weizenbaum
JOURNAL OF THE
DIGITAL SOCIETY

VOL.06
ISSUE
03

Volume 6 \ Issue 3 \ 09-28-2026
DOI 10.34669/WI.WJDS/6.3.4 \ ISSN 2748-5625
Licensed under Creative Commons Attribution 4.0 (CC BY-SA 4.0)

RESEARCH PAPER

Factors That Matter

An Integrative Review of Human-Machine Decision-Making Processes in Content Moderation

Daniel Pothmann[*,1] \\ **Philipp Mahlow**[1] \\ **Maurice Stenzel**[1]
These authors contributed equally to this work.

[1]Alexander von Humboldt Institute for Internet and Society, Berlin, Germany
[*]**Corresponding author:** daniel.pothmann@hiig.de

ABSTRACT

Moderating online content is essential to ensure that digital communication is safe and enjoyable. Although automation is necessary to manage the substantial volume of content on large digital platforms, human expertise is deemed indispensable in decision-making processes, as well as in the design, training, and implementation of automated solutions. The "human-in-the-loop" (Hilo) concept is a key approach to integrating humans into (semi-) automated decision-making systems. However, its definition and implementation vary considerably across academic disciplines. Furthermore, the factors influencing the quality of such hybrid decision-making processes are not sufficiently well understood. This integrative literature review builds on previous conceptual work and identifies 14 factors that influence meaningful human involvement in (semi-)automated decision-making in content moderation. The findings provide an empirically grounded basis for evaluating existing and planned Hilo implementations. They will also support the design of socio-technical systems in which humans improve automated decision-making and will inform discussions on effective Hilo requirements. To this end, the review offers actionable insights to help platforms and policymakers enhance human–machine collaboration and decision-making quality.

KEYWORDS

content moderation \\ human-in-the-loop \\ content governance \\ decision-making \\ automation

1 Introduction

Moderating the content published on online platforms has become a necessity to ensure the safe and enjoyable usage of these services. Although harmful content existed long before the internet took over the majority of social communication, the "ease, speed and anonymity" of online content sharing has boosted its distribution (Langvardt, 2017). Technological support for human moderators has thus become necessary, and the sheer quantity of content posted on large social media platforms has made computer-driven assistance necessary from the very beginning (Bruckman et al., 1994; Dibbell, 1994).

Since the rise of social media in the 2010s, the scale of content has constantly increased, alongside the demand for more reliable software solutions facilitating the handling of millions of posts per day. Although recent content moderation approaches have relied heavily on automating the moderation process, human expertise remains crucial. Humans are currently involved by making (a fraction of the) final decisions, as well as designing, training, and implementing automated solutions (Ruckenstein & Turunen, 2019).

Regulators require that "a human" be integrated into decision loops, as stated, for instance, in Art. 22 of the General Data Protection Regulation (GDPR), Art. 14 of the Artificial Intelligence Act (AIA), or Art. 20, para. 6 of the Digital Services Act (DSA). Further, Art. 16, para. 6 of the DSA requires platforms to process user notices regarding content that is potentially illegal or violates platforms' terms of service "in a timely, diligent, non-arbitrary and objective manner." To make a diligent decision, human involvement is deemed necessary.

In this regard, the "human-in-the-loop" (Hilo[1]) concept has become increasingly important to ensure human control in automated systems, especially in automated decision-making (ADM) and artificial intelligence (AI) systems (Crootof et al., 2023). However, there is no widely accepted, cross-disciplinary definition of this concept. In computer science, the term usually refers to human involvement in the development and training of AI systems, with the aim of improving their accuracy (Chen et al., 2023). In contrast, regulatory perspectives emphasize human intervention during the operation of deployed decision-making systems, although full-cycle involvement is often considered impractical (Binns, 2022; Enarsson et al., 2021). However, neither approach is comprehensive because they both fail to capture human impact throughout socio-technical systems. Building on previous conceptual work, a proposed definition of the term (see Stenzel et al., forthcoming), and findings from ongoing research,[2] the Hilo concept is understood to encompass the roles, relationships, and procedures involved in a (semi-)automated decision-making process. The goal of human involvement is to integrate human judgment and critical review effectively throughout the entire system life-cycle,

[1] Unlike the established abbreviation "HITL" used in other disciplines, the abbreviation "Hilo" is used here to avoid confusion because of the conceptual differences with other established terms.

[2] The study design includes field analyses, workshops, and dialogue formats in four scenarios to address the research question comparatively in different organizational and regulatory contexts. The four case studies deal with automation in lending, content moderation, aviation, and prosecution.

meaningfully intervening where necessary to improve decision-making quality. In this regard, a decision could be considered high quality if it aligns with human and moral reasoning while remaining consistent with norms and the prevailing legal framework (Santonio de Sio & van den Hoven, 2018). Furthermore, the absence of arbitrariness, the reliance on rational argumentation, and the anchoring of decisions in "well-founded general ideas of justice of the community" (Federal Constitutional Court of Germany, 1973) are regarded as key elements of decision quality. A comprehensive operationalization of decision quality in content moderation processes is beyond the scope of this paper because the criteria for assessing decision quality must be tailored to the circumstances of each decision-making process and cannot be broadly generalized. However, proxies for the empirical assessment of decision quality may include consistency rates across comparable cases, rates of appeal and override in human-review stages, and compliance audit outcomes against applicable legal standards.

Although there is consensus that humans can make a meaningful difference in co-decisionary architectures, it remains unclear how these individuals can fulfill their roles and obligations effectively and adequately (Kaminski & Price, 2023; Sheridan, 1995). More broadly, an empirical understanding of the real-world conditions and factors that influence decision-making processes in such hybrid systems is still lacking. Specifically in the field of content moderation, there is no comprehensive overview of the factors that contribute to meaningful human involvement. Therefore, this integrative literature review aims to identify and synthesize factors that influence the quality of human–machine decision-making in the context of content moderation. These factors can be used to evaluate the meaningful impact of existing or planned Hilo implementations on the decision-making process, as well as in the regulatory discourse to determine qualitative Hilo requirements.

2 Method for the Integrative Literature Review

2.1 Study Design

To identify the factors influencing the decision quality of content-moderation processes, we adapted integrative literature review approaches to our research objectives to aggregate, critically assess, and synthesize influencing factors in the field of interest (Snyder, 2019; Torraco, 2005). Based on a defined search string (see Table 1) and theoretically informed selection criteria, we followed an iterative process of extracting aspects that influence decision quality, analyzing their impact, clustering them into factors on a comparable level of abstraction where possible, and summarizing each factor's effect within the socio-technical system in focus.

2.2 Inclusion Criteria and Search Strategy

To minimize conflicting information and develop a sufficient understanding of the practical process at an adequate level of abstraction, the scope of the review was limited to specific contexts in the realm of content moderation. We decided to focus on content moderation on major platforms, within the meaning of Art. 3 (i), 33 para. 1 of the DSA, without any geographical or cultural limitations on our literature selection. To account for the limitations of singular search tools and their specific algorithms, we used the following search tools to aggregate relevant literature: Google Scholar, Primo Central (a library licensed by the University of Liverpool), and StabiKat (an alternative library discovery service used by public libraries in Berlin, Germany). The selection of databases was primarily due to the transdisciplinary, socio-technical lens of the research question. Disciplinary databases could have returned more results, but they might also have introduced disciplinary biases. For each of these search tools, the formal inclusion criteria detailed in Table 1 were applied.

