Rethinking AI Content Moderation through an Intersectional Feminist Lens

Alexandra Wudel

Payal Padhy

Author Note

The authors declare that they have no conflicts of interest and are currently affiliated with Fem AI. This is an independent, non-commissioned research work conducted solely by the authors.

Correspondence concerning this article should be addressed to the authors at info@fem-ai.com

Keywords : *Content moderation, intersectionality, feminism, AI , framework*

Abstract

This chapter contributes to the broader discourse on AI and culture by shedding light on the intersection of AI content moderation and cultural dynamics, offering an intersectional feminist understanding for rethinking how these systems work. The authors examine the impact of cultural values on the design and the decision-making processes of AI content moderation systems. This chapter aims to bridge the gap in understanding the limitations of current content moderation systems, which fail to prioritise human-centric, intersectional values. With a feminist lens, this chapter considers how cultural biases and norms permeate AI-driven content moderation systems and impact users' experiences. By incorporating Delphi method the authors aim to inform strategies for redefining existing content moderation practices.

The current AI content moderation system which is reminiscent of modern colonisation is rooted in colonial practices, focussed on white-privileged, mostly male users. Through this research, the authors highlight the urgency of centering marginalised voices to promote more inclusive, diverse and safe online space for everyone.

Key themes and concepts that are addressed in this chapter include:

- **AI, Culture, and Content Moderation**: An examination of the interplay between AI content moderation and cultural dynamics from a feminist perspective, emphasising how the technology and its algorithms often reflect and exacerbate societal biases, reinforcing dominant power structures.
- **(Gendered) Biases in Content Moderation**: A critical analysis of current AI content moderation systems, their community guidelines and their policies which play a significant role in silencing marginalised voices, reinforcing gender biases and stereotypes.
- **Rethinking Content Moderation through a feminist lens** : Proposing strategies for framework development to redefining existing content moderation practices and policies, grounded in feminist principles. A discussion of how feminist theory can be applied to AI content moderation by taking into account the diverse and intersecting values of content moderators.

1. Introduction

“The world’s biggest tech companies today have more power and money than many governments.” (Hellerstein 2023).

In the age of AI, content has become the lifeblood of the internet, shaping our online experiences and interactions. From social media posts to user-generated videos, content permeates every corner of the virtual world. Undeniably, content moderation becomes important to our society as a societal challenge. As noted by Gillespie (2018), it is not merely an ancillary aspect of what platforms do; it is essential, constitutional, and definitional. Content moderation, therefore, is the practice of monitoring, filtering, and managing this vast array of online content to ensure that it aligns with platform guidelines, community standards, and legal regulations (Grygiel & Brown, 2019; Hartmann, 2020).

Philosophical inquiries on content moderation route back to Gengle's "Fairwitness" concept in 1978: the concept introduced the idea of impartial citizens serving as guides of online conversations. In the context of content moderation, the Fairwitness role emphasises the importance of impartiality and fairness in managing online discussions and disputes. Fair Witnesses are tasked with observing and facilitating conversations without bias or personal agenda, ensuring that all participants are heard and respected. Gengle's Fairwitness concept laid the groundwork for subsequent discussions and research on content moderation, emphasising the importance of impartiality and fairness in managing online communities and discussions (Benedikt, 1991). This derivation highlights the importance of people in the AI value chain.

However, the concept of content moderation is not one-size-fits-all. It varies significantly depending on the platform and its specific goals and functions. For instance, the way content moderation is defined and implemented on a social media platform like Facebook may differ greatly from how it operates within the context of General Purpose Artificial Intelligence (GPAI). The enormous increase in content has given rise to myriad challenges in content moderation. Human moderators who have been at the centre of this for almost two decades now face additional obstacles with the continuous nature of their work. The sheer volume of content generated on platforms like Facebook and Twitter(now X) each day requires constant vigilance to ensure compliance with community standards (Roberts, 2016; Wolfangel, 2023).

The standards for content moderation are often ambiguous and subject to interpretation. As explained by Gillespie (2018), the platform operators are not a holistic representation of their users which is majorly skewed towards white, privileged male who are overwhelmingly technological in their skills and worldview. Prior research has also consistently shown the disproportionate impact of content moderation on marginalised communities. Haimson et al. (2021) identified Black people, transgender individuals and political conservatives as the most disproportionately moderated groups amongst social media users. It has also been evidenced that women - mainly women of colour are often subjected to the harshest rules of content moderation (Gerrard 2020; Nurik 2019).

Moderators often face challenges navigating complex and subjective moderation systems. They are tasked with making nuanced distinctions between acceptable and unacceptable content, often without clear guidance or clarity. As examples, Instagram's inconsistent moderation of images of women's bodies (Witt et.al, 2019), banning images of "female-presenting nipples" on Tumblr (Duguay, 2018) or Instagram's restricted hashtags for women of colour like #blackgirls, #mixedgirls, #mexicangirls (Drewe, 2016) can be mentioned.

This ambiguity can lead to inconsistencies in policy enforcement, as highlighted by The Guardian's investigation into Facebook's moderation practices in 2017 (Hopkins, 2017).What has changed since then?

Large scale platforms are now increasingly pushing Artificial Intelligence (AI) as the solution to address harmful content burgeoning at an immense scale to successfully imitate human judgements while shielding human moderators from traumatic content (Gillespie, 2020). From a feminist lens, the power dynamics inherent in AI as well as on platforms play a crucial role in shaping content moderation practices. As AI technologies increasingly mediate our online experiences, questions of control, authority, and bias come to the forefront. Understanding the distribution of power within AI systems is therefore essential for comprehending the complexities of content moderation in the digital age and will be addressed in the research methodology.

The challenges inherent in content moderation are compounded by the relentless pace of online discourse, where misinformation, hate speech, and other forms of harmful content can spread

rapidly. Current research suggests that algorithmic content moderation often reinforces dominant power structures by inheriting normative biases in their systems (Binns, R. et al., 2017). It also indicates that its opaque and dynamic nature often conceals the underlying value choices and tradeoffs, with the primary focus being to optimise a single goal in a particular context. (Elkin-Koren, N. 2020). While valuable insights can be gleaned from these localised studies, they may obscure broader patterns and trends in moderation practices.

A holistic approach to content moderation promises to uncover these broader insights. Recent research by Jiang et al. (2023) demonstrates the variability of moderation practices across different contexts. By examining content moderation through a more comprehensive lens, we can better understand its operation and impact across diverse online communities beyond the stereotyped white-male privileged user.

