STATUS QUO OF SPECIFIC MEASURES OF HOSTING SERVICES FOR CONTENT MODERATION

Study commissioned by the Federal Network Agency

Reference number: [_____]

Status: 23.11.2023

Contractor
Goldmedia GmbH Strategy Consulting
Prof Dr Klaus Goldhammer | Dr André Wiegand
Oranienburger Str. 27, 10117 Berlin
Phone +49 30 2462660
www.goldmedia.com

Machine translation

Contents

Gender note: For reasons of better readability, the simultaneous use of male and female language forms is avoided in the text. All personal designations (e.g. "user" or "moderator") nevertheless apply to both genders.

This does not imply any discrimination against the female gender, but should be understood as gender-neutral in the sense of linguistic simplification.

1 Situation and mission

1.1 Situation

The EU Regulation on combating terrorist content online (EU 2021/784; TCO Regulation) and its German accompanying legislation (Terroristische-Online-Inhalte-Bekämpfungsgesetz; TerrOlBG) have given the Federal Network Agency (Bundesnetzagentur - BNetzA) new tasks. In cooperation with the Federal Criminal Police Office (Bundeskriminalamt - BKA), the BNetzA is now responsible for requiring hosting services to take effective action against the dissemination of terrorist content through their services in accordance with the Digital Service Act (DSA) in the event of certain violations of the TCO Regulation.

To this end, the BNetzA can oblige hosting services that are exposed to terrorist content to take specific measures to systematically prevent the dissemination of terrorist content, among other things.

These specific measures must be appropriate to enable hosting services to identify and remove terrorist content themselves in a timely manner. Examples of specific measures include adequate staffing, appropriate technical means or user-friendly reporting mechanisms for the users of the respective services.

Such specific measures are colloquially referred to as "content moderation". Content moderation refers to the procedures and organised practices for reviewing user-generated content posted on online platforms or other hosting services to determine whether content is appropriate for, among other things, a service or under a particular jurisdiction. The procedures may result in user-generated content being removed by moderators acting on behalf of or as a proxy for a service. Online platforms, especially social media, generate large amounts of user-generated content of various types (text, video, audio, live content such as streams, chats, etc.). For this content, there is a need for the publishing services to enforce their own rules (hereinafter: "Community Guidelines") and the applicable legal framework, as the publication of inappropriate content can pose a significant liability risk for the service.[1]

In Germany, this legal framework essentially consists of the EU's Digital Services Act (DSA), which replaces the German Netzwerkdurchsetzungsgesetz (NetzDG), which previously only applied nationally, the TCO Regulation, which is the focus here, the provisions of the German Criminal Code and the provisions of the Youth Protection Act and the Interstate Treaty on the Protection of Minors in the Media.

[1] Cf. https://link.springer.com/referenceworkentry/10.1007/978-3-319-32001-4_44-1, accessed on 22/09/23

1.2 Study contract

According to the TerrOIBG, a hosting service that is exposed to terrorist content must first decide for itself which "specific measures" it will take to curb its dissemination. The hosting service is free to take any measures it deems appropriate to combat the availability of terrorist content on its services.

The BNetzA then assesses the effectiveness and appropriateness of "specific measures" to curb terrorist content in accordance with Art. 5 Para. 3 TCO Regulation, taking into account the size and financial strength of the respective hosting service.[2]

In order to make proportionate decisions, the BNetzA needs a basic understanding of the possible and appropriate content moderation measures that a hosting service can take to counteract the dissemination of illegal and, in particular, terrorist content via its service.

In order to be able to assess this properly and, if necessary, in a court of law, this study first presents the status quo of the existing, relevant market practices of hosting services for content moderation in order to counteract the dissemination of illegal content. In addition, the existing and developing technical procedures for content moderation and the associated costs are analysed (work package 1).

Subsequently, minimum standards of specific measures will be developed that can be demanded from hosting services in specific individual cases, especially if they are repeatedly exposed to terrorist content. The associated effort should be targeted and proportionate, particularly with regard to the individual financial strength of the hosting service (work package 2).

1.3 Methodology

The status quo analysis is based on the following steps:

This was preceded by a categorising normative analysis of the complex interplay of EU law with national, public and private, mandatory and soft-law-based standards in order to examine which types of providers must implement content moderation measures on the basis of which legal requirements, even independently of the new TCO regulation.

The next step was to evaluate the existing NetzDG reports and initial DSA reports in order to summarise the information already available on the measures taken to deal with the volume of reported/identified and remedied infringements and the content moderation measures used by the companies subject to reporting.

At the same time, desk research was carried out on the content moderation measures available and in use.

Based on this, expert interviews were conducted with hosting services as well as with external providers of manual and automated content moderation and reporting centres. To this end, an overview of the German hosting services market was first compiled. The

[2] If the assessment is negative, the hosting service must be requested to take further measures. In this case, too, the provider itself shall specify which further specific measures it wishes to take (Art. 5 para. 6 subpara. 2 TCO Regulation).

hosting services to be analysed were agreed with the client. The following expert interviews were then conducted:

Hosting services

- 8 Hosting service providers,
- including 2 very large online platforms

Moderation services

- 6 Providers of moderation services

Reporting centres

- eco - Association of the Internet Industry
- State Media Authority of North Rhine-Westphalia (Landesanstalt für Medien – LfM)

The following questions were discussed with the various interviewees, whereby the focus was variable and determined by the interest of the respective interview:

- Company size
- Technical and personnel dimension
- Effectiveness of the measures
- Delimitation / additional expenses due to TCO-VO
- Severity of the impact
- Consideration of Art. 5 para. 3 lit. c TCO Regulation
- Outlook

In combination with the results of the NetzDG report evaluation, differences, similarities and consensual market standards (best practice variants) of content moderation for different hosting service types and sizes were identified.