Table 1: Search criteria

Category	**Selection criterion**	**Qualification**
Content	Moderation of (written, audio-visual) content on (very large) online platforms	The full search string was constructed as follows: ((“content moderation” OR “content moderator*” OR “platform moderation” OR “platform moderator*” OR “online moderation” OR “online moderator*”) AND (“human-in-the-loop” OR “human in the loop” OR HITL OR HCI OR “human machine interaction” OR “human-machine-interaction” OR “human machine collaboration” OR “human-machine-collaboration” OR “human computer interaction” OR “human-computer-interaction” OR “human AI interaction” OR “human-AI-interaction” OR “human AI collaboration” OR “human-AI-collaboration” OR “decision making” OR “decision-making” OR judgment OR AI OR “artificial intelligence” OR “LLM” OR “large language model*” OR “human* impact” OR “human cost” OR “human oversight” OR “human supervision” OR pitfall* OR limitation* OR problem*))

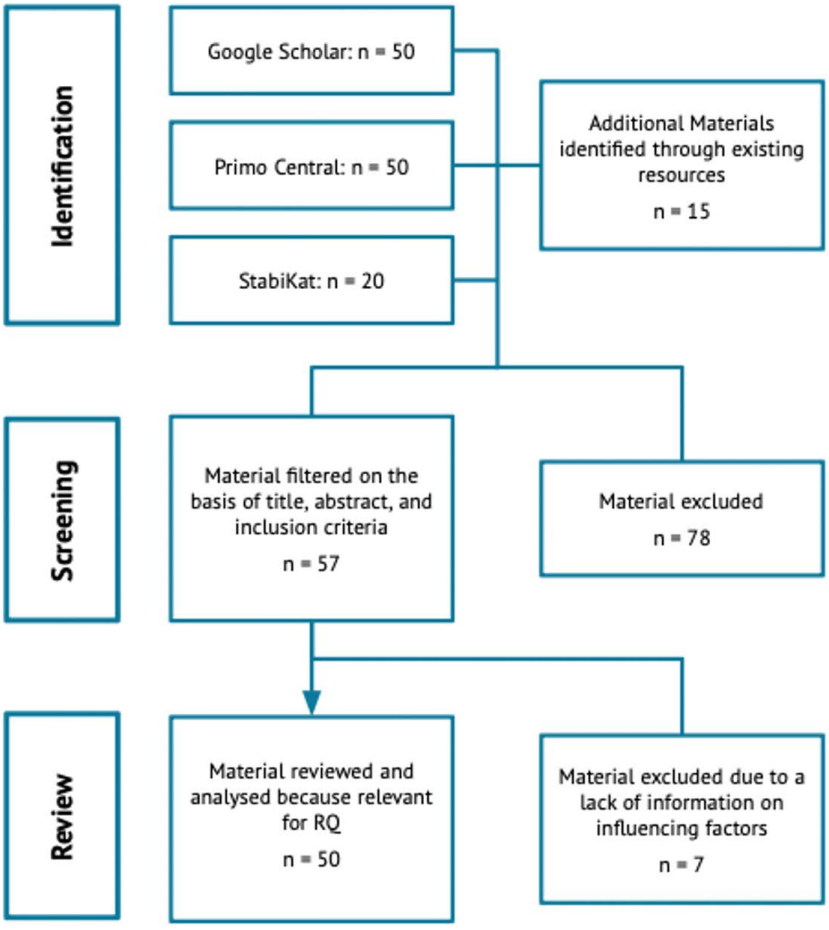
Category	Selection criterion	Qualification
Discipline	Not restricted explicitly	There were only implicit limitations due to the keywords included in the search string.
Publication year	2017–present	Necessity of recent literature due to substantial changes in the past 8+ years, particularly after US-election interference allegations in 2016 (François & Douek, 2021; Kumar, 2019). Additionally, the introduction of transformer architectures in 2017 advanced technological capabilities in content moderation (Kapoor, 2025).
Methods	Not restricted explicitly	Both qualitative and quantitative studies were considered.
Publication format	Not restricted explicitly	All written publications returned by the search tools were considered.
Language	German, English	Limited by the authors' language abilities.
Country/region	Worldwide	Influencing factors are not inherently region-specific and considering moderation practices in a variety of countries is crucial in this heavily globalized economy.
Sorting logic	By relevance, as defined by the search tool catalogues	Although this criterion depends on opaque sorting algorithms unique to the specific search tool, we deemed the general logic of relevance-sorting necessary to not miss important contributions and sacrifice either recent or impactful insights.

We used a structured search string for each search tool, which combined two thematic categories: one addressing the theme of content moderation (including variations), and the other covering the themes of human–computer interaction and decision-making (including variations). The breadth of this string was necessary due to the exploratory nature of the review. A preliminary screening prevented the inclusion of unfitting material.

2.3 Selection Process

Applying the search criteria listed above, we independently scanned through the first 50 listings (20 in the case of StabiKat) and selected publications based on their titles and abstracts, reviewing only those that appeared to fit our inclusion criteria. The scope of 50 (or 20) listings for more elaborate screening was determined to be sufficient based on a preliminary scan for criteria fit, which resulted in low confidence in relevant entries after said amount. Additionally, we analyzed an existing database of relevant literature that we had aggregated in the context of the overall research project. We then scanned the selection of 57 materials and excluded seven from our review due to a lack of information on influencing factors, leaving 50 publications for in-depth review and analysis. Figure 1 visualizes the process of identifying, screening, and reviewing materials.