A key aspect of content moderation is the trade-offs inherent in decision-making. Seering et al. (2020) suggest that the preference for automated tools in moderation decisions may stem from the evolving values within online communities. Moderators must balance the facilitation of cooperation with prevention of abuse, a task fraught with ethical and practical complexities. In our study design, we emphasise the importance of trade-offs between AI and Human content moderation approaches.

2. Types of Content Moderation

In the majority of our desk research, content moderation is analysed in relation to user platforms. It is typically used in platforms such as social media, dating, gaming or e-commerce. Content moderation can appear in several different forms depending on the type the product or service provided. Jiang et al. (2023) compares 86 content moderation articles to analyse the key characteristics of content moderation. In their findings, they present the trade-offs related to (1) moderation style, (2) moderation actions, (3) moderation philosophies and (4) moderation values which are tailored to cultural dimensions.

Social media companies heavily rely on "automated" content moderation, employing tools like keyword filters, spam detectors, and hash-matching algorithms.These tools, trained on vast amounts of online text data, supplement human moderation efforts. Despite significant investments in Large Language Models (LLMs) to identify violating content, challenges persist due to limited and lower-quality text data for many languages. Tech companies have attempted to address this by developing multilingual LLMs, trained on diverse text data from multiple languages simultaneously (Elswah, 2024). Our analysis will deepen the look into the interplay of Human and AI moderated content.

As a starting point, our research adopts the trade-off framework as the preliminary checklist provided by Jiang et al. 2023, to better comprehend the competing interests at play in content moderation decisions. It particularly focuses on moderator values as content moderation is

inherently dialectical, characterised by a continual negotiation of competing values and priorities.

A range of moderator values relevant to content moderation have been identified in the literature. Choi (2015) highlights the importance of helpfulness, informativeness, relevance, descriptiveness, and explicitness in assessing the quality of answers on social learning platforms. This points out the need for considering context-based factors while developing criteria for content moderation which can look different for different cultures and markets.

Amongst many researchers, Langvardt (2017) calls for legislation to ensure that content moderation practices align with First Amendment principles (U.S. Const. amend. I) of the United States Constitution, suggesting a focus on free speech and transparency. Singhal (2022) further explores the differences in content moderation between mainstream and narrow social media platforms, underscoring the need for tailored approaches. These studies collectively focus on the complexity of content moderation and the diverse values that must be considered. While existing literature often portrays moderation as a series of standalone actions, we argue for a more nuanced understanding that acknowledges the inherent trade-offs involved while observing an absence of feminist principles in current approach.

By examining the pros and cons of different moderation actions and interrogating the normative values that underpin them, we can gain a deeper insight into the complex dynamics of content moderation. In doing so, we emphasise on the urgency of reimagining content moderation

technology to center marginalised voices, fostering a more inclusive, diverse, and safer online environment for all. This aims to open pathways to devise strategies for redefining existing content moderation practices and policies, grounded in intersectional feminist principles. As part of our Delphi study, which will be introduced in the following chapter, an expert on content moderation will provide insights into her work to contribute practical and timely relevant insights.

3. Methodology

In our research, the authors employ an intersectional feminist approach as a method to minimise reductionism in research design by addressing unequal power distribution in AI (Helma, 2015). Intersectionality which originated in the late 20th century with the black feminist movement, was a term coined by Kimberlé Crenshaw (1989), an American civil rights advocate. Intersectionality is the idea that various aspects of a person's identity—like race, gender, class, and sexuality—intersect and interact, shaping their experiences of privilege and discrimination. It emphasises the interconnected nature of social categories that foster inequality and injustice. In applying intersectionality as a research method to minimise reductionism, the authors aim to build a more comprehensive understanding of the complexities of identity and power dynamics (Carastathis, 2014). Furthermore, “reflexivity” of methods and assumptions used will be applied to consider our own social position, role and power to set the direction of this study (Kirsch 1999, p. 3). In order to center marginalised voices, intersectionality encourages researchers and policy makers to look beyond single-identities or specific groups and enable more nuanced solutions rather than a “one-size-fits-all” approach (Hankivsky, O, 2022). Reflexivity of the

approach and analysis as well as the aim of centering marginalised voices in our research design will be the two main considerations of the applied feminist design.

However, the lack of transparency in AI systems presents significant challenges to a comprehensive feminist analysis. The lack of transparency is apparent in content moderation approaches, as the preliminary research phase for developing this book chapter has uncovered further obstacles such as the inaccessible nature of content moderation practices and the limited availability of open-source software for utilising AI in content moderation value chains.

In the initial stage, we conduct classic desk research to understand content moderation theories, their use cases and problems, drawing insights from academic journals and company websites. This desk research lays the groundwork for the subsequent phase, which is characterised by an intersectional research design framework where Delphi method is applied. Delphi methods are consensus-building techniques which, if designed from an intersectional feminist research approach, promote empowerment for vulnerable groups (Brady, S. R.,2015). Prioritising the perspectives of content moderators, the findings from the desk research will be enriched through engagement from diverse participants within a roundtable format. This roundtable hosts the execution of the Delphi study, supplementing the desk research through a collection of diverse views from domain experts in the field of content moderation. The “Feminist AI and Digital Policy Roundtable” is an existing format of FemAI for 24 months and will be re-designed for this study to tailor the Delphi design criteria.

By reframing moderation as more than just an occasional task that platforms undertake to making it a core component of their service and a crucial element of their role in public discourse, this book chapter aims to reassess the nature of platforms themselves from a feminist perspective by analysing power dimensions. This shift in perspective prompts to re-think societal influence wielded by these platforms.

3.1 Design Process

3.1.1 Delphi Study

The dynamics between the rapid pace of innovation in the field of AI and the necessity for a cautious approach, mindful of ethical and moral principles, presents a challenge that traditional analytical techniques alone cannot fully address (Linstone and Turoff, 1975). The Delphi method emerges as a potential avenue to navigate this tension. Originally conceived as a written survey technique (Mayer and Köck-Hódi, 2013), the conventional Delphi method typically involves a structured questionnaire circulated among a group of experts. The Delphi method can also be used to determine connections between socio-cultural issues or research into personal values (Linstone and Turoff, 1975). It allows a structured survey process of a group, which is primarily aimed at consensus building on complex issues (Linstone and Turoff, 1975; Häder, 2014).

