in order to achieve those objectives' (Joined Cases C–293/12 and C–594/12, Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others, 2014). In some situations, due to the relevance given by the ECtHR to fairness of proceedings and procedural safeguards afforded within them as a relevant factor for assessing proportionality (Cumhuriyet Vakfi and Others v. Turkey, 2013, para. 59), human involvement may be crucial depending on their qualifications for incorporating balance. This could, for instance, apply in situations where the applicable sanction is more severe, such as potential suspensions or terminations of accounts.

6. Summarising the framework

As explained in section 5, the framework for human involvement in online content moderation can be summarised in the following way: first, since automated content moderation entails decisions producing legal effects that are authorised under EU law, they require adequate safeguards. Second, as a general rule, content moderation would only demand human involvement at an appeals stage, but the way in which this is implemented is free for the platforms to decide as long as it implies supervision by appropriately qualified staff. Third, when the moderation decision relates to copyright enforcement, the complaint mechanism that is granted must include human review. Fourth, when moderation relates to the prevention of terrorism, it is meant to be subject to human oversight and verification in addition to further access to appeal stages that are provided as a general rule by the DSA. See Fig. 3 for an illustration. Finally, if moderation is done over content provided by media services, it has to incorporate specific safeguards, such as a compulsory statement of reasons before content removal and the opportunity for the media service provider to respond within 24 hours.

In that sense, while platforms are relatively free in the way in which they configure their moderation structures, these have some minimum requirements for human involvement. These are mostly related to 'when' is the intervention required. This can be interpreted as leaving a wide margin on the way in which the moderation involves humans prior to appeal stages, except when the content to be moderated is deemed as terrorist. As such, companies may apply full automation for uncontroversial subject matters, or matters where the accuracy of the systems is deemed safe to avoid over-censorship. However, when the automated tools flag content as terrorist, human oversight is required. At the appeals stage, the legal framework seems to be requiring companies to invest resources in securing the supervision by appropriately qualified staff, which at the same time must have specific allocations for applying direct human review in copyright disputes.

Furthermore, the human involvement within the EU secondary law affecting the moderation cycle has specific requirements that have direct human rights implications as explained in the previous section and which is illustrated to some extent in Fig. 4.

In that sense, the DSA requires online platforms to act **diligently**, namely responsibly and thoroughly, when enforcing their terms and conditions, processing user notices, and handling complaints. This means platforms must not solely focus on cost and efficiency but should allocate appropriate resources – both human and automated – to effectively moderate content. Diligence does not imply a one-size-fits-all approach but recognizes that different platforms may require different moderation
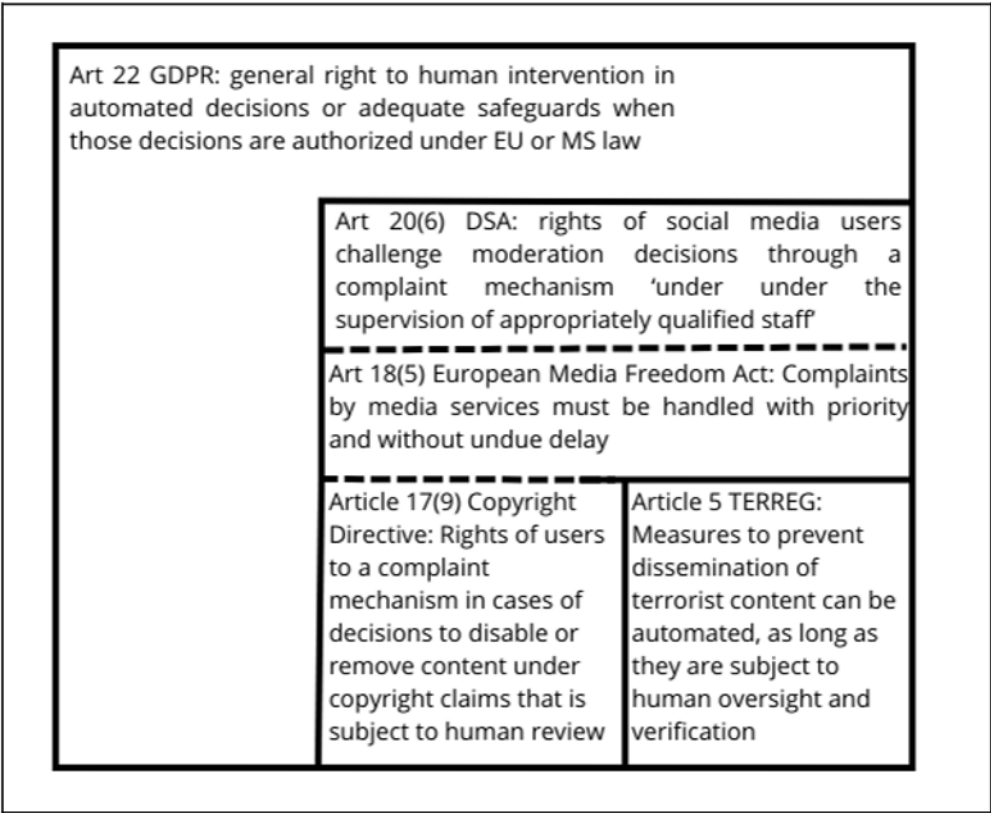
Figure 3. Illustration of the EU framework for human intervention in moderation.[18] Author's work.