Figure 1: Representation of our methodological approach



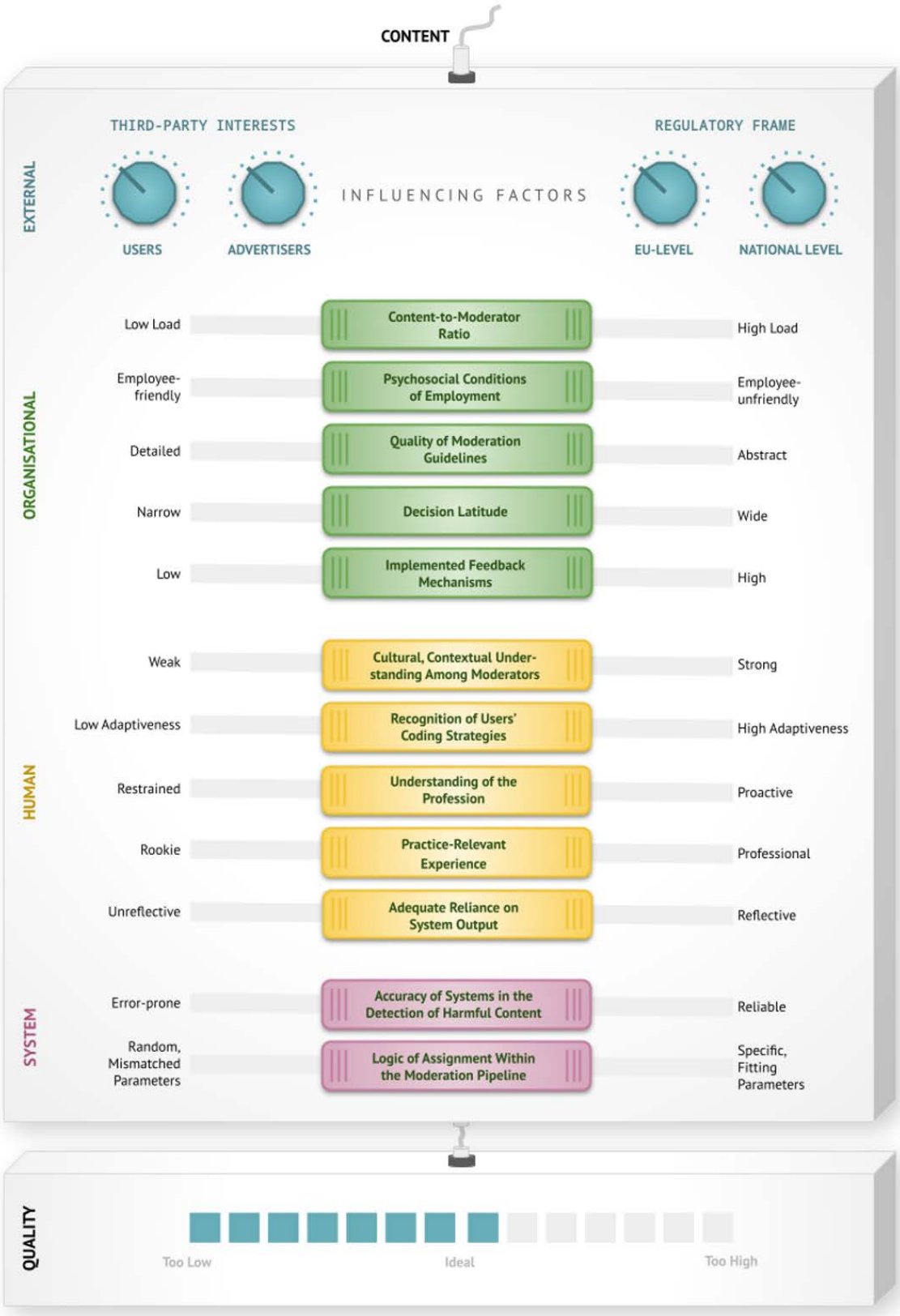
2.4 Data Extraction and Analysis

The synthesis of influencing factors from the extracted data followed an iterative review process involving multiple researchers before any factor was determined relevant and defined. The initial review was performed independently by multiple researchers, who extracted from the selected publications descriptions of aspects that either explicitly or implicitly influence the decision-making process. Saturation was evaluated and jointly considered sufficient because literature towards the tail section of the material reinforced previously surfaced aspects instead of introducing substantially novel elements. This also strengthened our confidence in the initial scoping of screened listings. These structured notes were then categorized, resulting in preliminary factors. For each entry, this step was done by a researcher who did not conduct the initial review of that particular material. After categorization, the structured notes were iteratively discussed in multiple sessions by all involved researchers, who have diverse academic backgrounds. Disagreements on factor definitions and categorizations were resolved qualitatively based on the coherence of the overall factor structure, the strength of support from the literature, and the logic of a factor's dynamics. This rigorous revision process produced a preliminary set of cohesive influencing factors, followed by a second step of aligning all constructed factors on a comparable level of abstraction.

3 Factors Influencing Decision-Making Quality in Content Moderation

The integrative literature review produced 14 influencing factors. Each factor describes functional dynamics and mechanisms on the external, human, organizational, or system level that affects decision quality, although the categorizations often are blurry. Importantly, the factors are not static or unchangeable but are characterized by a range of possible expressions and varying degrees of adjustability. In practice, there are set expressions of the factors within a socio-technical system that culminate in a certain decision quality. The mixing console in Figure 2 serves as a metaphorical illustration of their core logic of adjustment and interplay, which results in overall decision quality. Importantly, this is an oversimplification of the mechanisms at play, particularly in regard to binary dimensions, completeness, or causal implications. Additionally, the centering of the controls does not suggest the optimal setting but is meant to improve readability. Overall, the figure illustrates the interwovenness of the influencing factors discussed in more detail below.

Figure 2: An overview and metaphorical illustration of the influencing factors



3.1 Content-to-Moderator Ratio

The content-to-moderator ratio indicates the volume of content a moderator must process within a specified timeframe, thus reflecting each moderator's quantitative workload. An excessively high content-to-moderator ratio means that moderators may experience mental and physical exhaustion [21, 34, 19, 37, 15, 12, 3, 50],[3] while users may be exposed to more guideline-violating or illegal content, causing dissatisfaction [21]. Although hiring more moderators can lower this rate, it inevitably increases platform costs [25, 19, 15].

Because large platforms generate more content than humans alone can moderate, automated systems are increasingly used to lower the content-to-moderator ratio. Therefore, platforms increasingly depend on advancing AI systems. Critics note that AI implementation requires significant resources [7] and, being prone to errors in contextual understanding, still necessitates human oversight to ensure safe, high-quality moderation [18, 27]. There is also concern that human understanding of speech offenses and communication norms may eventually adapt to AI interpretations [4].

Some argue that, given the psychological toll of moderation, increased automation may be the only morally acceptable solution [7, 36]. Yet, as automation retains limitations, a sustainable, employee-friendly content-to-moderator ratio requires a balanced division of labor between humans and machines. To ensure high-quality decision-making within a platform's resources, it is crucial to determine which content can be delegated to machines and which requires humans to meet regulations and platform standards. When the content-to-moderator-ratio is tenable through the tailored delegation of work to machines, moderators are more willing to work, particularly in voluntary content moderation (VCM), and view their work as healthier and more efficient [12].

3.2 Psychosocial Conditions of Employment

Structural factors and working conditions affect the qualitative workload of human moderators, with physical, emotional, relational, and financial consequences [36]. These can be divided into structural conditions and psychological strain, both of which affect the quality of moderation and decision-making.

Structurally, moderators often find themselves in precarious situations involving low pay [34, 15,2], performance-based salaries and unpaid leave [34], short-term contracts [34], a lack of overall job security [2], and limited remote working options, even during the COVID-19 pandemic [19].

[3] The numbers in square brackets refer to the index of papers analyzed as part of the literature review (see Annex). Additional references in parentheses in this section are included to provide clarification or context for the individual factors and to support understanding, but they do not contribute directly to the core analytical framework of the review.

Psychologically, access to support varies. While some technical systems, such as the automatic pixelation of disturbing content, may reduce mental strain [47, 3], in-house psychological support is often unavailable [34]. Furthermore, health plans offered by the platforms rarely cover external therapy (Equidem, 2025). A lack of support can lead to mental exhaustion [15, 33, 2, 36, 13, 3, 42, 50], manifesting as symptoms such as insomnia, substance dependence, weight loss, anxiety, decision fatigue, depression, panic attacks, and suicidal thoughts [3, 42], as well as post-traumatic stress disorder and empathy fatigue. This can affect both personal life and work [3].

It has been suggested that the availability of support from the moderation team may improve the quality of moderation [35, 14]. However, opportunities for interpersonal exchange among moderators vary depending on the platform and moderation service provider. Some work alone, with limited contact with colleagues or supervisors [34], while others work in open-plan offices and language-based teams [19]. Nevertheless, confidentiality agreements can foster a "culture of silence," restricting discussions about work-related stress and hindering the relief of psychological strain [42].

3.3 Quality of Moderation Guidelines

High-quality moderation guidelines are essential to ensure the consistent and reliable enforcement of rules by moderation staff. They also give moderators confidence in their decisions [33]. Moderation guidelines encompass a range of materials, from general platform rules to detailed manuals with case-specific instructions. For instance, a broad rule such as "Spam, fraud, and other deceptive practices that mislead the platform community are not permitted on this platform"[4] can be supplemented by specific examples or criteria to guide interpretation.