The design of the Delphi process should be tailored to suit the specific problem and objectives at hand. While this adaptability enhances the method's flexibility, it also introduces complexity and diminishes the clarity of existing Delphi methodologies. Nevertheless, this flexibility offers researchers the opportunity for a more customised and individualised approach to their research

endeavours (Häder, 2014.). The following diagram operationalises the Delphi process by applying it to an intersectional feminist research approach.

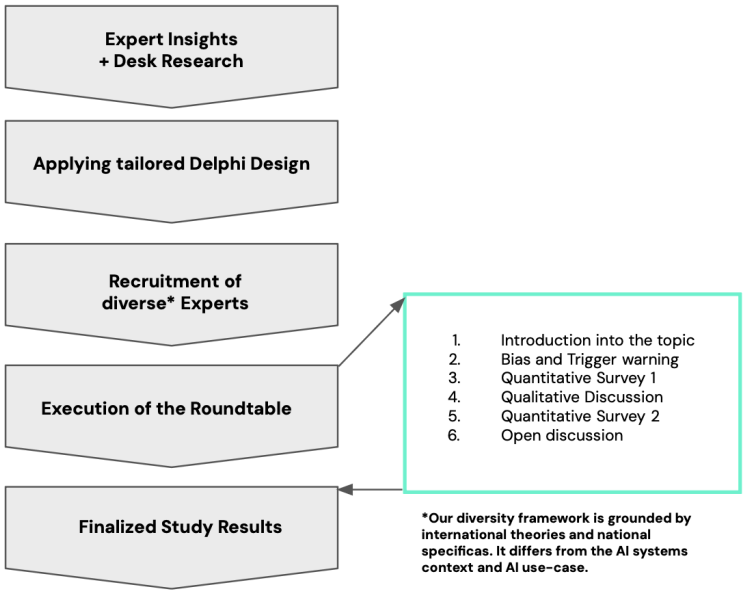


Fig 1. Intersectional feminist Delphi process

Next to a feminist research design prioritising the diverse nature of the expert selection including civil society and using accessible language, the selection of the sources presented are based on a feminist lens.

The relation between AI content moderation, racism and decolonisation is a critical factor to consider in terms of cultural dimensions. Anibal Quijano (2007a) proposed the concept of the coloniality of power to elucidate and conceptualise the emerging patterns of authority in a post-colonial milieu. Despite attaining independence, former colonies remain ensnared in the grip of Western dominance. Quijano's perspective holds relevance today as it places race at the forefront and acknowledges its fluid role as a tool of power within the global framework.

New technologies do not arise independently of existing systems; instead, they are intertwined with current frameworks of economic, political, and social power. Put simply, race and racism are inherent within technologies, and critical researchers must grapple with understanding their role in perpetuating dominant power structures (Gengler et. al. 2023).

Certainly! Here's a revised version:

The employment of content moderators in lower-cost countries, particularly from the Global South, often working under precarious conditions, evokes parallels with modern-day colonisation (Elswah, 2024). In the case of ChatGPT-3, the training method "Reinforcement Learning from Human Feedback (RLHF)" has been utilised in 2023. The media shed light on this pressing concern: an article published in Zeit (Wolfangel, 2023) revealed that Kenyan workers were required to view content depicting violence and racism for ChatGPT to learn. Kenyan content moderation workers are screening “recordings of people being beheaded, burned alive, eaten” for a payment of 2 Dollars a day causing mental harm such as sleeping disorder and anxiety (Hellerstein 2023). This underscores the risk of AI remaining a technology primarily serving the interests of the powerful in the future while building their success on outsourcing models to the Global South. Our Delphi process emphasises the importance of the Global South perspective by providing a 10 minutes insights presentation from the content moderation expert Adi Engel working on mental health solutions for content moderators in the Global South (Frame Ex, 2024).

3.1.2 Addressing potential design biases

To increase the robustness of the study results, the following section will address potential biases occurring in Delphi studies. Compared to classic interviews or focus groups, the Delphi method has a lower susceptibility to bias (Mukherjee et al., 2018). To further increase the robustness of the survey results, the main four biases will be addressed in the following section. (Winkler and Moser, 2016). Significant biases in Delphi surveys are:

- Framing
- Desirability Bias
- Bandwagon effect
- Persistence of faith (Winkler and Moser, 2016)

Framing

Framing biases describe the differences in responses based on how information is presented or framed. The most effective tool to prevent framing biases in forward-looking Delphi studies is maximisation of the participants' heterogeneity. Study participants who like to take a loner perspective are particularly special in terms of addressing this bias. (Winkler and Moser, 2016).
To further address potential framing effects, a warning at the start of the survey helps to indicate the arisement of framing effects (Cheng and Chin-Shan, 2010). Due to a high level of expertise and involvement of the experts, the participants' attention span may increase, which leads to a reduction of framing biases (Cheng and Chin-Shan, 2010, Winkler and Moser, 2016).

Desirability Bias

The desirability bias describes the potential bias on opinions due to optimistic or pessimistic expectations in the relation to the desired outcome. In addition to bias due to personal attitude, strategic pessimism or optimism may also lead to distortions. If survey participants hold decision-making roles and influence strategic planning, their personal expectations can significantly impact the outcomes, making them challenging to mitigate (Zlatan and Windschitl, 2007). Desirability bias in Delphi studies arises unavoidably due to the research design, which relies on the selection of experts. However due to the reduction of the considered time horizon, the reduction of desirability biases can be further addressed (Ecken et al., 2011).

Bandwagon effect

The objective of the Delphi method is finding consensus (Linstone and Turoff, 1975; Häder, 2014). This goal contradicts the appearance of the bandwagon effect as one of the study's main potential biases. A study by Bolger et. al., 2011 points to the value of the best argument within consensus building. The heterogeneity of the expert group helps with this reduction of bandwagon effects. (Winkler and Moser, 2016). However, these circumstances present a contradiction between method and susceptibility to error.

The bandwagon is one of the most influential biases in classic group interviews. By design, the Delphi study reduces the extent of the bandwagon effect due to the mechanics of controlled feedback. Compared to the classic group discussion, Delphi study has the advantage of fewer dominant participants that may bias opinions of less influential participants' (Mukherjee et al., 2018). The bandwagon effect is therefore of less importance than in a face-to-face research

design (Winkler and Moser, 2016). In addition, anonymity in Delphi studies also leads to a reduction in potential (Bolger et al., 2011).