structures suited to their unique contexts, thereby upholding fundamental rights without imposing unnecessary uniformity. **Objectivity** involves making moderation decisions based on established rules and standards rather than personal opinions, aiming for consistency in enforcement. However, an objective approach does not automatically prevent discrimination. **Non-discrimination** requires that moderation decisions do not unjustly favour or disadvantage any individual or group. In practice, this means that platforms, at the complaint-handling level, are called to be vigilant about biases that can exist at the previous stage. Human involvement is crucial here, as humans can contextualise and assess complex situations more effectively than automated tools, helping to prevent or revert discriminatory outcomes. Moreover, the DSA mandates that platforms address notices and complaints in a **timely** manner. In practice, this requires platforms to establish efficient moderation workflows that prioritise urgent or sensitive cases, such as those reported by trusted flaggers or involving journalistic accounts. While speed is important, platforms must balance it with thoroughness; complex cases may justifiably take longer to resolve to ensure accurate decisions, which may in some instances – like the assessment of potentially terrorist content-verification by a human moderator. To **avoid arbitrary** decisions, platforms should have clear guidelines that limit moderator discretion, provide transparency, and ensure consistency in moderation actions. The principle of **proportionality** requires that any restrictions on users' rights are appropriate and necessary, matching the severity of the violation and, as such, human involvement may turn crucial in those cases where proportionality may be at stake.