It was then determined for small, medium-sized and large companies which measures are possible for each company and which measures depend on the respective company size and in what way.

At the same time, an assessment was also made as to whether the identified measures and systems for content moderation are sufficient in type and scope, in particular to remove terrorist online content in accordance with Art. 5 TCO Regulation.

A final assessment was made on this basis,

a) which providers should implement which measures for the implementation of the requirements of Art. 5 (2) and (3) TCO Regulation, taking into account the principle of proportionality pursuant to Art. 5 (3) (b) and (c) TCO Regulation, and
b) which measures are also fundamentally required by other regulatory requirements and can be adapted or extended to take account of the TCO Regulation.

2 General conditions

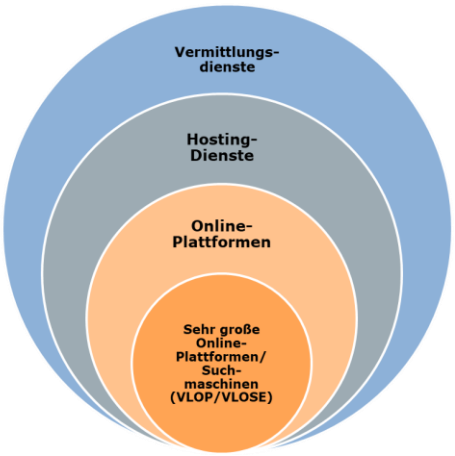
This introductory chapter summarises the legal framework for content moderation and shows which providers in Germany are covered by the TCO Regulation. This presentation forms the basis for analysing the content moderation measures used or to be examined in the context of the TCO Regulation.

2.1 Legal framework

2.1.1 Conceptual system of the Digital Service Act

To classify various digital services, the EU definition in the Digital Services Act (DSA) categorises various types of services, including online platforms, under the generic term "intermediary services". A special form of online platform is the very large online platform (VLOP) or very large online search engine (VLOSE). Intermediary services, on the other hand, fall into three sub-categories: pure conduit, caching and hosting services. The following diagram serves to illustrate the categories.

Fig. 1 Differentiation of digital services according to DSAcategories



Source: BNetzA 2023

It is important to note that the categories shown in the diagram build on each other and contain each other: A very large online platform thus always fulfils the criteria for an online platform, as well as those of a hosting service and those of an intermediary service. A hosting service, for example, is always a subset of the intermediary services, but a hosting service is not necessarily an online platform or a very large online platform.

In contrast to this, the three sub-categories for **intermediary services** do not build on each other. A service is an intermediary service if it fulfils the criteria for either a pure transmission service, a caching service or a hosting service. Pure conduit and caching services can claim disclaimers as long as they do not alter the integrity of the information transmitted or provided.

A hosting service stores information provided by users on their behalf. In this context, users can be any natural or legal person who utilises hosting services with the aim of

making information accessible. Hosting services[3] are exempt from liability as long as they promptly remove content or block access to it as soon as they receive actual knowledge of illegal activities or illegal content.[4] These procedures for obtaining disclaimers were introduced in the EU e-commerce directive for hosting services under the term "notice and takedown"[5] and conceptually expanded in the DSA as the "notice and action" mechanism.[6]

A hosting service is also **an online platform** if its core business involves the storage and public dissemination of information on behalf of a user. Excluded from this definition are hosting services where the dissemination of user-generated content is only a secondary function of another service or an insignificant auxiliary function of the main service.[7] In the case of an online platform, the public dissemination is necessarily carried out by the service, whereas in the case of a hosting service that is not an online platform (e.g. pure cloud services), the users (or third parties) decide on the public dissemination of data.

According to the DSA, very large online platforms (VLOPs) and very large online search engines (VLOSEs) are services that have an average monthly number of at least 45 million active users in the EU (and therefore reach more than 10 per cent of the 450 million consumers in the EU). To date, the following 19 companies have been categorised as very large online platforms by the EU Commission:

3 Exceptions to this apply in particular to online platforms that operate e-commerce.

4 Cf. recitals 21, 22 DSA

5 Cf. Directive 2000/31/EC "Directive on electronic commerce" Art. 14

6 Cf. Art. 14 DSA

7 Cf. Art. 3 para. 1 lit. i DSA

Tab. 1 Very large online platforms and search engines in the EU

Very large online platforms	Average users per month in the EU (company data)
Alibaba AliExpress	>45 million
Amazon Store	>45 million
Apple AppStore	>45 million
Booking.com	k. A.
Facebook	255 million
Google Play	274.6 million
Google Maps	278.6 million
Google Shopping	74.9 million
Instagram	250 million
LinkedIn	k. A.
Pinterest	k. A.
Snapchat	k. A.
TikTok	100.9 million
Twitter/X	100.9 million
Wikipedia	k. A.
YouTube	401.7 million
Zalando*	27,449 million (76,247 million for retail service and platform service)
Very large online search engines	**Average users per month in the EU (company information)**
Bing	107 million
Google Search	278.6 million

Sources: European Commission 2023; user data according to Reuters, see https://www.reuters.com/technology/google-twitter-meta-face-tougher-eu-online-content-rules-2023-02-17/, retrieved on 31/08/2023

**Zalando: https://en.zalando.de/legal-notice/*

Note: Zalando is currently defending itself against classification as a VLOP

See: https://www.handelsblatt.com/unternehmen/handel-konsumgueter/online-modehaendler-zalando-geht-mit-klage-gegen-strengere-eu-regulierung-vor/29227268.html, retrieved on 09/10/2023