Persistence of faith

Similar to the anchoring effect, the persistence of faith bias describes how experts tend to maintain their initial viewpoints despite encountering a wider range of insights. Participants insist on his or her opinion (Yaniv and Milyavsky, 2007). The lack of transparency about individual value sets and personal motivation may influence the intensity of persistence of faith (Bolger et al., 2011).

The following table provides an overview of the biases presented, categorises them and points out appropriate avoidance strategies. The highlighted suggestions for reducing potential bias are adapted to the Delphi design of this survey with the aim of increasing the robustness of our results.

Bias	Description	Avoidance Strategy
Framing	The way a question is presented alters the outcome.	● **Warning of framing and anchoring effects** ● **Expertise** ● **Involvement** ● **Heterogeneity** ● Lone wolf mentality ● Avoiding anchor points
Desirability Bias	The likelihood of an event is perceived as higher/lower due to personal preferences.	● Integration of a question regarding personal attitude (optimistic or

		pessimistic) ● Identification of questions with high desirability bias susceptibility ● **Reduction of time horizon**
Bandwagon effect	Conformity between individual opinion and group assessment.	● **Anonymity** ● Filtered reasoning in the foreground ● **Heterogeneity**
Persistence of faith	Loyalty to one's opinion despite conflicting information.	● Examples of good and bad arguments ● **Controlled feedback**

Table 1. Potential biases of the Delphi study and their avoidance strategies

3.2 Planning and Executing the delphi design

After laying down the key principles for the Delphi design process, this section outlines feminist design principles of the roundtable executing the Delphi study. As explained in the beginning of the section, Intersectionality, a central concept in feminism, offers a nuanced understanding of power dynamics (Ferguson, 2017). It acknowledges the complex interplay of patriarchal systems, affecting individuals in various ways and encompasses diverse aspects of marginalised, under-represented and under-privileged groups such as gender, age, religion, disability, sexual identity, ethnicity, and appearance (Ferguson, 2017). It illustrates how individuals face discrimination based on multiple attributes simultaneously. This intersection of multiple marginalised identities often results in compounded discrimination (Crenshaw, 1989). As AI tends to mirror traits devalued in patriarchal systems, such as those of women of colour, they are particularly vulnerable to discrimination by AI systems (Buolamwini & Gebru, 2018; Crenshaw,

1989). Intersectional feminism offers a lens to analyse and reform AI's operation at both the system and societal levels.

The roundtable is intended to spark conversation and build a consensus between the participants for finding an answer to our research question :

" *What would an AI content moderation system with an intersectional feminist lens look like*?"

For the purpose of this study and to have a common understanding amongst the roundtable participants, we have adopted the following definitions :

- Intersectionality : *It is a framework that acknowledges the interconnectedness of various forms of oppression such as race, gender, class and sexuality* (Ferguson, 2017)
- Content Moderation : *The governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse* (Grimmelman, 2015)

3.2.1 Selection of Participants

The selection of experts in Delphi studies is a vulnerable part as it directly influences the robustness of the results. Participant selection has prioritised diversity across multidimensionality, trans-disciplinarity, and intersectionality, thereby facilitating the incorporation and examination of disparate viewpoints (Sundermeier et al., 2020).

We applied this theory through an open recruitment call for participants across various digital platforms and including demographic information about participants which provided insights

about their intersectional values. While sensitive information like name, organisation and designation were kept anonymous, the following demographic details were collected to better understand the intersectional values that impact tradeoffs in decision making.

Gender/ Gender Pronouns	**Expertise in Content Moderation**	**Professional Expertise** (These are moderator identities defined by their social roles)	**Community Identity**	**Country**
Example : Male/ Female/ Trans/ Do not wish to respond	Expert ,or Learning the ropes, or General User	Example: Policy, Feminism, Human Rights etc	Academia, or Industry, or Civil Society, or Government, or Other	Participants were asked to mention the country that they most closely associated themselves with (For example : Country at birth, or most lived/researched, adopted country)

Table 2. Table showing the demographic details captured for each questionnaire participant

3.2.2 The Feminist AI and Digital Policy Roundtable Design

The roundtable session was compiled and moderated by both authors. The moderators maintained a passive role, ensuring inclusivity, respectfulness, and relevance in the discourse with a special regard on empowering all people to engage. The session was documented through video recording with unanimous participant consent.

The authors set the context by introducing the participants to the aim of the topic and the roundtable which was augmented by the insights of Adi Engel (Frame Ex, 2024). The roundtable consisted of two rounds of questionnaires as mentioned in Fig.1 in section 3.1. These questionnaires were designed to probe the participants on 20 different content types which were based on the content categories developed by Scheuerman et.al. (2021).

Category Number	Content Category	Example
C1	Adult sexual exploitation	sextortion, adult non-consensual sexual images
C2	Platform Abuse	spam, fake accounts
C3	Targeted attacks on an individual person	bullying, harassment
C4	Pornography	pornographic images of adults
C5	Coordinated attacks that promote harming people	terrorist propaganda, hate groups
C6	Violence	credible threats of injury or murdeR
C7	Self Injury and Eating disorder promotion	images of pro-anorexic content
C8	Graphic violence imagery	Images of gore violence
C9	Direct harm to children	child pornography, child grooming
C10	Animal Abuse	videos or images of animals being harmed physically
C11	Sale of firearms and sexwork (prostitution)	things that are illegal or highly regulated to be sold online
C12	Hate speech	slurs, hateful language used against protected groups
C13	Coordinating scams and political attacks	misinformation, fraud, voter suppression
C14	Hard drug sales	advertisement for hard drug sales
C15	Informative images of graphic violence	disturbing journalistic images of war
C16	Suicide promotion	posting videos that promote suicide
C17	Regulated goods	selling marijuana, alcohol, tobacco etc
C18	Child Nudity	images posted by trusted guardians of their infants or children like in a bathtub
C19	Adult Non-sexual nudity and Digital nudity	images or drawings of nude adults
C20	Coordinated attacks that directly harm people	human trafficking, mass murder coordination

Table 3. Table showing the code number, content categories and their specific examples for questionnaire 1 and 2.