The level of detail is crucial to the quality of moderation guidelines [21, 13]. A balance between abstraction and specificity is needed because abstraction allows adaptability to new cases, while specificity enables clearer classification. However, overly abstract guidelines can make moderation tedious [21], lead to inconsistent rule enforcement [30], or even encourage moderators to bend the rules for personal gain, particularly in VCM [21].

Beyond abstraction, other factors influence guideline quality. Their scope should align with the specific features of the platform and subsequent moderation requirements [33] while also reflecting diverse cultural contexts [41, 13]. However, there can be no "one-size-fits-all" solution for context-dependent cases. Involving affected groups can help address evolving forms of harmful speech [41]. Although consistency is important, moderation guidelines must remain flexible enough to address new content phenomena [7]. Also, the process, frequency, and the individuals entrusted with revising moderation guidelines affect their overall quality. This is particularly important in VCM, as inexperienced volunteer moderators create low-quality rules that ultimately lead to over-moderation [16].

[4] This example is based on the wording YouTube uses in its moderation guidelines regarding spam, deceptive practices, and scam policies (YouTube, 2025).

3.4 Decision Latitude

For moderators to fulfill a meaningful role, they need some amount of leeway in their decision-making. In content moderation, this latitude is two-fold, as it concerns both interpreting flagged content and choosing suitable measures (e. g., deletion or shadow-banning). Without such leeway, moderators would merely be rubberstamping automated decisions.

Platforms facilitate unique communities characterized by their members' values. Consequently, if membership changes, values shift as well [21]. Moderation rules cannot be made specific enough for completely uniform application without becoming outdated quickly. Only through human-led discovering, testing, and reassertion of values can adequate rule-application be achieved. For example, forbidding hate speech might be a universally applicable rule, but determining whether something constitutes hate speech requires carefully reconciling the fundamental right to freedom of speech with the right not to be harmed [17, 7].

This leeway places significant responsibility on moderators, raising questions of qualification, bias, and required cultural literacy. Interpretive freedom definitionally entails subjective preferences [28, 49]. For example, humans are prone to sympathize with users to whom they can relate more, which may cloud their judgment [35]. Further, having little time per decision increases doubts regarding the actual use of decision latitude, with moderators reporting having to decide on 700–1,000 cases per shift (Equidem, 2025). Still, some degree of latitude is unequivocally beneficial to decision quality given that the meanings of language and symbols constantly evolve.

3.5 Implemented Feedback Mechanisms

The literature indicates that the existence and design of adequate feedback mechanisms can influence the quality of the content-moderation process. In terms of design, it is important to consider which employees are involved in the feedback process. If feedback is determined purely automatically, it must first be clarified whether the quality of the moderation guidelines can be assessed automatically. One source highlights that Reddit's "Automod" help system does not provide information about how frequently a specific system rule leads to moderation. Consequently, moderators can only evaluate the effectiveness of a rule based on user feedback [16]. Conversely, if human moderators are given the opportunity to provide feedback, studies show that they can create their own moderation rules more quickly and accurately than AI systems can learn through reinforcement learning [22].

3.6 Cultural and Contextual Understanding Among Human Moderators

Human moderators play a crucial role in assessing content that automated systems cannot, particularly for context-sensitive posts. The quality of moderation depends on the moderators' contextual understanding and cultural literacy, as well as the time they are given to understand the context and their access to tools for researching unfamiliar contexts.

Crucially, the amount of contextual information provided must be balanced with the time allotted per decision. Limited time per decision increases moderators' stress [34, 6], especially considering that a minimum amount of time is needed to interpret the context adequately. Consequently, insufficient or excessive context information without sufficient time harms decision quality. The time needed to assess content also depends on the moderators' cultural literacy, which encompasses an understanding of symbols, values, narratives, historical events, and social codes, as well as linguistic and semiotic knowledge [24, 35, 44, 34, 40, 29, 33, 28, 17, 14]. However, language skills alone do not guarantee high-quality moderation due to dialects, slang, generational differences, and humor [35, 27]. Partial contextual or cultural literacy can lead to the misclassification of satire, covertly illegal content, or practices like reclaiming. Conversely, users' coding strategies, such as dog-whistling, may remain undetected. Using linguistic markers to identify constructive contributions can facilitate comprehension of the context, but this approach has been criticized for being too focused on Western values [40].

3.7 Recognition of Users' Coding Strategies

Knowledge about filter systems' weaknesses empowers users to exploit them. If users manage to prevent their harmful content from being detected by these systems, it will not reach the critical eye of a moderator. Using codified language to express restricted speech is often simple and unoriginal. Yet what any human could easily decipher is often still difficult for systems to catch. Modern classifiers continue to struggle with combinations of words, misspellings, satire, changing syntax, and coded language [44, 20]. Oftentimes, the "code" might not even be strategic, for example, in the case of region-specific racial slurs that have harmless meanings in other contexts [14]. In other instances, there is obvious intent to circumvent filters, such as referring to Hitler as "the painter," which is understandable to humans but undetectable to systems lacking contextual inference. Users' strategies for circumventing filters while moderators try to catch up lead to a cat-and-mouse game. This is where timely and reliable human–machine collaboration becomes indispensable. Thus, being able to quickly adapt systems through moderators providing feedback on newly discovered coding strategies is key to building resilient moderation systems at scale. However, coding strategies are a powerful excuse for platforms to not publish their exact moderation criteria [16, 34], even though increased transparency could benefit the perceived fairness and acceptance of moderation practices [35, 13].

3.8 Understanding of the Profession

Content moderators' understanding of their profession depends on their perception of their position and tasks within the socio-technical system, which shapes their moderation decisions within the decision latitude afforded to them. The degree of invasiveness chosen for the specific moderation process, that is, how "hierarchical" moderators can act towards platform users, is a decisive factor [47, 26]. "Dogmatic" moderation involves no explanation of the moderation, "authoritative-interactive" moderation informs users in advance of moderation measures and the standards of enforcement, and "discursive-interactive" moderation allows user statements [47]. The final subgroup is visibility moderation, where content is not removed from the platform, but its visibility is reduced for certain user groups through methods such as geo-blocking, shadow banning, or age restrictions [47].

Moderators' views of their activities' objectives are also relevant. They may perceive their role as to facilitate civil and productive discourse or to ensure the quality of content on the platform [37] (e.g., the moderation of the "Discussion" tab of a Wikipedia article) [37]. Surveyed moderators consider it crucial for the information in online encyclopedia entries to be accurate and, therefore, lean towards accepting disruptions to the discussion resulting from their moderation [37]. Voluntary content moderators describe roles ranging from that of a "police/manager" to a "team member/discussion facilitator" and favor less confrontational, direct engagement [37].

In terms of how content moderators understand their role, it is important to consider whether they view the use of systems as supportive or as a replacement justified by economic reasons [35]. The latter is perceived by content moderators as criticism or devaluation of their work, which can result in lower acceptance of introduced systems [35].

3.9 Practice-Relevant Experience

As with any task, practicing content moderation might not result in perfection but in a substantial improvement in decision quality. Employees who have been moderating content for longer have a better intuition for the task [35]. Especially in difficult cases, experience with similarly complex moderation decisions affects a moderator's ability to effectively solve them [33]. In addition to proficiency in core moderation tasks, experience increases abstract awareness of current problems within practice, including the holistic moderation process [35].

Besides increased proficiency, performing a repetitive task like content moderation over an extended period involves a risky trade-off. On the one hand, developing a routine might increase workflow efficiency (e.g., increased speed in the operation of software). On the other hand, repetition can decrease attention to detail (rubberstamping) since humans inherently seek to reapply known patterns to preserve precious cognitive resources.