2.1.2 Requirements for content moderation according to DSA

The DSA applies to all intermediary services from 17 February 2024. For very large online platforms and search engines already identified by the EU Commission, it has already applied since 25 August 2023.[8] New VLOPs designated in the future will be granted a period of 4 months to comply with the requirements of the DSA.[9]

All intermediary services must comply with orders from the competent national judicial or administrative authorities to take action against "illegal content"[10] . These "orders to take action against illegal content" must contain information about the reason why the content in question is illegal and the territorial scope of the order. Depending on the territorial scope, the intermediary service must then either delete the content completely or block access to it in certain countries. The intermediary service must inform the authority when the order has been complied with. However, there are no direct time limits associated with the order.[11] There is explicitly no general obligation to proactively monitor user contributions or actively investigate them.[12]

All intermediary services must explain to users in their general terms and conditions what restrictions apply to the information provided by users. The guidelines, procedures, measures and tools used to moderate content, including algorithmic decision-making and human review, as well as the procedural rules of the service's internal complaints management system must be made transparent.[13]

In order to ensure an appropriate level of transparency and accountability, all intermediary services above the threshold of small and micro enterprises within the meaning of Recommendation 2003/361/EC of the EU Commission are obliged to prepare an annual public report on their content moderation activities.[14] This report must show in aggregated form which type of violations (illegal content, violations of terms and conditions/community rules) were identified on which (legal) basis and how these violations were dealt with in which median periods.[15]

This "transparency report" should also contain a qualified description of automated means of content moderation, stating the exact purposes, accuracy indicators and the error rate of these automated means.[16]

---

8 In this context, adjustments are currently being made to the terms of use of the very large online platforms in order to comply with the regulatory framework of the DSA.
Cf. https://www.facebook.com/legal/terms_preview_DSA, accessed on 20/09/2013

9 Cf. https://digital-strategy.ec.europa.eu/de/policies/dsa-vlops, accessed on 22/09/23

10 Illegal content" is defined as any content that "... does not comply with Union law or the law of a Member State". Cf. Art. 3 lit. h DSA

11 Cf. Art. 9 DSA

12 Cf. Art. 8 DSA

13 Cf. Art. 14 para. 1 DSA

14 Cf. recital 49 DSA

15 Depending on whether the content is "illegal" or "only" a violation of the community rules, the service can take various moderation measures to curb undesirable content, e.g. low distribution, demonetisation, blocking of access or removal of content. User accounts can be closed or a warning issued ("strike"). See Art. 3 lit. t DSA

16 Cf. art. 15 para. 1 lit. e DSA

All hosting services, regardless of size, must also provide easily accessible and user-friendly reporting and remediation procedures that make it easy to report to the hosting service certain information that a reporting party considers to be illegal content.[17]

If a hosting service also fulfils the criteria of an online platform, it must also set up a complaints management system to enable internal dispute resolution.[18] In addition, all online platforms must submit each individual case of a moderation decision that leads to restrictions for a user (downgrading the visibility of content, spatial restriction of visibility, removal of content, blocking of users) to the newly established DSA Online Transparency Database immediately after the conclusion of the procedure.[19] The following graphic shows an example of how these decisions are to be transmitted to the database:

Fig. 2 Example message from TikTok to the DSA Transparency Database

Statement of reason details: 626604f4-ce5f-4cbd-945d-0a1eef5fac88

Platform name	TikTok
Received	2023-10-26 09:49:59 UTC
Visibility restriction of specific items of information provided by the recipient of the service	Removal of content
Facts and circumstances relied on in taking the decision	The decision was taken pursuant to own-initiative investigations.
Ground for Decision	Content incompatible with terms and conditions
Reference to contractual ground	Harassment and Bullying
Explanation of why the content is considered as incompatible on that ground	We welcome the respectful expression of different viewpoints, but not toxicity or trolling. We do not allow language or behavior that harasses, humiliates, threatens, or doxes anyone. This also includes responding to such acts with retaliatory harassment (but excludes non-harassing counterspeech).We proactively enforce our Community Guidelines through a mix of technology and human moderation. We have detected this policy violation using automated measures. We have used automated measures in making this decision.
Is the content considered as illegal?	No
Territorial scope of the decision	Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden
Content Type	Text
When the content was posted or uploaded	2023-10-26
Category	Illegal or harmful speech
Information source	Own voluntary initiative
Was the content detected/identified using automated means?	Yes
Was the decision taken using other automated means?	Fully automated
Application date of the decision	2023-10-26

Source: https://transparency.dsa.ec.europa.eu/statement/626604f4-ce5f-4cbd-945d-0a1eef5fac88, retrieved on 22/09/23

Very large online platforms harbour particular (systemic) risks for the distribution of illegal content and for damage to society.[20] For this reason, these companies are subject

[17] Cf. recital 50 DSA

[18] Cf. art. 20 para. 3 DSA

[19] Cf.: Art. 24 para. 5 DSA

[20] Cf. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_de, accessed on 22/09/23

to particularly strict requirements. They are required to assess their systemic risks, establish simple reporting channels and provide robust content moderation tools. They must take risk mitigation measures - for example to counteract the spread of disinformation - and have their risk management audited externally.[21] The following table provides an overview of the graduated requirements of the DSA at the level of intermediary services, hosting services, online platforms and VLOPs.