The framework does not appear to mandate a specific choice between HOC, HIC, or HIL systems, at least not in the initial moderation stage. In the complaint-handling stage, the requirement
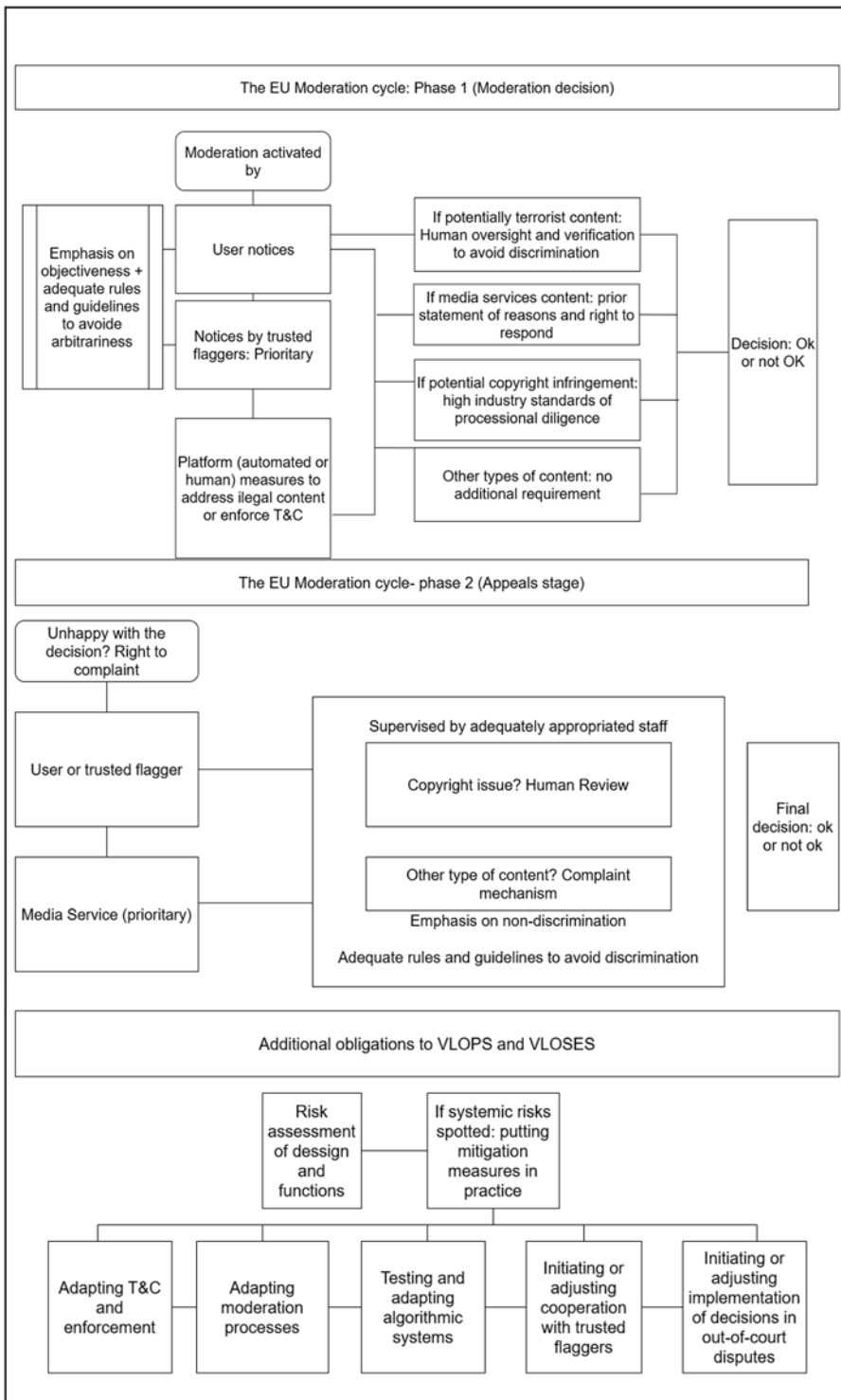
Figure 4. The moderation workflow and its requirements.[19]

for 'supervision under qualified staff' suggests a preference for HIC due to its managerial approach, although the language allows room for interpretation. Additionally, as noted earlier, 'diligence' may vary across companies based on factors like size and service type.

7. Conclusions

Policy and corporate decisions to automate, to not automate, or to semi-automate moderation processes involve a complex balance of factors. Automation may be attractive due to the possibility of speeding up large-scale processes, but in the moderation context may come at a risk of endangering human rights. On the contrary, moderation performed solely by humans can bear an economic and well-being cost. Semi-automation comes as a balanced but still complex solution in which the aim should be avoiding the inclusion of humans as mere rubber-stampers of automated tools.

While some ethical and human rights oriented documents tend to point at human intervention of different degrees as a key safeguard for freedom of expression in content moderation, its value does not rely on the mere incorporation of a rubber stamper. Instead, a meaningful human intervention in these processes requires a thoughtful analysis of different factors to determine things like the number of moderators, their qualifications, and the moments in which they should intervene. Neither automated tools or human moderators are likely to work in a perfect way, as different contexts and the passing of time is likely to continuously bring new challenges. As such, the discussion should not be centred on debating whether human or automated moderation is better, but instead how to make the best combination of both.