3.10 Adequate Reliance on System Output

Human moderators are prone to excessively trusting the outputs of machine systems, even when these outputs are flawed or clearly incorrect. This phenomenon is often referred to as "automation bias" (Goddard et al., 2014). In the context of content moderation, human moderators tend to rely on systems the most in areas where those systems perform the worst, for example, when dealing with difficult, context-dependent cases [22]. System-generated attributions of meaning are simply adopted; for instance, keywords flagged by systems are "automatically" perceived as problematic by human moderators [22]. Furthermore, developers of moderation systems and processes are significantly more skeptical about them than the moderators who use them [35].

To develop an appropriate level of trust in technical systems, human moderators must understand the systems and their limitations sufficiently well. To this end, technical experts must 'translate' the relevant contexts of AI systems into understandable explanations for moderators [17], which has been called the "interpretive function" of a Hilo (Crootof et al., 2023). In this regard, it is helpful to educate moderators about what the system does and why, as well as its limitations. Another remedy currently under discussion is the use of large language models (e.g., LLM-as-a-judge [32]) because these are believed to be capable of generating textual explanations of their outputs.

3.11 Accuracy of Systems in the Detection of Harmful Content

A key prerequisite for moderation is the detection of suspected content by filtering systems. The accuracy of these filters determines false positive and negative rates, overall reliability, and the confidence conveyed to moderators and users. Accurate, transparent evaluations help moderators make better decisions as a Hilo. These filters typically include predictive classifiers or hash-based matching systems.

The literature on the fallibility of moderation systems primarily focuses on predictive classifiers because hash-matching systems, especially "perceptual hashing," are relatively tolerant of deviations (Gorwa et al., 2020). The main cause of error in classifiers lies in the limited availability of high-quality training data [35, 44]. This may be due to biased data collection (e.g., if restricted to certain geographical, linguistic, or cultural contexts) or technical constraints that prevent full contextual representation [24, 47, 4, 18, 19, 30, 17, 7, 49, 14, 27, 13]. Consequently, classifiers perform well only within familiar contexts and substantially worse than humans outside them [22], which leads to biased outcomes for underrepresented data [25, 44].

Linguistic nuances, such as mock politeness, euphemisms, and hidden or semantically veiled speech, also present challenges for detection [14, 31, 4]. Additionally, some systems assess toxicity relative to post length, with sensitivity varying depending on the topic or group [35, 11]. However, models that rely too heavily on keywords (e. g., swear words) and other specific

markers (e. g., Black-aligned English) may over-block content [31, 25, 44]. Keyword-based systems (e. g., Perspective API) mainly serve an assisting function [8, 37, 33]; to make informed decisions, moderators need transparent tools that clarify contextual harm [11]. Furthermore, as humans train the systems intended to correct their biases [35], disagreement among human annotators can lead to false learnings [15].

Additionally, errors can be trivial, such as mistaking spilled red paint for blood. This could be avoided by considering the context (e. g., a nearby paint bucket). However, contextual reliance can be exploited, as in the case of filters misinterpreting topless photos with baby dolls as breastfeeding. Accurate predictions require up-to-date training data containing sufficient contextual information [24, 17, 27]. Importantly, the findings outlined above focus on errors in the training data and do not cover the inherent statistical limitations of classifiers or design choices in the system, such as the setting of precision thresholds.

3.12 The Logic of Assignment Within the Moderation Pipeline

Assigning detected content to human moderators or machines, taking into account expertise and experience or automated handling, affects the quality of moderation through both accuracy and "soft" effects on moderators. The "moderation pipeline," or the sequence of steps in a moderation process, begins after algorithmic detection. Depending on the format (text, image, or video), suspected rule violations (e.g., nudity, hate speech, or terrorism), language, or regional/cultural context, it may assign content to moderators or systems for moderation. However, the exact pipelines are opaque to the public, making this literature-based list non-exhaustive.

The literature identifies regular and high-priority queues. These are based on moderator seniority and expertise [19], with content being assigned directly or escalated from junior to senior moderators when needed. The assignment may also depend on contextual urgency, for instance, the frequency of similar flagged content or its potential impact [13]. Asynchronous processes, caused by filtering delays [35] or private-life obligations in VCM [21], can limit the ability to intervene promptly and affect both user experience and moderator workload [37].

Maladaptive assignment can have several effects. Random assignment can cause stress for moderators due to constant adaptation and exposure to graphic content [6], and it may reduce efficiency because of a lack of specialization [24]. Repetitive work may result in lower rigor [35], while relying solely on keywords can lead to false positives [37] and the transmission of bias (Vicente et al., 2025). The prioritization of frequently flagged content as an alert is vulnerable to abuse by coordinated groups or "troll armies" who flag content they want removed [34]. This strategy is also referred to as "coordinated flagging" (Griffin & Stallman, 2024). Moreover, users may become frustrated if moderation occurs hours after posting due to asynchronous filtering delays, assignment pipelines, or other obligations, while preemptive deletion is often perceived as censorship [35].

3.13 Third-Party Interests

Beyond the legal frameworks and platforms' approaches to moderation governance, moderation processes are shaped by third-party interests, particularly those of users and advertisers. Platforms' business models depend on user engagement to determine the value of ad placement for revenue generation (Meta Investor Relations, 2025). As a result, platforms align their content guidelines, the consequences of user engagement, and the interests of advertisers.

This reflects the attention economy, where content that drives engagement creates value for advertisers [13], with mechanisms focusing on retaining both users and advertisers. Erroneous moderation can lead to discontent [36], particularly when users have invested effort in creating content [35] or when they are public figures or popular content creators [37]. Minimizing the time from posting or flagging to an actual moderation decision reduces frustration [34]. However, speed may conflict with careful, diligent decision-making and transparent justifications, which are crucial to procedural fairness and perceived legitimacy [15, 46]. Without these, overall platform governance as a superordinate framework for the quality of content moderation is negatively affected [15].

For advertisers, content must align with brand values [36], and declining user engagement may deter ad investment. Therefore, platforms must strike a balance between showing engaging ads and not annoying users, while maintaining a "brand-friendly" environment that will retain financially strong advertising customers in the long term. This ambivalence is exacerbated by the positive association between the emotionality of content and user engagement (Schreiner et al., 2019).

3.14 Regulatory Frame

The regulatory framework affects the way platforms operating in Europe moderate their content in three ways. First, general legal frameworks govern the structure and operation of platforms. These include contractual and corporate legal requirements and rules on working conditions and accounting.

Second, requirements regarding the formal design of content-moderation processes can be found in the DSA and the GDPR, in particular. Art. 16 of the DSA, for example, sets a requirement for platforms to implement a reporting system, and Art. 22(1) of the GDPR imposes a general prohibition on the use of fully automated decision-making systems (Penagos, 2025).

Third, in practical terms, legal requirements influence the moderation decisions that must be made. According to Art. 3(t) of the DSA, content that is unlawful or violates the platforms' terms of use must be moderated. Whether content is illegal is determined by the laws of the European Union's Member States.

Importantly, inadequate regulatory requirements can create false incentives. For instance, the effectiveness of moderation could be of secondary importance if the legal framework considers a sufficient number of moderated posts to be indicative of an effective content-moderation process [43]. Therefore, the incentives set by the legal framework and the extent to which compliance can be verified impact the quality of the process.

4 Discussion of the Results

In this non-exhaustive, integrative literature review, we searched multiple academic databases for publications on the collaboration of humans and machines in the field of content moderation. We aimed to identify the factors influencing the quality of content moderation decision-making processes embedded in their socio-technical systems.