Tab. 2 New obligations for services according to DSA

New (cumulative) obligations	Mediation services	Hosting services	Online platforms	Very large platforms
Transparency reporting obligations	●	●	●	●
Message DSA Transparency Database			●	●
Co-operation with national authorities for orders	●	●	●	●
Details of contact points and legal representation, if applicable	●	●	●	●
Notification and remedy and obligation to inform users	-	●	●	●
Reporting of criminal offences that may pose a "threat to the life or safety of a person"	-	●	●	●
Complaints and redress mechanism, out-of-court dispute resolution	-	-	●	●
Trustworthy whistleblowers	-	-	●	●
Measures against abus. Reports and counter-reports	-	-	●	●
Special obligations for e-commerce platforms, e.g. checking the authorisations of third-party providers, compliance by design, random checks	-	-	●	●
Prohibition of advertising aimed specifically at children or using special personal data	-	-	●	●
Transparency of the recommendation systems	-	-	●	●
Transparency of online advertising towards users	-	-	●	●
Commitment to risk management and crisis response	-	-	-	●
External and independent review, internal compliance function and public accountability (audits)	-	-	-	●
Data access for Digital Services Coordinator, EU-KOM, lic. Researcher	-	-	-	●
Option for users to reject recommendations based on profiling	-	-	-	●
Codes of conduct	-	-	-	●
Cooperation in the event of a crisis	-	-	-	●

[21] Cf. https://ec.europa.eu/commission/presscorner/detail/de/ip_23_2413, accessed on 22/09/23

Source: European Commission: "Digital Services Act: more security and responsibility in the online environment" (as of September 2023)

2.1.3 Requirements for the moderation of content in accordance with the TCO Regulation

Content that directly or indirectly incites terrorist offences by advocating or glorifying them or otherwise contributing to their commission[22] is generally covered by the DSA as illegal content. However, the TCO Regulation expands the instruments for curbing terrorist content and places higher requirements on hosting services with regard to the moderation of terrorist content than the DSA.

The TCO Regulation created the new instrument of the "removal order". In contrast to an "order to take action against illegal content" pursuant to Art. 9 DSA, hosting services are obliged to remove terrorist content or block access to terrorist content within one hour of receiving the removal order. [23]

The deadline for the removal of terrorist content is significantly shorter than comparable, previous time periods that have applied to the removal of content for copyright infringements, hate speech or criminal content, for example under the Digital Millennium Content Act (DMCA), the Copyright Service Provider Act (UrhDaG), the Netzwerkdurchsetzungsgesetz (NetzDG) or the Digital Service Act (DSA).

The TCO Regulation applies in principle to hosting services within the meaning of Art. 1 (1) (b) of EU Directive 2015/1535 on information society services, whereby the Federal Network Agency interprets the wording "generally for remuneration" in such a way that it also covers services that are offered free of charge, provided that they are very similar to services that are usually provided for remuneration. In the opinion of the Federal Network Agency, the decisive factor is that comparable services in the same service category are usually provided for a fee. The criterion is also fulfilled, for example, if a user authorises the service provider to use their data (e.g. for marketing to advertisers) in return for the service. It also includes cases of indirect financing, e.g. where the service provider offers the service without direct monetary remuneration for marketing or advertising purposes and cross-subsidises it with other products or services.

The TCO Regulation also applies to small and micro-enterprises as defined in Commission Recommendation 2003/361/EC[24] . However, the size of the company is to be taken into account in any penalties, along with other factors (Art. 18 (2) (f) TCO Regulation).

In Germany, the BKA is the competent authority that can issue such removal orders. The IT system PERCI[25] (see Annex, Chapter 10.2) is to be used to transmit removal orders between the BKA and hosting services, competent authorities in other EU countries and Europol.

---

22 Cf. Art. 2 para. 7 TCO Regulation

23 Cf. Art. 3 para. 3 TCO Regulation

24 Less than 250 employees and a maximum annual turnover of EUR 50 million or a maximum annual balance sheet total of EUR 43 million.

25 Plateforme Européenne de Retraits de Contenus Illégaux sur Internet

The extent to which hosting services or "suspended" hosting services are able to respond to removal orders in a timely manner cannot yet be answered empirically. In 2022, no removal orders were issued by the Federal Criminal Police Office on the basis of Art. 5 TCO Regulation. The Federal Criminal Police Office also did not receive any removal orders from foreign authorities to check their legality. The PERCI IT system, which is used for the transmission of removal orders between the competent authorities and hosting services, only went live in summer 2023. Public information on the volume of reports is not yet available.

As no removal orders were issued in 2022, there were consequently no hosting services categorised as "exposed to terrorist content" in 2022. This means that no German hosting services were obliged to take "specific measures" under Art. 5 TCO Regulation in 2022.[26]

By contrast, the Federal Criminal Police Office sent 10,472 referrals, deletion requests to remove or block content, to hosting services in 2022. Unlike removal orders, referrals are not legally binding. Deletions as a result of referrals are carried out by hosting services on a voluntary basis. In 2022, hosting services complied with the BKA's deletion requests by removing or blocking content in 88 per cent of cases (9,207).[27]

The lack of removal orders in the remaining cases was partly due to the fact that removal requests related to content that was criminally relevant but did not constitute terrorist content within the meaning of the TCO Regulation, or that hosting services were based outside the EU.[28]

2.1.4 NetzDG - Review

In Germany, a legal obligation to moderate content has been in force since 2018 with the Netzwerkdurchsetzungsgesetz (NetzDG). Social networks[29] that have at least two million registered users in Germany are subject to the NetzDG.

The NetzDG lacked the instrument of official orders. However, it did oblige social networks to submit half-yearly transparency reports. These reports had to state the frequency with which complaints were received for specific offences under the Criminal Code and how these complaints were dealt with.[30]

The offences covered by the NetzDG are enumerated in Section 3a (2) NetzDG. The following table provides an overview. Some of the enumerated offences were directly or indirectly related to terrorist acts and thus to the scope of the TCO Regulation.

[26] ibid.

[27] ibid.

[28] ibid.

[29] The definition of a social network in the NetzDG essentially corresponds to an online platform in the DSA (cf. Section 1 (1) NetzDG).