While the DSA has emerged as a general rule providing leeway to the implementation of human intervention in content moderation, some elements of a piecemeal approach can be seen in the way in which legal requirements were introduced to very specific scenarios, such as copyright and anti-terrorism enforcement. As such, the DSA and its complementary legislation, read under a human rights lens, emphasise the call for a more qualified involvement. While the incorporation of these and other similar standards into other forms of automated decision-making processes in the public and private sector is something that needs further research, it can be said at this point that the way in which platforms incorporate humans into the moderation cycle to ensure diligence, objectivity, non-discrimination, timeliness, non-arbitrariness, and proportionality in their processes will be essential to keep them off the hook.

References

Addink, H. (2019). Good governance: An introduction. In H. Addink (Ed.), *Good Governance* (1st, 3–14). Oxford University Press: Oxford.

Adiletta and Others v. Italy, 13978/88, 14236/88, 14237/88 (ECtHR 19 February 1991). https://hudoc.echr.coe.int/eng?i=001-57671

Ajunwa, I. (2020). The paradox of automation as anti-bias intervention. *Cardozo Law Review*, *41*(5), 1671–1742.

Al-Dulimi and Montana Management Inc. v. Switzerland, 5809/08 (ECtHR [GC] 21 June 2016). https://hudoc.echr.coe.int/eng?i=001-164515

Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and

Amending Certain Union Legislative Acts, P9_TA(2023)0236, European Parliament (2023). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Article 29 Data Protection Working Party. (2016). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (17/EN WP 251).

Aza, A. (2024). Scores as decisions? Article 22 GDPR and the judgment of the CJEU in *SCHUFA holding (scoring)* in the labour context. *Industrial Law Journal*, dwae035. https://doi.org/10.1093/indlaw/dwae035

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Barral Martínez, M. (2023). Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union. *JIPITEC*, *14*(1). http://www.jipitec.eu/issues/jipitec-14-1-2023/5716

Bayer, J. (2024). The place of content ranking algorithms on the AI risk spectrum. *Telecommunications Policy*, *48*(5), 102741. https://doi.org/10.1016/j.telpol.2024.102741

Bertuzzi, L. (2023). *European Union squares the circle on the world's first AI rulebook – Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/european-union-squares-the-circle-on-the-worlds-first-ai-rulebook/

Binns, R., & Veale, M. (2021). Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR. *International Data Privacy Law*, *11*(4), 319–332. https://doi.org/10.1093/idpl/ipab020

Bonnitcha, J., & McCorquodale, R. (2017). The Concept of 'Due Diligence' in the UN Guiding Principles on Business and Human Rights. *European Journal of International Law*, *28*(3), 899–919. https://doi.org/10.1093/ejil/chx042

Bovenzi, G. M. (2024). Content moderation in (decentralized) metaverses. *Proceedings of the International Congress Towards a Responsible Development of the Metaverse*, 13-14 June 2024. https://catedrametaverso.ua.es/papers/

Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology*, *27*(2), 91–121. https://doi.org/10.1093/ijlit/eay017

Bulgakov v. *Russia*, *20159*/15 (ECtHR 23 June 2020). https://hudoc.echr.coe.int/fre?i=001-203181

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Case C–401/19 Republic of Poland v European Parliament, Council of the European Union, ECLI:EU:C:2021:613, Opinion of AG ØE (ECJ 2021).

Case C–634/21 OQ v Land Hessen, Case C–634/21 OQ v Land Hessen, Intervener: SCHUFA Holding AG (ECJ 7 December 2023). https://curia.europa.eu/juris/document/document.jsf;jsessionid=3E818B5F5D8196E9F077021685F35116?text=&docid=280426&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=569037

Case C-136/17 GC, AF, BH, ED v Commission Nationale de l'informatique et Des Libertés (CNIL) (ECJ 24 September 2019). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOC_2019_399_R_0002.