Each participant was asked to rank the content categories (C1-C20) across a Likert scale of 1-5 where

1- represented completely human moderated

2- represented more human moderation and less AI moderation

3- represented equal level of moderation from both humans and AI

4- represented more AI moderation and less human moderation

5- represented complete AI moderation

These two questionnaires which were similar in design and conducted in two stages were interjected by an open discussion to contemplate trade-offs. To steer the discussion and reduce biases like desirability bias and persistence of faith as mentioned in section 3.1.2, the authors collected qualitative information on the following three questions:

1. What are the practical strategies for integrating feminist values into AI content moderation ?
2. What are the potential challenges in integrating intersectional feminist values into AI content moderation systems?
3. What are the ethical considerations in integrating intersectional feminist values into AI content moderation systems ?

The participants were given a trigger warning before being shown the questionnaire as some of the content might be disturbing for them. By integrating expert sessions, two stages of questionnaires and open discussions, the authors aimed to minimise the four biases of Delphi method as mentioned in section 3.1. It also comprehensively encourages reflexivity among both researchers and the participants by fostering critical reflection, iterative analysis and dialogue between the participants.

4. Results

The data analysis followed by the outcomes and the macro trends aims at rigorous categorisation of data as we identify relevant patterns and categorise them into broader content moderation themes that would form the basis of the feminist framework. It also suggests that this study has incorporated methods to mitigate bias and include reflexivity at each stage of the process.

- The researchers were constantly self-reflective about their identity and role and ensured diverse participation with acknowledgement of participant background and identities.
- The inclusion of an open discussion between two questionnaires allowed participants to share their perspective and insights. It encouraged participants to acknowledge their own biases and accept alternate viewpoints thus mitigating persistence of faith bias.
- The anonymity of the questionnaire helped in reducing the bandwagon effect. Moreover, it effectively minimised researcher bias and increased trust among the roundtable participants. Importantly, this approach aligns with feminist research principles by empowering marginalised voices to express themselves freely without prejudice or adhering to societal norms.
- In addition to quantitative insights, qualitative insights were also collected through the open discussion which would provide a comprehensive understanding of the experts' perception about content moderation. These insights would serve as a fundamental element to design the feminist framework for content moderation.

4.1 Data analysis

11 experts participated in the roundtable with 10 of them consenting to fill out the questionnaires. Of these participants, 6 were female and 4 were male. Half of the participants were experts in content moderation, while 40% had some exposure to content moderation. Academia represented the highest percentage participation at 40%, followed by equal percentage of participation from industry, civil society and government. With a heterogenous set of experts, we were able to include a diverse set of social roles into our study and validate its importance in defining moderator identities.

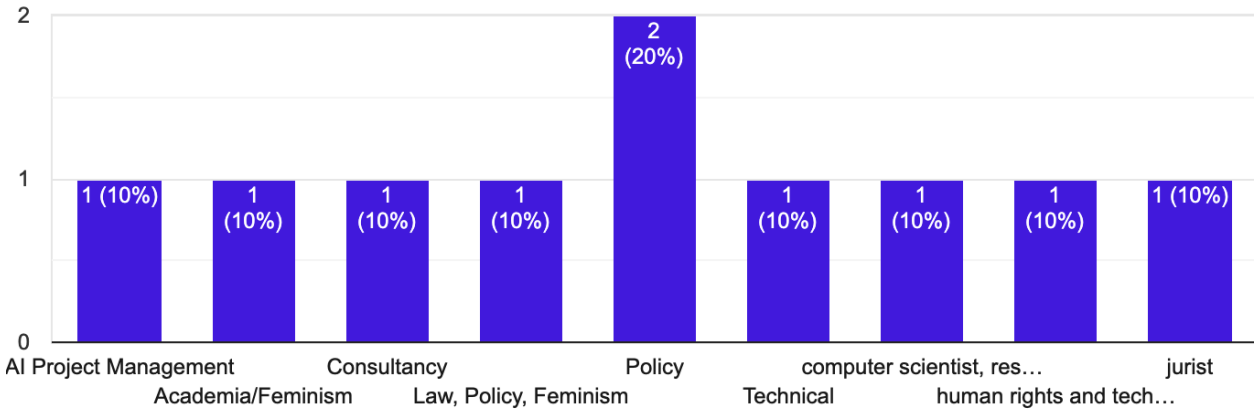


Table 4. Table showing number of participants and their corresponding professional roles which are considered as their moderator identities for the purpose of this study

We organised the content categories (C1- C20) into five broad categories which helped to contextualise the questionnaires into themes for content moderation. It also aims to discover practical strategies for balancing varied value systems in content moderation rule settings while establishing the intersectional feminist framework for content moderation.

Content moderation categories	**Content categories(from questionnaire)**
Sexual Exploitation and Harassment	C1. Adult sexual exploitation C3. Targeted attacks on an individual person C4. Pornography C7. Self Injury and Eating disorder promotion C16. Suicide promotion
Violence and Gore	C6. Violence C8. Graphic violence imagery C10. Animal Abuse C15. Informative images of graphic violence
Harmful Coordinated Attacks	C5. Coordinated attacks that promote harming people C20. Coordinated attacks that directly harm people C13. Coordinating scams and political attacks
Illegal or Regulated Activities	C2. Platform Abuse C11. Sale of firearms and sex work (prostitution) C14. Hard drug sales C17. Regulated goods
Hate Speech and Controversial Content	C9: Direct Harm to Child C12. Hate speech C18. Child Nudity C19. Adult Non-sexual nudity and Digital nudity

Table 5 : Table showing the broad content moderation categories proposed by the authors for the content (C1-C20) that were ranked by participants.

Due to the small sample size of our study and the robustness of the median to outliers, we used median as the statistical tool for our Delphi method analysis. This allowed us to effectively analyse the ordinal data and investigate if tradeoffs in moderator values impacted decision making in both the rounds.

In the following table, the content categories have been sorted in ascending order based on their median values in round 2 to showcase the development between the two rounds.

Ranking
1 represents completely human moderated
2 represents more human moderation and less AI moderation
3 represents equal level of moderation from both humans and AI
4 represents more AI moderation and less human moderation
5 represents completely AI moderated

Code	Round 1	Round 2	Development2
3	2,20	2,25	0,05
13	2,50	2,38	-0,13
20	2,40	2,63	0,23
5	2,80	2,63	-0,18
18	3,10	2,75	-0,35
16	3,00	2,88	-0,13
1	2,80	3,00	0,20
6	2,80	3,00	0,20
9	3,10	3,00	-0,10
19	3,10	3,00	-0,10
7	3,00	3,13	0,13
11	3,10	3,13	0,02
4	3,60	3,25	-0,35
15	2,60	3,38	0,78
17	3,50	3,38	-0,13
14	3,60	3,50	-0,10
12	3,20	3,63	0,43
2,0	3,90	3,63	-0,28
10	3,80	3,88	0,08
8	3,90	3,88	-0,02

Table 7 : The table shows the content categories (C1- C20), their corresponding median values for Round 1 and Round 2 and the development from Round 1 to Round 2. It is sorted in ascending order according to the median value in Round 2.