[30] Cf. § 2 NetzDG

Tab. 3 Enumerated offences in the NetzDG

Paragraph	Offence	Reference to terrorism
§ Section 86 StGB	Distributing propaganda material of unconstitutional organisations	indirect
§ Section 86a StGB	Use of signs of unconstitutional organisations	indirect
§ Section 89a StGB	Preparation of a serious act of violence endangering the state	direct
§ Section 91 StGB	Instruction to commit a serious act of violence endangering the state	direct
§ Section 100a StGB	Treasonous forgery	indirect
§ Section 111 StGB	Public incitement to commit criminal offences	indirect
§ Section 126 StGB	Disturbing the public peace by threatening to commit criminal offences	indirect
§ Section 129 StGB	Formation of criminal organisations	no
§ Section 129a StGB	Formation of terrorist organisations	direct
§ Section 129b StGB	Criminal and terrorist organisations abroad; confiscation	direct
§ Section 130 StGB	Incitement to hatred	indirect
§ 131 StGB	Depiction of violence	indirect
§ Section 140 StGB	Rewarding and condoning criminal offences	indirect
§ Section 166 StGB	Insulting denominations, religious communities and ideological organisations	no
§ 184b i. V. m. § Section 184d StGB	Distribution, acquisition and possession of child pornographic content / making pornographic content available by means of broadcasting or telemedia; retrieval of child and youth pornographic content by means of telemedia	no
§ 185 StGB	Insult	no
§ Section 186 StGB	Defamation	no
§ Section 187 StGB	Slander	no
§ Section 201a StGB	Violation of the highly personal sphere of life and personal rights through image recordings	no
§ Section 241 StGB	Threat	no
§ Section 269 StGB	Falsification of evidentiary data	no

Source: Goldmedia on the basis of Section 3a (2) NetzDG

While the NetzDG only provides for moderation with regard to certain criminal laws, the DSA will introduce the obligation to moderate all illegal content. In principle, this includes the entire legal system.

The draft regulation for the Digital Services Act (DSA) repeals the Netzwerkdurchsetzungsgesetz (NetzDG) and the Telemedia Act (TMG) (see Article 37 of the DDG Authorisation Act). The NetzDG is currently no longer valid for very large online platforms.

2.2 Hosting services with a substantial connection to Germany

Hosting services that offer and disseminate information publicly in the European Union are subject to the TCO Regulation.[31] If there is a significant connection to Germany, e.g. due to its establishment, a significant number of users in Germany or a specific focus on the German market, German authorities such as the BKA and the BNetzA[32] are generally responsible. [33]

2.2.1 Types of hosting services

The following is a rough overview of hosting services that have a significant connection to Germany. Tab. 4 provides an overview of the 50 hosting services with the widest reach in Germany in 2022, clearly showing the diversity of topics among the hosting services with the widest reach in Germany.

Tab. 4 Top 50 most accessed hosting services in Germany in 2022

Domain	Users per month in million, Total January - December 2022	Average users in million per month 2022
google.com	890,6	74,2
youtube.com	300,2	25,0
facebook.com	281,3	23,4
amazon.com	246,2	20,5
wikipedia.org	232,8	19,4
google.com	185,0	15,4
bild.de	166,3	13,9
ebay.de	123,1	10,3
instagram.com	96,1	8,0
t-online.de	90,2	7,5
ebay-kleinanzeigen.de	86,2	7,2
spiegel.de	66,2	5,5
web.de	61,5	5,1
xhamster.com	59,1	4,9
dhl.de	57,7	4,8
tagesschau.de	57,5	4,8
twitch.tv	56,0	4,7

[31] Cf. Art. 1 para. 2 TCO Regulation
[32] Cf. § 1 TerrOIBG
[33] Cf. Art. 2 para. 5 TCO Regulation

Domain	Users per month in million, Total January - December 2022	Average users in million per month 2022
twitter.com/X	55,8	4,6
focus.de	55,4	4,6
paypal.com	50,6	4,2
samsung.com	49,8	4,2
gmx.net	46,4	3,9
otto.de	43,6	3,6
welt.de	42,8	3,6
weather.com	41,7	3,5
pornhub.com	40,4	3,4
n-tv.de	40,4	3,4
sport1.de	33,8	2,8
live.com	33,5	2,8
fandom.com	33,1	2,8
wetteronline.de	32,7	2,7
derwesten.de	32,3	2,7
netflix.com	32,3	2,7
merkur.de	32,2	2,7
zdf.de	31,6	2,6
chefkoch.de	31,2	2,6
idealo.de	30,9	2,6
whatsapp.com	30,1	2,5
yahoo.com	29,8	2,5
immobilienscout24.de	29,5	2,5
chip.de	29,2	2,4
accuweather.com	28,9	2,4
vodafone.com	28,8	2,4
mobile.de	27,8	2,3
booking.com	27,4	2,3
ndr.de	26,5	2,2
teads.tv	23,8	2,0
presscompass.net	23,3	1,9
kaufland.de	23,2	1,9
tz.de	23,2	1,9

*Source: Semrush 2022: The most visited and most viewed websites in Germany in 2022*https://de.serush.com/blog/top-der-meistbesuchten-webseiten/

In order to capture the diversity and complexity of these hosting services in more detail, they have been divided into different categories below, each of which reflects a specific function and focus. The categories are divided into DSA upper categories and services categories, whereby upper categories include hosting services, online search engines and online platforms. Service categories usually contain several hosting services or online platforms with a similar function, which can sometimes differ greatly from one another in terms of size, user base and functionalities.