Case C-740/22 Endemol Shine Finland Oy (ECJ 7 March 2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62022CJ0740.

Charter of Fundamental Rights of the European Union, 326 OJ C (2012). http://data.europa.eu/eli/treaty/char_2012/oj/eng

Coche, E. (2023). Countering Terrorism Propaganda Online Through TERREG and DSA: A Battlefield or a Breath of Hope for Our Fundamental Human Rights? In D. M. Vicente, S. D. V. Casimiro & C. Chen (Eds.), *The Legal Challenges of the Fourth Industrial Revolution* (vol 57, 313–333). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-40516-7_16

Commissioner for Human Rights. (2019). *Unboxing artificial intelligence: 10 steps to protect human rights*. https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights

Consolidated Version of the Treaty on European Union (2016). http://data.europa.eu/eli/treaty/teu_2016/oj/eng

Consolidated Version of the Treaty on the Functioning of the European Union (2016). http://data.europa.eu/eli/treaty/tfeu_2016/oj/eng

Constantino, J. (2022). Exploring article 14 of the EU AI proposal: Human in the loop challenges when overseeing high-risk AI systems in public service organisations. *Amsterdam Law Forum*, *14*(3), 1–17.

Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) (1950).

Council Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 328 OJ L (2008). http://data.europa.eu/eli/dec_framw/2008/913/oj/eng

Crootof, R., Kaminski, M. E., & Price Ii, W. N. (2022). Humans in the Loop. *Vanderbilt Law Review*, *76*(2), 429–510. https://doi.org/10.2139/ssrn.4066781

Cumhuriyet Vakfi and Others v. Turkey, 28255/07 (ECtHR 8 October 2013). https://hudoc.echr.coe.int/eng?i=001-126797

De Lemos, R. (2020). Human in the loop: What is the point of no return? *2020 IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, 165–166. https://doi.org/10.1145/3387939.3391597

Delfi as v. Estonia, 64569/09 (ECtHR [GC] 16 June 2015). https://hudoc.echr.coe.int/eng?i=001-155105

Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others, Joined Cases C–293/12 and C–594/12 (ECJ 8 April 2014). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0293

Douek, E. (2020). What kind of oversight board have you given us? *University of Chicago Law Review Online*, *2020*, 1–11.

Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, *136*(2), 526–607.

Enarsson, T. (2024). *Online Content Moderation in and Beyond the EU Digital Services Act – Exploring the Tension between Automated Speed and Human Contextuality* (SSRN Scholarly Paper 4806704). https://doi.org/10.2139/ssrn.4806704

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop – Legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123–153. https://doi.org/10.1080/13600834.2021.1958860

Endorsed WP29 Guidelines | European Data Protection Board. (n.d.). Retrieved 17 September 2024, from https://www.edpb.europa.eu/our-work-tools/general-guidance/endorsed-wp29-guidelines_en

Endsley, M. R. (2023). Ironies of artificial intelligence. *Ergonomics*, *66*(11), 1656–1668. https://doi.org/10.1080/00140139.2023.2243404

Enqvist, L. (2023). Human oversight' in the EU artificial intelligence act: What, when and by whom? *Law, Innovation and Technology*, *15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

EU Commission. (2020, December 15). Commission Staff Working Document—Annexes Accompanying the document proposal for a Regulation of the European Parliament and the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act

EU Commission. (n.d.). *Supervision of the designated very large online platforms and search engines under DSA | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses

Eweida and Others v. the United Kingdom, 48420/10, 59842/10, 51671/10, 36516/10 (ECtHR 15 January 2013). https://hudoc.echr.coe.int/eng?i=001-115881

Fiss, O. M. (1982). Objectivity and Interpretation. *Stanford Law Review*, *34*(4), 739. https://doi.org/10.2307/1228384

Frosio, G., & Geiger, C. (2023). Taking fundamental rights seriously in the Digital Services Act's platform liability regime. *European Law Journal*, *29*(1–2), 31–77. https://doi.org/10.1111/eulj.12475