In sum, 14 such factors were derived in an iterative process of synthesizing explicit and implicit information, categorizing gathered information into cohesive clusters, and aligning these on a matching level of abstraction. The factors address the following aspects: the content-to-moderator ratio, the psychosocial conditions of employment, the quality of moderation guidelines, decision latitude, implemented feedback mechanisms, cultural and contextual understanding among human moderators, the recognition of users' coding strategies, understanding of the profession, practice-relevant experience, adequate reliance on system output, the accuracy of systems in the detection of harmful content, the logic of assignment within the moderation pipeline, third-party interests, and the regulatory framework.

Our method only allowed for the identification of factors influencing decision-making quality. Influence here is understood as the factors' effects, as described in the material, being deemed relevant to the research question. The relationship between the factors and decision-making quality does not enable causal interpretations, an estimation of the strength of the relationship, or conclusions about how the factors interact. Such inferences will require empirical investigations designed to detect causal patterns. Further, there are noticeable differences in the levels of abstraction between a majority of specific factors and a minority of seemingly more abstract factors, namely, third-party interests and the regulatory framework. This is because the factors result from categorization, which was influenced by the contextual proximity of identified aspects, producing more concrete factors where the material provided more detail and more abstract factors where the material was less detailed. Excluding certain aspects, for example, third-party interests, because a differentiated discussion of economic incentives or business practices was missing, would not be coherent with the exploratory approach and socio-technical lens of the research question.

Quantitatively, the strongest support in the material was found for the content-to-moderator ratio, the psychosocial conditions of employment, and the accuracy of systems in the detection of harmful content. In practical terms, this points to three critical areas affecting

the content-moderation process. First, considering the roles assigned to human moderators and the immense responsibility that comes with them, it is crucial for employers to carefully assess the amount of human resources needed to diligently fulfill them. Although this is certainly limited by the availability of funds and qualified personnel, the ratio should be proportionate to the risks involved.

Second, given the severity of the psychological consequences of moderation work, appropriate support measures are imperative. Support should not be limited to post-hoc treatment via counseling or therapy but also include actions to reduce the induction of stress and trauma in the first place. Nonetheless, we acknowledge the difficulty of balancing extensive human review of harmful content with limiting the exposure to graphic material. However, this dilemma does not undermine the criticality of effectively tackling the issue.

Third, the review above addresses a multitude of flaws embedded in technological systems that assist moderators in detecting harmful content. Importantly, we only shine a dim light into the black box of shortcomings of predictive systems in particular. In the context of content moderation, the data being filtered is highly complex and context dependent and demands well-annotated, relevant training data that adequately represents the affected populations and are continuously updated based on moderators' feedback. At the same time, technological and statistical limitations must be considered in the distribution of tasks and responsibilities.

As Europe's leading regulatory tool for large online platforms, the DSA sets specific process requirements for sufficient notice and action mechanisms in content moderation. Yet it does not fully provide for specific rules for semi-automated decision-making. Art. 16, para. 6 of the DSA requires that providers of hosting services and platforms process notices in a "timely, diligent, non-arbitrary and objective manner." Unlike other European regulations, such as the AIA, the DSA does not contain specific requirements for a diligent process design. Art. 14, para. 1 of the AIA, for example, requires high-risk AI systems to be designed and developed to be effectively overseen by natural persons while in use. A checklist of AI-system features that enable effective oversight is provided in Art. 14, para. 4 of the AIA. This includes a "stop button" that enables overseers to halt the system in a safe state. According to Art. 26, para. 2 of the AIA, deployers must assign human oversight to natural persons with the necessary competence, training, authority, and support to fulfil oversight obligations.

Given automated content-moderation systems' current susceptibility to errors and the consensus that humans can make a meaningful difference in co-decisionary architectures, only a process design incorporating Hilos can be considered truly diligent within the meaning of Art. 16, para. 6 of the DSA. As the analyzed literature makes clear, key factors must be attended to sufficiently to optimize semi-automated content-moderation processes. At the time of writing, the authors of this paper are drafting a DSA reform proposal in light of the current DSA review stage, stressing the need to include a specific requirement focused on the process design. This could be achieved by adding the following addendum (bold, in brackets) to Art. 16, para. 6, sentence 2 of the DSA:

> Providers of hosting services shall process any notices that they receive under the mechanisms referred to in paragraph 1 and take their decisions in respect of the information to which the notices relate, in a timely, diligent, non-arbitrary and objective manner. Where they use automated means for that processing or decision-making, they shall include information on such use in the notification referred to in paragraph 5 **[and shall design the decision-making process in such a way as to ensure meaningful human involvement]** (Digital Services Act, 2022, Art. 16(6)).

To help norm addressees implement this process design, regulators, digital service providers, civil society, and academia could collaborate on co-regulatory measures like codes of conduct that enable responsible content-moderation practices beyond abstract regulatory requirements. As an example of a soft-law measure, the authors iteratively developed a Code of Conduct on Human-Machine Decision-Making in Content Moderation (Kettemann et al., 2025) in collaboration with various stakeholders to further specify the regulation of content moderation, focusing on effective and responsible cooperation between humans and machines. This crucial yet often overlooked relationship is addressed in 10 codes containing practical measures that serve as suggestions for action to promote responsible, transparent, and fundamental rights-preserving content moderation. The identified factors provided the scientific backing for the drafting process.

Overall, the reviewed literature sample had a clear focus on aspects closer to the concrete moderation process performed by moderators (e.g., effects on moderators or flaws of the technical system) and less so on aspects more distant from the concrete process (e.g., broader content governance by platforms and external actors, organizational decision-making processes, or business model incentives). This lack of holistic analyses could be attributed to the fact that most material deals with particular issues in content moderation rather than abstract processes and the grounds of their configurations. As concerns the predominant disciplines, most material was published in the fields of social sciences, law, and political science, while economics or psychology, for example, were substantially less prevalent. However, the pool of considered materials was overtly affected by the search terms, which concentrated on topics more akin to social sciences than other fields, such as computer science. Consequently, the methodology was dominated by qualitative approaches, limiting conclusions about the scale of the identified dynamics, as discussed above.

As concerns the issues we expected to be addressed based on our initial, surface-level modelling of the moderation processes (Züger et al., 2025), a few areas are worth mentioning that were not elaborated on in the material. Between the two broad categories of systems utilized for content moderation (i.e., predictive classifiers and hash-matching), there was an almost exclusive focus on predictive systems, which are arguably more prone to errors, including the widely discussed risks of harmfully biased training data. Further, discussions of system-related factors predominantly concerned the training data rather than aspects intrinsic to system functionality (e.g., limitations of statistical logic) or the intentional design of these systems (e.g., the setting of accuracy thresholds). Relatedly, the material contained almost no information on the prevalence and design of mechanisms for moderators to provide feedback on the systems' outputs and how this may be implemented.

With its holistic view of the socio-technical system in content moderation, the synthesis and comprehensive systematization of influencing factors provided in this review represents a novel approach to understanding collaborative human–machine decision-making processes. Prior studies have either examined the roles of humans in (semi-)automated decision-making, described the implications of practical processes (e.g., moderating content), or critically discussed the risks involved in specific cases. In contrast, this review is a first attempt to holistically map influencing factors derived from the relevant literature, building a foundation for determining vectors of improvement and formulating actionable steps towards higher-quality decision-making. Only by identifying critical interdependencies within and around the decision-making process can aspects detrimental to rational judgment and socially negotiated justice, which ultimately corrupt the overall quality of the process, be reflected on. Research investigating more isolated dimensions of these processes in-depth is a crucial prerequisite for the present approach, but sufficient complexity in the understanding of (semi-)automated decision-making can only be achieved through the evaluation of socio-technical systems. Otherwise, recommendations could overlook blind spots related to any applied measures.