Round 1 - represents the initial assessment from the survey participants, and

Round 2 - represents a subsequent round of responses, after the open discussion as mentioned in section 3.2.3

Development - indicates the change between Round 1 and Round 2.

Positive values indicate improvement, while negative values indicate deterioration. An increase indicates that the participants are leaning towards a more AI moderated approach. A decrease indicates that the participants are leaning towards a more human moderated approach.

4.2 Outcome

Across the five broad categories some pertinent insights were observed through qualitative and quantitative analysis.

4.2.1 Sexual Exploitation and Harassment

Overall, there are mixed perceptions about the type of content moderation for this category. Across both the rounds, the experts preferred more human moderation than AI moderation. Content like pornography and suicide promotion showed inclination for more human moderation from round 1 to round 2, but there was a positive shift with a slight increase in preference for more AI moderation for content like adult sexual content, targeted attacks on an individual and content related to self injury/eating disorder.

- C1: Round 1 = 2.8, Round 2 = 3.0, Development = 0.20
- C3: Round 1 = 2.2, Round 2 = 2.25, Development = 0.05
- C4: Round 1 = 3.6, Round 2 = 3.25, Development = -0.35
- C7: Round 1 = 3.0, Round 2 = 3.13, Development = 0.13
- C16: Round 1 = 3.0, Round 2 = 2.88, Development = -0.13

4.2.2 Violence and Gore

Content related to animal abuse and graphic violence imagery show a leaning towards more AI content moderation than human moderation. There is a significant spike from human moderation to AI moderation for informative images of graphic violence.

- C6: Round 1 = 2.8, Round 2 = 3.0, Development = 0.20

- C8: Round 1 = 3.9, Round 2 = 3.88, Development = -0.02
- C10: Round 1 = 3.8, Round 2 = 3.88, Development = 0.08
- C15: Round 1 = 2.6, Round 2 = 3.38, Development = 0.78

4.2.3 Harmful Coordinated Attacks

Overall the preference for more human moderation is consistent across this category with content like promotion of coordinated attacks on people and political scam show a slight increase in preference for human moderation in round 2 as compared to round 1.

- C5: Round 1 = 2.8, Round 2 = 2.63, Development = -0.18
- C20: Round 1 = 2.4, Round 2 = 2.63, Development = 0.23
- C13: Round 1 = 2.5, Round 2 = 2.38, Development = -0.13

4.2.4 Illegal or Regulated Activities

Median scores generally decreased from round 2 to round 1 with a tendency to have equal involvement from both humans and AI for sale of firearms and sexwork but preference for increased AI content moderation for other content like platform abuse, hard drug sales or regulated goods.

- C2: Round 1 = 3.9, Round 2 = 3.63, Development = -0.28
- C11: Round 1 = 3.1, Round 2 = 3.13, Development = 0.02
- C14: Round 1 = 3.6, Round 2 = 3.5, Development = -0.10
- C17: Round 1 = 3.5, Round 2 = 3.38, Development = -0.13

4.2.5 Hate Speech and Controversial Content

Preference for AI moderation sees a significant jump in case of hate speech but all the other content which includes child nudity, harm to children and adult non-consensual nudity see a slight dip and demand equal participation from humans and AI.

- C9: Round 1 = 3.1, Round 2 = 3.0, Development = -0.10
- C12: Round 1 = 3.2, Round 2 = 3.63, Development = 0.43
- C18: Round 1 = 3.1, Round 2 = 2.75, Development = -0.35
- C19: Round 1 = 3.1, Round 2 = 3.0, Development = -0.10

More AI Moderated	**More Human Moderated**
Round 1 = 3.9 C2: Platform abuse Category: Illegal or regulated activities	Round 1 = 2.2 C3: Targeted attacks on an individual person Category: Sexual exploitation and harassment
Round 1 = 3.9 C8: Graphic violence imagery Category: Violence and Gore	Round 2 = 2.38 C13: Coordinated scams and political attacks Category: Harmful coordinated attacks
Round 2 = 3.88 C10: Animal Abuse Category: Violence and Gore	Round 1 = 2.4 C20: Coordinated attacks that directly harm people Category: Harmful coordinated attacks

Table 8 : The table shows the content which is most preferred to be AI moderated and most preferred to be human moderated by the experts at the roundtable

4.3 Macro Trends

The results show a small difference between the rounds. The authors analyse these results as:

Robustness of Interpretation: With similar deviation across categories, the interpretation of the survey results becomes more robust and reliable. Participants are less likely to be influenced by outlier responses or inconsistencies in ratings, leading to a more coherent analysis of the data.

Consensus Building: By ensuring similar deviation across categories, the Delphi-like study design facilitates consensus building among participants. This approach allows for a more balanced interpretation of the survey results, as participants are likely to converge toward a common understanding of the importance of different content moderation categories.

- The results in round 1 range between 2,2 and 3,9 with an average of 3,1. The span between these results is 1,7 which can be rated as moderate.
- The results in round 2 range between 2,3 and 3,9 with an average of 3,1. The span between these results is 1,6 which can be rated as moderate.
- The deviation from round 1 to round 2 ranks between 0,78 and -0,35 and has a span of 1,13 and a median of -0,10

These results indicate that the experts suggest a collaboration between human and AI content moderation. There are no responses supporting moderation by only AI or humans.

5. Rethinking content moderation

5.1 Outlook

The applied research design centred a German perspective on content moderation due to the regional background of the selected experts. Although the Delphi study was nurtured with insights from the Global South and complemented with on-ground insights from Adi Engler (Frame Ex, 2024), the cultural dimensions were not addressed significantly enough. This is why the authors suggest executing this research design on a national level in emerging countries to foster a culturally applicable framework of rethinking content moderation.