Tab. 5 Overview of hosting services according to DSA classification

DSA-Super category	Service categories
Hosting services	
hosting services, which are not online platforms	Web hosting (incl. file hosting and website construction kits)
	e-mail services, short message services and other interpersonal communication services
	Cloud computing (Hardware-, Computing-Power- and Security-as-a-Service)
	Cloud storage
	Collaboration platforms (Software-as-a-Service)
Online search engines	Online search engines
Online platforms, hosting services that publish customer/user content	
Online platforms for distance selling	Online trading centres
	VoD platforms (incl. transaction-based VoD)
	Audio streaming platforms (Spotify)
	E-book/e-journal platforms
	App stores
	Collaborative economy platforms: marketplaces (accommodation, small businesses, cars, property, doctors, craftsmen, crowdfunding ...) and sharing economy
	Dating sites/dating platforms
	Data marketplaces
	...
Online platforms (social media)	Generic social networks (Facebook, Instagram, TikTok ...)
	Social business networks (LinkedIn, XING ...)
	Social video platforms incl. live streaming (YouTube, Twitch ...)
	Social audio platforms (Soundcloud, ...)
	Short message services/micro-blogging (X ...)
	Special interest communities/forums/image boards (e.g. on gaming, Q&A, DIY, books, music, recipes, sport, art ...)
	Social gaming (Roblox, Steam ...)
	Dating
	Social inspiration (Pinterest, Flickr ...)
	Online encyclopaedias
	Location-based services (Google Maps, Yelp, Tripadvisor ...)
	Document portals (Scribd, Doc-Player, Slide-Player ...)
	...

Source: Goldmedia analysis 2023

The requirements for content moderation vary depending on the focus of the platform. Platforms that discuss politics or controversial social issues generally require more moderation than travel, recipe or career portals, for example. Online platforms with a focus on gaming can also be moderation-intensive, although this depends heavily on the respective game title and its gaming culture. Platforms aimed at children and young people are also moderation-intensive, as an age-appropriate environment must be maintained in addition to the usual moderation tasks. In the case of online platforms for distance selling, on the other hand, the moderation effort consists primarily of preventing and combating fraud and less in the moderation of violent or hateful content.

Tab. 6 Monthly reach of selected domains (desktop and online) according to Similarweb in August 2023, in million visits

Service category	Online platform	million visits per month
Social video platform	YouTube.com	33.900,0
Social network	Facebook.com	17.400,0
Social network	Instagram.com	6.700,0
Micro-blogging platform	Twitter.com/X	6.400,0
Social commerce platform	Amazon.com	2.500,0
Social video platform	TikTok.com	2.300,0
Micro-blogging platform	Reddit.com	1.900,0
Social video platform	Twitch.tv	1.100,0
Social gaming platform	Roblox.com	866,2
Q&A platform	Quora.com	712,6
Social commerce platform	Amazon.de*	445,8
Micro-blogging platform	Tumblr.com	217,0
Social network	Snapchat.com	187,6
Social gaming platform	Steampowered.com	161,2
Music platform	SoundCloud.com	126,0
Social video platform	Vimeo.com	73,3
Q&A platform	Good question.net	40,5
Social video platform	Bitchute.com	17,9
Special Interest Forum	moviepilot.com	17,4
Special Interest Forum	Gamestar.de	12,3
Special Interest Forum	Computerbase.de	9,6
Micro-blogging platform	Mastodon.social	4,6
Social network	Nebenan.de	4,5
Social network	Jappy.com	2,4
Social network	StayFriends.com	2,4
Special Interest Forum	Boersennews.com	1,9
Q&A platform	who-knows-what.com	1,1
Social video platform	YouNow.com	0,8
Special Interest Forum	Musician-board.com	0,7
Special Interest Forum	Worldofplayers.com	0,7
Special Interest Forum	Bvb-forum.de	0,6

** additional specification of the DE top-level domain to the COM top-level domain*

Source: https:www.similarweb.com, retrieved on 15.09.23

2.2.2 Further hosting services

Hosting services that realise the technical hosting on behalf of online platforms are largely excluded from liability for the content they provide as long as they do not change the integrity of the information provided. These include web and file hosters, data centres, cloud storage providers, CDN services and similar intermediary services. They do not have to moderate any content themselves.

However, these hosting services, which act on behalf of other hosting services, are also obliged to promptly remove content or block access to it as soon as they become aware of illegal activities or content.

3 Status quo of content moderation

3.1 Introduction

Content moderation is based on the general terms and conditions of the services and the community guidelines ("Community Standards", "Netiquette", "Code of Conduct", etc.) anchored therein. As a rule, the regulations enshrined in a service's community guidelines go well beyond the legal requirements. Illegal content is therefore often already moderated via the general processes that ensure compliance with a service's community guidelines.[34]

The permissibility of content can vary greatly from service to service, depending on the focus and target group of the service. A service that is primarily aimed at children and young people, for example, usually has stricter moderation requirements, particularly with regard to vulgar language or sexual content, than a service that is primarily aimed at adults.

Specific areas present particular moderation challenges. For example, in-game chats in certain computer games, which are highly competitive, often lead to hate and inhumane comments. In the area of services aimed at children and young people, cybergrooming is one of the biggest challenges.

While easily identifiable violations of the community guidelines in text, image or video form are already recognised by upload filters and blocked without manual review, depending on the platform, the moderation of more complex content (e.g. longer text contributions on topics that tend to be critical), which are "flagged" by the moderation software or reported by users, sometimes takes place via a multi-stage process.