5 Conclusion and Outlook

The 14 factors derived from this integrative literature review contribute to a holistic understanding of the factors influencing the quality of content-moderation decision-making processes. By laying bare the co-dependent dynamics and adjustable mechanisms that either directly or indirectly affect the goal of timely, diligent, non-arbitrary, and objective content moderation (as demanded by Art. 16 of the DSA), platforms and policymakers can work to better align practice with said goals. In particular, ensuring sufficient human resources, adequate psychological support, and adaptive technological systems with proper feedback mechanisms appears to be key to high decision-making quality.

In the context of the research effort in which this review is embedded, the findings also contribute to a prospective meta-analysis of influencing factors in four different case studies on (semi-)automated decision-making processes. Based on these case studies, similarities and abstract factors relevant to collaborative human–machine decision-making in general will be analyzed, culminating in a taxonomy applicable to a broader range of practical cases. More directly, this review also informed the drafting of a Code of Conduct on Human-Machine Decision-Making in Content Moderation (Kettemann et al., 2025), which formalizes 10 guiding principles for the organization of content-moderation processes by large online platforms. With this code, we propose a voluntary set of actionable measures platforms can commit to, as stipulated in Art. 45 of the DSA, to demonstrate legal compliance beyond minimum obligations. Lastly, the recognition of the multifaceted factors that influence the quality of the decision-making process led the authors to draft a DSA reform proposal to help the legislator better reflect the complexity of content moderation in the DSA.

References

Binns, R. (2022). Human judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance, 16*, 197–211. https://doi.org/10.1111/rego.12358

Bruckman, A., Curtis, P., Figallo, C., & Laurel, B. (1994). Approaches to managing deviant behaviour in virtual communities. *Conference Companion on Human Factors in Computing Systems (CHI '94), Association for Computing Machinery*, 183–184. https://doi.org/10.1145/259963.260231

Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "human-in-the-loop" system. *Systems, 11*(11), Article 548. https://doi.org/10.3390/systems11110548

Crootof, R., Kaminski, M. E., & Price II, W. N. (2023). Humans in the loop. *Vanderbilt Law Review, 76*(2), 429–510. http://dx.doi.org/10.2139/ssrn.4066781

Dibbell, J. (1994). A rape in cyberspace; or, how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. In M. Dery (Ed.), *Flame wars: The discourse of cyberculture* (1st ed., pp. 237–262). Duke University Press. https://doi.org/10.1215/9780822396765

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2021). 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

Equidem. (2025). *Scroll. Click. Suffer. The hidden human cost of content moderation and data labelling*. https://equidem.org/wp-content/uploads/2025/05/Equidem-Data-Workers-Report-May-29.pdf

Federal Constitutional Court of Germany. (1973, February 14). *Soraya* (BVerfGE 34, 269).

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics, 83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

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). https://doi.org/10.1177/2053951719897945

Griffin, R. & Stallman, E. (2024, February 22). *A systemic approach to implementing the DSA's human-in-the-loop requirement*. VerfBlog. https://verfassungsblog.de/a-systemic-approach-to-implementing-the-dsas-human-in-the-loop-requirement/

Kettemann, M. C., Mosene, K., Stenzel, M., Mahlow, P., Pothmann, D. & Spitz, S. (2025). *Strengthening trust. Code of conduct on human–machine decision-making in content moderation*. Alexander von Humboldt Institute for Internet and Society. https://doi.org/10.5281/zenodo.17650988

Langvardt, K. (2017). Regulating online content moderation. *Georgetown Law Journal, 106*(5), 1353–1388. http://dx.doi.org/10.2139/ssrn.3024739

Penagos, V. E. (2025). Platforms on the hook? EU and human rights requirements for human involvement in content moderation. *Cambridge Forum on AI: Law and Governance, 1*, Article e23. https://doi.org/10.1017/cfl.2025.3

Ruckenstein, M., & Turunen, L. L. M. (2019). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society, 22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990

Santoni de Sio, F. & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*(February), Article 15. https://doi.org/10.3389/frobt.2018.00015

Sheridan, T. B. (1995). Human centered automation: Oxymoron or common sense? *IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, 823–828. https://doi.org/10.1109/ICSMC.1995.537867

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stenzel, M., Kettemann, M. C., Mosene, K., Pothmann, D., Spitz, S. & Mahlow, P. (forthcoming). *(Teil-)Automatisierte Content Moderation: Macht, Recht und die Rolle menschlicher Entscheidungen. Abschlussbericht der zweiten Fallstudie des Projekts „Human in the Loop? Autonomie und Automation in sozio-technischen Systemen"*. HIIG Impact Publication Series

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review, 4*(3), 356–367. https://doi.org/10.1177/1534484305278283

YouTube (2025). *Spam, deceptive practices, & scams policies*. https://support.google.com/youtube/answer/2801973?hl=en

Züger, T., Mahlow, P., Pothmann, D., Mosene, K., Burmeister, F., Kettemann, M. C., & Schulz, W. (2025). Crediting humans: A systematic assessment of influencing factors for human-in-the-loop figurations in consumer credit lending decisions. *FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1281–1292. https://doi.org/10.1145/3715275.3732086

Date received: 20 February 2026
Date accepted: 19 August 2026

Annex

Indexed text corpus of the literature review

[1] Araujo, T., Brosius, A., Goldberg, A. C., & Moller, J. (2023). Humans vs. AI: The role of trust, political attitudes, and individual characteristics on perceptions about automated decision making across Europe. *International Journal of Communication, 17*, 6222–6249. https://ijoc.org/index.php/ijoc/article/view/20612/4348

[2] Barnes, M. R. (2022). Online extremism, AI, and (human) content moderation. *Feminist Philosophy Quarterly, 8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14295

[3] Cook, C. L., Cai, J., & Wohn, D. Y. (2022). Awe versus aww: The effectiveness of two kinds of positive emotional stimulation on stress reduction for online content moderators. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), Article 3555168, 1–19. https://doi.org/10.1145/3555168

[4] Dias Oliva, T., Antonialli, D.M. & Gomes, A. (2021). Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture, 25*, 700–732. https://doi.org/10.1007/s12119-020-09790-w

[5] Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2021). 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

[6] Gier-Reinartz, N. R., Zimmermann-Janssen, V. E. M., Kenning, P., Léger, P.-M., Randolph, A. B., Davis, F. D., Müller-Putz, G. R., Brocke, J. vom, & Riedl, R. (2024). AI-assisted hate speech moderation—how information on AI-based classification affects the human brain-in-the-loop. *Information Systems and Neuroscience, 68*, 45–56. https://doi.org/10.1007/978-3-031-58396-4_5

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

[8] 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). https://doi.org/10.1177/2053951719897945

[9] Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law & Technology, 42*.

[10] Hartmann, I. A. (2020). A new framework for online content moderation. *Computer Law & Security Review, 36*, Article 105376. https://doi.org/10.1016/j.clsr.2019.105376

[11] Hashir, M. H., Memoona, & Kim, S. W. (2025). TARGE: Large language model-powered explainable hate speech detection. PeerJ *Computer Science, 11*, Article e2911. https://doi.org/10.7717/peerj-cs.2911

[12] He, Q., Hong, Y. K., & Raghu, T. S. (2022). The effect of machine-powered content moderation: An empirical study on Reddit. *Proceedings of the 55th Hawaii International Conference on System Sciences, HICSS 2022, 5963–5972*.