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, *22*(7), 1266–1286. https://doi.org/10.1177/1461444820912540

Gillan and Quinton v. the United Kingdom, 4158/05 (ECtHR 12 January 2010). https://hudoc.echr.coe.int/eng?i=001-96585

Goldenfein, J. (2024). Forthcoming in Gavin Sullivan, Fleur Johns, and Dimitri Van Den Meerssche (eds) Global Governance by Data (Cambridge University Press, 2024). Lost in the loop: Who is the 'human' of the human in the loop. https://www.ssrn.com/abstract=4750634

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Griffin, R. (2022). The Sanitised Platform. *JIPITEC*, *13*(1). https://www.jipitec.eu/issues/jipitec-13-1-2022/5514

Griffin, R., & Stallman, E. (2024). Verfassungsblog.

Hacker, P. (2024) The AI Act between Digital and Sectoral Regulations, Gütersloh: Bertelsmann Stiftung. https://doi.org/10.11586/2024188

Holmberg, L. (2021). *Human In Command Machine Learning*. Malmö: [Malmö University, Department of Computer Science and Media Technology (DVMT)]. https://doi.org/10.24834/isbn.9789178771875

IAB Europe v Gegevensbeschermingsautoriteit, Case C–604/22 (ECJ 7 March 2024). https://curia.europa.eu/juris/liste.jsf?num=C-604/22&language=en

Independent High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for trustworthy AI*. Brussels: European Commission.

Jaidka, K., Chen, T., Chesterman, S., Hsu, W., Kan, M.-Y., Kankanhalli, M., … **Yue, A.** (2024). Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy. *Digital Government: Research and Practice*, *3689372*. https://doi.org/10.1145/3689372

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction*, *26*(5), 31:1–31:35. https://doi.org/10.1145/3338243

Jones, M. (2015). The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles. *Vanderbilt Journal of Entertainment & Technology Law*, *18*(1), 77.

Khadija Ismayilova v. Azerbaijan, 65286/13, 57270/14 (ECtHR 10 January 2019). https://hudoc.echr.coe.int/eng?i=001-188993

Kou, Y., & Gui, X. (2021). Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3411764.3445279

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022). Human-AI collaboration via conditional delegation: A case study of content moderation. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3491102.3501999

Langvardt, K. (2017). Regulating online content moderation. *Georgetown Law Journal*, *106*(5), 1353–1388.

Laux, J., Wachter, S., & Mittelstadt, B. (2021). Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Computer Law & Security Review*, *43*, 105613. https://doi.org/10.1016/j.clsr.2021.105613

Link, D., Hellingrath, B., & Ling, J.(2016). A human-is-the-loop approach for semi-automated content moderation: ISCRAM Association. http://idl.iscram.org/files/daniellink/2016/1401_DanielLink_etal2016.pdf Accessed October 10, 2024

Lupeni Greek Catholic Parish and Others v. Romania, 76943/11 (ECtHR [GC] 29 November 2016). https://hudoc.echr.coe.int/eng?i=001-169054

Malaihollo, M. (2021). Due diligence in international environmental law and international human rights law: A comparative legal study of the nationally determined contributions under the Paris agreement and positive obligations under the European convention on human rights. *Netherlands International Law Review*, *68*(1), 121–155. https://doi.org/10.1007/s40802-021-00188-5

Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019

McGonagle, T. (2019). The council of Europe and internet intermediaries: A case study of tentative posturing. In R. F. Jørgensen (Ed.), *Human Rights in the Age of Platforms* (227–253). Cambridge: The MIT Press.

Member States | UNESCO. (n.d.). Retrieved 23 September 2024, from https://www.unesco.org/en/countries.