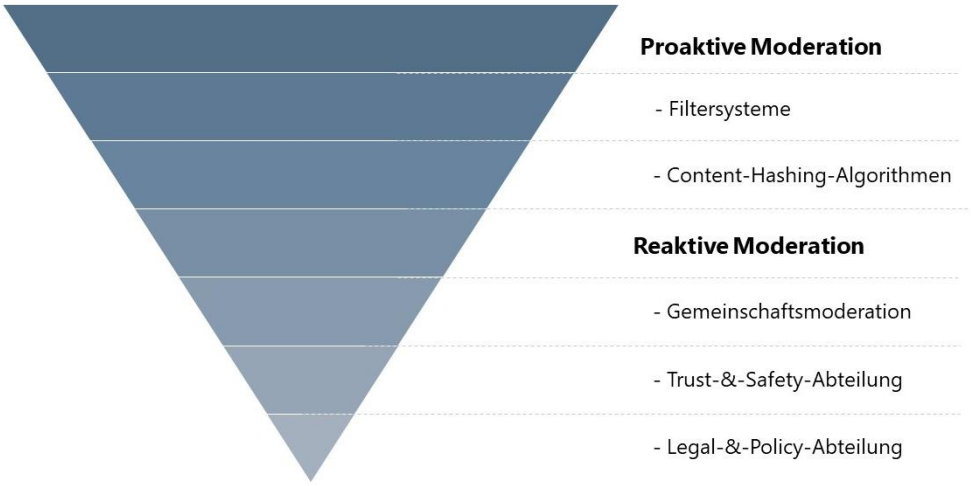
This process involves individual decisions by a moderator, obtaining second and third opinions from other subject area experts within the trust and safety team and legal review steps, some of which can take place at several levels.[35] For some providers, for

[34] If necessary, more specialised moderation teams are also deployed to ensure compliance with (country-) specific moderation requirements. For example, reports under the NetzDG are moderated by specialised NetzDG teams as long as the reporting party makes reference to the NetzDG within the report.

[35] In the case of the YouTube service, the escalation process ranges from referral to the company's own legal department to the involvement of the higher-level legal department of Google Germany and the involvement of external criminal law legal advice (see Google 2023: YouTube Transparency Report 2nd half-year 2022).

example, particularly sensitive decisions are no longer made by the trust and safety team but by the higher-level legal and policy team, in some cases only after consultation with external legal advice.

Fig. 3 Content moderation process (schematic diagram)



Source: Goldmedia Analysis 2023

Moderation guidelines must be revised periodically, as the services are constantly challenged by new forms of abusive content, e.g. in the area of disinformation. For larger services, this takes place in structured updating processes, sometimes with legal advice from the legal and policy team. The exchange on new challenges at the various levels of content moderation also takes place horizontally in order to be able to react to new moderation situations with sufficient agility. Non-harmonised moderation decisions are therefore most likely to occur in new moderation situations, especially if different teams at different locations or countries are involved. It is then the task of the legal and policy team to ensure that the community guidelines are applied uniformly.

3.1.1 Proactive moderation procedures

In the area of proactive moderation procedures, automated systems are mainly used, as described in section 3.3.1. A distinction can be made here between fully automated filter systems that directly prevent a post from being published and filters that "flag" posts for subsequent manual review.

Even though numerous moderation tools are offered commercially by service providers, most services either also or even primarily use systems developed in-house. There are many reasons for this: some of the services have been on the market longer than solutions from service providers, while others are too specific in their content for a generic moderation solution to be of much use. Another key point is that services that require moderation solutions usually have sufficient internal IT development capacity to develop customised solutions for their own purposes. In the best case scenario, this is done in a safety-by-design approach even before or parallel to the development of user-centred service features.

Manual moderation by humans is also sometimes used as a proactive moderation process, even if it is much less important than automated systems. Manual moderation is primarily used to monitor general trends and developments on the platform, particularly in sensitive areas that require moderation, such as politics, or in special situations.

3.1.2 Reactive moderation process

In the area of reactive moderation, moderation is carried out exclusively manually at the time the study is created, i.e. by human decisions of the so-called "content reviewers".

Reactive moderation procedures presuppose that a content

a) has either been categorised as problematic by an automated system with sufficient probability (see section 3.1.1),

b) or that content has been reported.

Reports are made either by users or the community, by authorities or by reporting centres.

Users report problematic content via the contact address specifically set up on the website of the hosting service or via an input mask set up for this purpose. Services that provide for reports via input masks use these for the initial classification of complaints by asking the user to specify the respective legal area affected by the report. Depending on the initial classification of the reason for the report, different moderation processes ("cues") are triggered depending on the size of the hosting service and, for example, the report is passed directly to a team of moderators entrusted/specialised with the legal area.

However, the origin of the report also defines how a report is handled:

Public authorities (public flaggers) usually communicate with the hosting services via their own interfaces or contact points (examples: Contact point of the BKA/contact interface PERCI or the BNetzA list of legal representatives in accordance with the TCO Regulation).

Reports from users who have a certain status or a moderator role in a community or who have often provided reliable reports in the past are prioritised. In some cases, these users also have dedicated reporting channels to the hosting service as user moderators or even have restricted access to the moderation system.

In addition to these "private flaggers", many platforms also assign a special status to NGOs operating in the market that are committed to combating hate speech or child abuse. The "YouTube Priority Flagger Programme", for example, prioritises not only reports from authorities (public flaggers) but also reports from NGOs that have already been particularly effective in the past in reporting YouTube content that violates the community guidelines.[36] In 2020, around 30 organisations in Germany were part of the YouTube Priority Flagger Program, compared to around 180 organisations worldwide.[37]

[36] Cf. YouTube Help "The YouTube Priority Flagger Programme", online at: https://support.google.com/youtube/answer/7554338?hl=de#:~:text=The%20YouTube%20Priority%20Flagger%20Program%20offers%20powerful%20tools%20for%C3%BCr%20complaints%20against%20the%20Community%2Guidelines%20infringing%C3%9F, accessed on 18.09.23

[37] Cf. HateAid "Trusted Flagger", online at: https://hateaid.org/trusted-flagger/, accessed on 08.09.23

Art. 22 of the DSA now stipulates that the state body of the Digital Services Coordinator (DDK/DSC) may appoint "trusted flaggers". Transparency obligations are imposed on these publicly appointed trusted flaggers in the DSA.