[13] Heldt, A., & Dreyer, S. (2021). Competent third parties and content moderation on platforms: potentials of independent decision-making bodies from a governance structure perspective. *Journal of Information Policy, 11*, 266–300. https://doi.org/10.5325/jinfopoli.11.2021.0266

[14] Hettiachchi, D., & Goncalves, J. (2020). Towards effective crowd-powered online content moderation. *Proceedings of the 31st Australian Conference on Human-Computer-Interaction (OzCHI '19).* (pp. 342–346). Association for Computing Machinery. https://doi.org/10.1145/3369457.3369491

[15] Huang, T. (2024). Content moderation by LLM: From accuracy to legitimacy. *arXiv:2409.03219 [cs.CY]*. https://doi.org/10.48550/arxiv.2409.03219

[16] 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), 1–35. https://doi.org/10.1145/3338243

[17] Kapoor, M. (2025). The Evolving Role of Human-in-the-Loop Evaluations in Advanced AI Systems. *European Journal of Computer Science and Information Technology, 13*(9), 115–126. https://doi.org/10.37745/ejcsit.2013/vol13n9115126

[18] Katsaros, M., Kim, J., and Tyler, T. (2024). Online content moderation: Does justice need a human face? *International Journal of Human–Computer Interaction, 40*(1), 66–77. https://doi.org/10.1080/10447318.2023.2210879

[19] Katzenbach, C., Pentzold, C., & Viejo Otero, P. (2024). Smoothing out smart tech's rough edges: Imperfect automation and the human fix. *Human-Machine Communication, 7*, 23–43. https://doi.org/10.30658/hmc.7.2

[20] Kim, S., Oh, C., Cho, W. I., Shin, D., Suh, B., & Lee, J. (2021). Trkic G00gle: Why and how users game translation algorithms. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–24. https://doi.org/10.1145/3476085

[21] Kolla, M., Salunkhe, S., Chandrasekharan, E., Saha, K., Sas, C., Mueller, F. F., Williamson, J. R., & Kyburz, P. (2024). LLM-Mod: Can large language models assist content moderation? *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3613905.3650828

[22] 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*. https://doi.org/10.48550/arxiv.2204.11788

[23] Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society, 7*(1). https://doi.org/10.1177/2053951720920686

[24] Lykouris, T., & Weng, W. (2025). *Learning to defer in congested systems: The AI-human interplay*. https://doi.org/10.48550/arXiv.2402.12237

[25] Martínez, M. B. (2023). Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union. *Journal of Intellectual Property, Information Technology, and Electronic Commerce Law, 14*(1), Article 211.

[26] Meerson, R., Koban, K., & Matthes, J. (2025). Platform-led content moderation through the bystander lens: A systematic scoping review. *Information, Communication & Society*, 1–18. https://doi.org/10.1080/1369118X.2025.2483836

[27] Mehta, M., & Giunchiglia, F. (2025). *Understanding Gen Alpha digital language: Evaluation of LLM safety systems for content moderation*. arXiv:2505.10588 [cs.CY]. https://doi.org/10.48550/arxiv.2505.10588

[28] Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication, 27*(4), Article zmac010. https://doi.org/10.1093/jcmc/zmac010

[29] Molina, M. D., & Sundar, S. S. (2024). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society, 26*(6), 3638–3656. https://doi.org/10.1177/14614448221103534

[30] Nguyen, A., Rai, A., & Maruping, L. (2024). Understanding the unintended effects of human–machine moderation in addressing harassment within online communities. *Journal of Management Information Systems, 41*(2), 341–366. https://doi.org/10.1080/07421222.2024.2340831

[31] Oh, D., & Downey, J. (2024). Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI. *Information, Communication & Society, 28*(7), 1157–1176. https://doi.org/10.1080/1369118X.2024.2346531

[32] Pasch, S. (2025). *AI vs. human judgment of content moderation: LLM-as-a-judge and ethics-based response refusa*ls. arXiv:2505.15365 [cs.HC]. https://doi.org/10.48550/arXiv.2505.15365

[33] Petrakaki, D., & Kornelakis, A. (2025). What do content moderators do? Emotion work and control on a digital health platform. *Journal of Management Studies, 63*(6), 2434–2465. https://doi.org/10.1111/joms.13219

[34] Petricca, P. (2020). Commercial content moderation: An opaque maze for freedom of expression and customers' opinions. *Rivista internazionale di Filosofia e Psicologia, 11*(3), 307–326. https://doi.org/10.4453/rifp.2020.0021

[35] Ruckenstein, M., & Turunen, L. L. M. (2019). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society, 22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990

[36] 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. https://doi.org/10.1145/3479512

[37] Schluger, C., Chang, J. P., Danescu-Niculescu-Mizil, C., & Levy, K. (2022). Proactive moderation of online discussions: Existing practices and the potential for algorithmic support. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–27. https://doi.org/10.1145/3555095

[38] Schöpke-Gonzalez, A. M., Atreja, S., Shin, H. N., Ahmed, N., & Hemphill, L. (2022). Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society, 26*(10), 5677–5701. https://doi.org/10.1177/14614448221138529

[39] Seering, J. (2020). Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), Article 3415178, 1–28. https://doi.org/10.1145/3415178

[40] Shahid, F. (2024). Human-AI collaboration to facilitate culturally-aware content moderation. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, (pp. 9–12). Association for Computing Machinery. https://doi.org/10.1145/3678884.3682057

[41] Siapera, E. (2022). AI content moderation, racism and (de)coloniality. *International Journal of Bullying Prevention, 4*, 55–65. https://doi.org/10.1007/s42380-021-00105-7

[42] Spence, R., & DeMarco, J. (2025). Content moderator mental health and associations with coping styles: Replication and extension of previous studies. *Behavioral Sciences, 15*(4), Article 487. https://doi.org/10.3390/bs15040487

[43] Tarvin, E., & Stanfill, M. (2022). "YouTube's predator problem": Platform moderation as governance-washing, and user resistance. *Convergence: The International Journal of*

Research into New Media Technologies, 28(3), 822–837. https://doi.org/10.1177/13548565211066490

[44] Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society, 10*(1). https://doi.org/10.1177/20539517231172424

[45] Wang, W., Huang, J., Huang, J., Chen, C., Gu, J., He, P., & Lyu, M. R. (2023). An image is worth a thousand toxic words: A metamorphic testing framework for content moderation software. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering*, v(pp. 1339-1351). IEEE. . https://doi.org/10.1109/ASE56229.2023.00189

[46] Wang, S., & Kim, K. J. (2023). Content moderation on social media: Does it matter who and why moderates hate speech? *Cyberpsychology, Behavior and Social Networking, 26*(7), 527–534. https://doi.org/10.1089/cyber.2022.0158

[47] Wiesner, A., Schäfer, S., & Lecheler, S. (2023). Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. *New Media & Society, 27*(3), 1215–1234. https://doi.org/10.1177/14614448231190901

[48] Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., & Boon, M. (2021). Can AI enhance people's support for online moderation and their openness to dissimilar political views? *Journal of Computer-Mediated Communication, 26*(4), 223–243. https://doi.org/10.1093/jcmc/zmab006

[49] Young, G. K. (2021). 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

[50] Yu, Z., Otto, L., Assenmacher, D., & Wagner, C. (2024). A systematic review of the effects of AI-assisted moderation on individuals and groups. *Human-Machine Communication, 9*, 167–188. https://doi.org/10.30658/hmc.9.10