Mendoza, I., and Bygrave, L. A. (2017). The right not to be subject to automated decisions based on profiling. In T.-E. Synodinou, P. Jougleux, C. Markou, T. Prastitou (Eds.), *EU Internet Law* (77–98). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-64955-9_4

Meta. (2023a). *Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook*. Meta. https://transparency.meta.com/sr/dsa-transparency-report-oct2023-facebook/

Meta. (2023b). *Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram*. Author. https://scontent-arn2-1.xx.fbcdn.net/v/t39.8562-6/447971060_1481740992549061_1404827436118992020_n.pdf?_nc_cat=100&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=sh73tBNSQgQQ7kNvgFbfDbd&_nc_ht=scontent-arn2-1.xx&oh=00_AYDZ1a7VO8NqAotRDE669f6uL9682Aa5skNqBPhHhk4kiA&oe=66A202AC

Meta. (2025, January 7). *More Speech and Fewer Mistakes | Meta*. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Milton, J. (1918). *Aeropagitica, with a Commentary by Sir Richard C. Jebb and with Supplementary Material* Jebb, Richard C Cambridge, United Kingdom: Cambridge at the University Press.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Muniz Da Conceição, L. H. (2024). A constitutional reflector? Assessing societal and digital constitutionalism in Meta's Oversight Board. *Global Constitutionalism*, 1–34. https://doi.org/10.1017/s2045381723000394

Nahavandi, S. (2017). Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, *3*(1), 10–17. https://doi.org/10.1109/MSMC.2016.2623867

Nourooz Pour, H. (2024). Voices and values: The challenging odyssey of Meta to harmonize human rights with content moderation. *International Journal of Law and Information Technology*, *32*(1), eaae009. https://doi.org/10.1093/ijlit/eaae009

Observer and Guardian v. the United Kingdom, 13585/88 (ECtHR 26 November 1991). https://hudoc.echr.coe.int/eng?i=001-57705

Ortolani, P. (2023). If you build it, they will come the DSA "procedure before substance" approach. In J. van Hoboken, J. Quintais, N. Appelman, R. Fahy, I. Buri & M. Straub, (Eds.), *Putting the DSA into practice* (151–166). Berlin: Verfassungsbooks. https://doi.org/10.17176/20230208-093135-0

Quintais, J. P., Frosio, G., van Gompel, S., Hugenholtz, B., Husovec, M., Jütte, B. J., & Senftleben, M. (2020). Safeguarding user freedoms in implementing article 17 of the copyright in the digital single market directive: recommendations from European academics. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, *10*(3), 277–282.

Recommendation of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, CM/Rec(2018)2, Council of Europe (Committee of Ministers) 1309th meeting of the Ministers' Deputies (2018). https://search.coe.int/cm#{%22CoEIdentifier%22:[%220900001680790e14%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}

Reporting on Pakistani Parliament Speech, 2023-038-FB-MR (Meta Oversight Board 4 April 2024). https://www.oversightboard.com/decision/fb-57spp63y/

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale: Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Roig, A. (2017). Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR). *European Journal of Law and Technology*, *8*(3). https://www.ejlt.org/index.php/ejlt/article/view/570

Romero Moreno, F. (2020). 'Upload filters' and human rights: Implementing article 17 of the directive on copyright in the digital single market. *International Review of Law, Computers & Technology*, *34*(2), 153–182. https://doi.org/10.1080/13600869.2020.1733760

Roose, K. (2024, March 21). *Reddit's IPO Is a Content Moderation Success Story—The New York Times*. https://www.nytimes.com/2024/03/21/technology/reddit-ipo-public-content-moderation.html

Rouas, V. (2022). Achieving access to justice in Europe through mandatory due diligence legislation. In *Achieving Access to Justice in a Business and Human Rights Context* 287–331. London: University of London Press. https://www.jstor.org/stable/j.ctv293p4bn.13

Sanchez, V. France 45581/15 (ECtHR [GC] 15 May 2023). https://hudoc.echr.coe.int/eng?i=001-224928

Sander, B. (2021). Democratic disruption in the age of social media: Between marketized and structural conceptions of human rights law. *European Journal of International Law*, *32*(1), 159–193. https://doi.org/10.1093/ejil/chab022