Tech companies have been found to apply policies that hinder individuals from freely expressing themselves or accessing information in their native language, particularly in non-English content moderation. This disproportion is even more glaring for low-resource languages, leading to account suspensions, amplification of hate speech, and spread of misinformation in the Global South (Elswah, 2024).

Thus, there is a pressing need to examine how technology firms moderate non-English languages. Data annotation, a process heralded for its potential to generate more job opportunities in low-income countries (Kshetri, 2021) can serve as an approach in developing data sets that are representative of diverse populations. This strategy holds promise in addressing the current biases and the power dynamics plaguing the industry and eventually fostering equity (Wang, 2022).

Especially in countries that lack regulation on content moderation values and practices, investigation by research and civil society organisations plays a crucial role to monitor potential harm. From a European perspective, the authors suggest further research to apply feminist research findings on content moderation with current regulations such as the Digital Services Act on digital platforms (European Commission, 2023).

In terms of the consensus building, a clear value set can be interpreted. Participants have strong opinions on value sets and can agree on a view of content moderation from a feminist lens from a German perspective.

As technical compromises are needed in certain cases, the survey suggests an order of how to design a framework around feminist content moderation, in addressing these questions to be moderated by AI more than by humans.

5.2 Feminist Perspective

With the proliferating AI-assisted world of content moderation exemplified by the use of GPT-4 (Open AI, 2023) for content moderation and Meta's utilisation of AI systems to tackle harmful content (Meta, 2021), the imperative of integrating intersectional feminist principles into content moderation has become increasingly urgent.

Drawing on the insights from the research and building upon Jiang et al's. (2023) content moderation trade-off framework, the authors, in collaboration with their roundtable experts have identified potential strategies to rethink current content moderation systems through a feminist lens.

5.2.1 <u>People and Culture</u>

While the existence of a universally applicable global value system for content moderation is debatable , the International Human Rights Law (Steiner, 2008) can be a good starting point due to its global reach and acceptance. However, the authors acknowledge and call to address the inherent biases of power dynamics and adoption of Western norms as the default standard. Contextual cultural considerations, alongside the inclusion of moderators and data annotators

from diverse regions, plays a crucial role in ensuring a more inclusive and culturally sensitive content moderation system.

5.2.2 Transparency and Accountability

To truly understand the tradeoffs encountered by moderators, it is imperative to improve transparency and accountability in current moderation systems, which are often poorly understood and lacking in accountability (Gorwa, 2020). The moderators, often working remotely from emerging countries, face issues such as unfair pay and poor working conditions. In response, an intersectional feminist approach advocates for the inclusion of feminist voices and marginalised groups not only in tech development but also in the establishment of broad regulations for the industry and rule setting such as the development of community guidelines.

5.2.3 Balancing interests

The concept of an intersectional feminist framework aims to balance values and business viability, offering a holistic approach to content moderation. One challenge highlighted by this research is the need to balance the financial costs of moderation and at the same time mitigate the ethical risks inherent in the business. For instance, investing in mental health support for human moderators and data annotators of AI algorithms is crucial for mitigating these risks. As feminist researchers, we actively challenge the status quo and confront the inherent power imbalances present in content moderation practices.

Conclusion

In a landscape where tech giants wield immense power, content moderation has emerged as a critical societal challenge, influencing online experiences and interactions.

The conclusion of this comprehensive discussion on content moderation in the age of AI and its intersection with feminist principles highlights the urgent need for a reevaluation of existing systems. The authors attempt to build upon prior research and utilise the trade-off centred framework of content moderation, to forge a nuanced yet equitable feminist framework.

The complexities of content moderation are longstanding, with platforms grappling with inconsistent standards and disproportionate impacts on marginalised communities. The shift towards AI-driven moderation poses both opportunities and challenges, as algorithmic biases reinforce dominant power structures. The research indicates that the debate of human vs automated moderation systems is far from straightforward. While AI systems are increasingly taking over the tasks previously done by human moderators, the emergence of new roles like content labelling and data annotation does not inherently improve the working conditions. The necessity of the human-in-the-loop in content moderation systems, albeit challenging, demands solutions to ensure sustainability and fairness for those undertaking such roles. It also highlights the lack of transparency in the content moderation value chain, resulting in underrepresentation of marginalised communities in designing content moderation systems. These systematic shortcomings are exacerbated in the AI systems, which are currently heralded as solutions to the current moderation challenges. Consequently, there is increased censorship of such communities on platforms. This is why a holistic approach, incorporating diverse perspectives and addressing

power imbalances, promises to uncover broader insights and inform more inclusive moderation practices.

In light of these considerations, the integration of intersectional feminist principles into content moderation can be used to combat threats of modern content moderation. Strategies such as incorporating contextual cultural considerations through the complete value chain of content moderation systems, improving transparency and accountability measures through regulations and advocating for best practices to be followed by organisations, as well as balancing interests between values and business viability, are proposed to rethink current systems. Ultimately, this approach seeks to challenge the status quo, confront power imbalances, and create more equitable online spaces for all users.

The study, which currently takes into account a German perspective suggests the need for further exploration of feminist perspectives within regulatory frameworks such as the Digital Services Act in Europe. It also proposes broadening the scope of this study to incorporate national level research to identify common themes and elucidate cultural nuances that remain uncaptured. The authors advocate for think tanks, academia and the industry itself to critically analyse the western norms of content moderation, acknowledging that while content can be global but moderation needs to be contextualised.

References

Benedikt, M. (Ed.). (1991). *Cyberspace: first steps*. Mit Press.

Binns, R., Veale, M., Kleek, M.V., & Shadbolt, N. (2017). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. ArXiv, abs/1707.01477.

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological forecasting and social change*, *78*(9), 1500-1513.

Brady, S. R. (2015). The Delphi Method. In L. A. Jason & D. S. Glenwick (Eds.), Handbook of Methodological Approaches to Community-Based Research: Qualitative, Quantitative, and Mixed Methods(pp.97-107).NewYork:OxfordAcademic. https://doi.org/10.1093/med:psych/9780190243654.003.0007

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Carastathis, A. (2014). The Concept of Intersectionality in Feminist Theory. *Philosophy Compass, 9*, 304-314.

Cheng F.-F, Chin-Shan W (2010) Debiasing the framing effect: the effect of warning and involvement. Decision Support Systems 49 (3), S. 328–334.