Schwemer, S. F., Mahler, T., & Styri, H. (2021). Liability exemptions of non-hosting intermediaries: Sideshow in the Digital Services Act? *Oslo Law Review*, *8*(1), 4–29. https://doi.org/10.18261/ISSN.2387-3299-2021-01-01

Scordino v. Italy, (No. 1), 36813/97 (ECtHR [GC] 29 March 2006). https://hudoc.echr.coe.int/eng?i=001-72925

Senftleben, M. (2023). Guardians of the UGC galaxy – Human rights obligations of online platforms, copyright holders, member states and the European Commission under the CDSM Directive and the Digital Services Act. *JIPITEC*, *14*(2023). http://www.jipitec.eu/issues/jipitec-14-3-2023/5847

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of Frontier AI, Generative AI, and Large AI Models. *2023 International Conference on Computer and Applications (ICCA)*, 1–7. https://doi.org/10.1109/ICCA59364.2023.10401723

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023). SoK: content moderation in social media, from guidelines to enforcement, and research to practice. *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 868–895. https://doi.org/10.1109/EuroSP57164.2023.00056

Solove, D. J., & Matsumi, H. (2024). AI, algorithms, and awful humans. *Fordham Law Review*, *92*, 1923–1940.

Steering Committee for Media and Information Society (CDMSI). (2021). *Content Moderation—Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation*. Council of Europe. Guidance Note https://rm.coe.int/content-moderation-en/1680a2cc18

Thompson, N. (2017). Instagram's CEO wants to clean up the internet—But is that a good @$ing Idea? *Wired*. https://www.wired.com/2017/08/instagram-kevin-systrom-wants-to-clean-up-the-internet/

Tosoni, L. (2021). The right to object to automated individual decisions: Resolving the ambiguity of Article 22(1) of the General Data Protection Regulation. *International Data Privacy Law*, *11*(2), 145–162. https://doi.org/10.1093/idpl/ipaa024

UNCHR. (2007). *General comment no. 32, Article 14, Right to equality before courts and tribunals and to fair trial*. UN Doc CCPR/C/GC/32.

UNCHR. (2018). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (UN Doc A/HRC/38/35).

UNCHR. (2021). *Disinformation and freedom of opinion and expression—Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (UN doc. A/HRC/47/25). https://www.ohchr.org/en/documents/thematic-reports/ahrc4725-disinformation-and-freedom-opinion-and-expression-report

UNCHR. (2022). *Terrorism and human rights - Report of the United Nations High Commissioner for Human Rights* (UN Doc A/HRC/50/49).

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNGA. (2018). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (UN Doc A/73/348).

UNGA. (2021). *Gender justice and freedom of expression—Report of Special Rapporteur on the promotion and protection of freedom of opinion and expression* (Un doc. A/76/258).

UNGA. (2022). *Disinformation and freedom of opinion and expression during armed conflicts—Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (UN Doc. A/77/288).

Vargas Penagos, E. (2024). ChatGPT, can you solve the content moderation dilemma? *International Journal of Law and Information Technology*, *32*(1), eaae028. https://doi.org/10.1093/ijlit/eaae028

Veale, M., Matus, K., & Gorwa, R. (2023). AI and global governance: Modalities, rationales, tensions. *Annual Review of Law and Social Science*, *19*(1), 255–275. https://doi.org/10.1146/annurev-lawsocsci-020223-040749

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, *11*(1), 104–122. https://doi.org/10.1002/poi3.198

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Young, G. K. (2022). How much is too much: The difficulties of social media content moderation. *Information & Communications Technology Law*, *31*(1), 1–16. https://doi.org/10.1080/13600834.2021.1905593

Emmanuel Vargas Penagos is a PhD researcher in Law at Örebro University, focusing on the Human Rights implications of automated content moderation. A lawyer from Los Andes University with an LLM in Information Law from the University of Amsterdam, his research spans freedom of expression, privacy, and AI. He has worked with organizations like UNESCO, Media Defence, the Inter-American Commission on Human Rights, and the Foundation for Press Freedom in Colombia. He also founded El Veinte, a Colombian NGO for freedom of expression.