Choi, E., Borkowski, M., Zakoian, J., Sagan, K., Scholla, K., Ponti, C., ... & Bielski, M. (2015). Utilizing content moderators to investigate critical factors for assessing the quality of answers on brainly, social learning Q&A platform for students: a pilot study. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Crenshaw, K. W. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine (pp. 139–168). In *University of Chicago legal forum*.Drewe, N. (2016). The hilarious list of hashtags Instagram won't let you search. *The Data Pack*, *10*.

Duguay, S. (2018). Why Tumblr's ban on adult content is bad for LGBTQ youth. *The Conversation*, *6*.

Ecken, P., Gnatzy, T., & von der Gracht, H. A. (2011). Desirability bias in foresight: Consequences for decision quality based on Delphi results. *Technological Forecasting and Social Change*, *78*(9), 1654-1670.

Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society*, *7*(2), 2053951720932296.

Elswah, M. (2024). Investigating Content Moderation Systems in the Global South. Center for Democracy and Technology. Retrieved from: https://cdt.org/insights/investigating-content-moderation-systems-in-the-global-south/

European Commission (2023). The impact of the Digital Services Act on digital platforms. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms.

Ferguson, K. E. (2017). Feminist theory today. *Annual Review of Political Science*, *20*, 269-286.

Frame Ex (2024). Individual growth at scale. Retrieved from: https://frameapp.io/.

Gengler, E.J.; Wedel, M.; Wudel, A.; and Laumer, S. (2023). Power Imbalances in Society and AI: On the Need to Expand the Feminist Approach. Wirtschaftsinformatik 2023 Proceedings. 37. Retrieved from: https://aisel.aisnet.org/wi2023/37

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, *22*(7), 1266-1286.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Microsoft. doi: 10.12987/9780300235029.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2). doi: https://doi.org/10.1177/2053951720943234

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

Grimmelmann, J. (2015). The virtues of moderation. *Yale JL & Tech.*, *17*, 42

Grygiel, J.; Brown, N. (2019). Are social media companies motivated to be good corporate citizens? Examination of the connection between corporate social responsibility and social media safety, Telecommunications Policy, 43 (5). doi:https://doi.org/10.1016/j.telpol.2018.12.003.

Häder, M. (2014). Delphi surveys: A workbook.

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-35.

Hankivsky, O. (2022). INTERSECTIONALITY 101.

Hartmann I.A. (2020), A new framework for online content moderation. *Computer Law & Security Review*, 26, doi: https://doi.org/10.1016/j.clsr.2019.105376.

Hellerstein, E. (2023). Silicon Savanna: The workers taking on Africa's digital sweatshops. Coda Media Inc. Retrieved from: https://www.codastory.com/authoritarian-tech/kenya-content-moderators/

Helma, L. (2015). Intersectionality as Method. DiGeSt. *Journal of Diversity and Gender Studies*, 2(1–2), 39–44. doi: https://doi.org/10.11116/jdivegendstud.2.1-2.0039

Hopkins, N. (2017) Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*. Retrieved from: https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

Jiang, J.A., Nie, P., Brubaker, J.R., & Fiesler, C. (2023). A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction*, 30 (1), 1–34. doi: https://doi.org/10.1145/3534929

Kirsch, G., & Kirsch, G. E. (1999). *Ethical dilemmas in feminist research: The politics of location, interpretation, and publication*. Suny press.

Köck-Hódi, S., & Mayer, H. (2013). Die Delphi-Methode: Übereinstimmung zu einem Thema durch die Befragung von Expertinnen. *ProCare*, *18*, 16-20.

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological bulletin*, *133*(1), 95.

Kshetri, N. (2021). Data labelling for the artificial intelligence industry: economic impacts in developing countries. *IT Professional*, *23*(2), 96-99.

Langvardt, K. (2017). Regulating online content moderation. Geo. LJ, 106, 1353.

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The delphi method* (pp. 3-12). Reading, MA: Addison-Wesley.

Meta. (2021, December). Meta's new AI system tackles harmful content [Press release]. Retrieved from https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/

Michael Benedikt (Ed.). 1991. Cyberspace: first steps. MIT Press, Cambridge, USA.

Mukherjee N, Zabala A, Huge J, Nyumba O-T, Esmail A-B, Sutherland W (2018) Comparison of techniques for eliciting views and judgements in decision-making. Methods in Ecology and Evolution 9, S. 54–63.

Nurik, C. (2019). “Men are scum”: Self-regulation, hate speech, and gender-based censorship on Facebook. *International Journal of Communication*, *13*, 21.

OpenAI. (2023, August 15). Using GPT-4 for content moderation [Blog post]. Retrieved from https://openai.com/blog/using-gpt-4-for-content-moderation

Quijano, A. (2007). Coloniality and modernity/rationality. *Cultural Studie*s, 21(2–3), 168–178. doi: https://doi.org/10.1080/09502380601164353

Roberts, S.T. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publication*s. Retrieved From: https://ir.lib.uwo.ca/commpub/12

Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-33.

Seering, J. (2020). Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the Human.-Computer Interacion.* 4, Article 107 doi: https://doi.org/10.1145/3415178.

Singhal, M., Ling, C., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2022). SoK: content moderation in social media, from guidelines to enforcement, and research to practice. arXiv. *Preprint posted online on June*, *29*, 2022.

Steiner, H. J., Alston, P., & Goodman, R. (2008). *International human rights in context: law, politics, morals: text and materials*. Oxford University Press, USA.

Sundermeier, J., Birkner, S., Ettl, K., Kensbock, J., & Tegtmeier, S. (2020). Hello Diversity! Opportunities and Challenges of Entrepreneurial Diversity in the Digital Age. *Communications of the Association for Information Systems*, 47, doi: https://doi.org/10.17705/1CAIS.04732

U.S. Const. amend. I

Wang, D., Prabhat, S., & Sambasivan, N. (2022). Whose AI Dream? In search of the aspiration in data annotation. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Winkler, J., & Moser, R. (2016). Biases in future-oriented Delphi studies: A cognitive perspective. *Technological forecasting and social change*, *105*, 63-76.

Witt, A., Suzor, N.P., & Huggins, A. (2019). The Rule of Law on Instagram: An Evaluation of the Moderation of Images Depicting Women's Bodies. *University of New South Wales Law Journal*

Wolfangel (2023). ChatGPT. Ausgebeutet, um KI zu zähmen. *Zeit Online.* Retrieved from: https://www.zeit.de/digital/2023-01/chatgpt-ki-training-arbeitsbedingungen-kenia

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104-120.