Tab. 7 Categories of private and public trustworthy whistleblowers

Model	Terms of Service	Liability
Private flagger	Hate Speech	Copyright holders; INHOPE
Private flagger with public endorsement	'Trusted flaggers' appointed under the Digital Services Act or NetzDGauto	
Public flagger	Police IRU	

IRU = Internet Referral Units
Source: Appelman, N. & Leerssen, P. (2022) "On "Trusted" Flaggers". Yale-Wikimedia Initiative on Intermediaries & Information, online at: https://law.yale.edu/sites/default/files/area/center/isp/documents/trustedflaggers_ispessayseries_2022.pdf, retrieved on 07.09.23

This emphasised position in the DSA further enhances the work of trustworthy whistleblowers.

The list of "trusted flaggers" named by the DDK is likely to have a high degree of overlap with the "trusted flaggers" already identified by the hosting services. In the further course of the study, however, the term "trusted flagger" will not be used in the sense of the DSA, but in the sense of the reporting centres and reporting individuals already identified as trustworthy and reliable by the respective services.

The appendix to this study describes all authorities and reporting centres relevant to German hosting services.

Ensuring standardised moderation decisions is a major challenge for the services, especially when it comes to very large online platforms that offer their services globally and have them moderated from different countries. The quality assurance of moderation decisions is therefore an integral part of the moderation process, even for smaller providers. Quality assurance is usually carried out by special teams made up of experienced moderators. [38]

3.1.3 Moderation decisions

Depending on the severity of a detected violation of the community guidelines, moderators can sanction the content or the user who posted the content. To this end, the service provider determines what an appropriate response to an offence should be; the moderator has little discretion in this regard.

The design of the exact sanction regime differs from platform to platform. However, some basic principles and instruments can be found in one form or another on all services: Firstly, the algorithmic distribution of content can be curbed or prevented if the

[38] Take YouTube, for example: Around 30 per cent of the content reviewed is checked by quality assurance teams. See Google 2023: YouTube transparency report 2nd half of 2022

content is undesirable but does not directly violate the community guidelines. Such content can then still be accessible to certain (private) user groups, but will no longer be marketed commercially. Content can also be deleted from the platform if the content violates the community guidelines or legal requirements.

3.1.4 Dealing with unauthorised content

If content violates the community guidelines of a service, it is removed and the posting user is informed of the removal. It is common for individual offences by a user to be counted and, if necessary, further sanctions imposed. Depending on which guidelines the content violates and depending on which and how many violations or warnings a user has already received, the account may be restricted in its functionality and temporarily or permanently deactivated.

Content that does not violate a service's community guidelines, but is still problematic or otherwise of low quality, may be restricted in its distribution. Platforms also use upstream notices of potentially sensitive or misleading content, even if it does not explicitly violate a service's community guidelines, to provide additional context to such content.

3.1.5 Dispute resolution

Online platforms maintain dispute resolution processes if users are of the opinion that content has been unjustifiably removed or not authorised for publication or that they have been unjustifiably warned. Services that are subject to the NetzDG treat complaints against moderation decisions on the basis of the NetzDG differently than on the basis of the other community guidelines. support mailbox. In the event of violations of a provision listed in the NetzDG, users receive an email stating the legal provision of the German Criminal Code that has been violated and the measures taken by the platform.

In accordance with Art. 17 DSA, Facebook now informs all users whose content has been categorised as unlawful or a violation of community guidelines by email about the moderation measure taken, including the reasons. In addition, in accordance with Art. 20 DSA, Facebook grants the affected users and, in the case of reported content, the reporting persons, access to the internal complaints management system for six months from receipt of the complaint. Complaints against the service's decision can be submitted via this system.

3.2 Operational process of content moderation

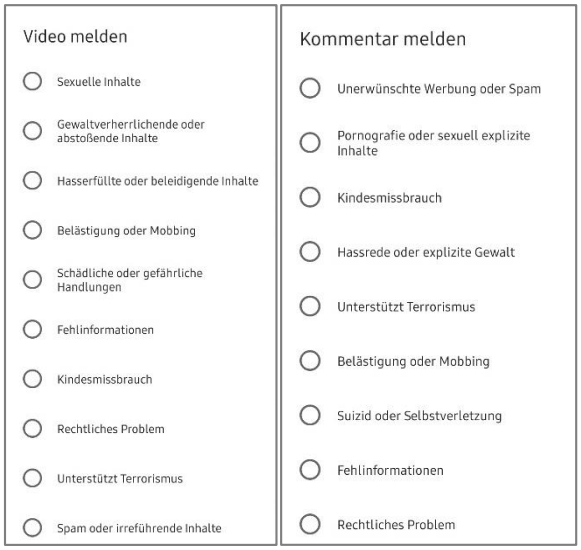
Content moderation removes content defined as undesirable by the community guidelines as well as illegal content or does not authorise it for publication, or access is restricted/downgraded, thus preventing (further) dissemination. However, a genuine "deletion" will not take place until the six-month objection period stipulated in the DSA has expired at the earliest.[39]

To facilitate processing, the moderation facts are first categorised. The initial categorisation is often carried out by the notifier during the notification process. Compact, intu-

[39] Cf. Art. 20 DSA

itively understandable taxonomies of up to 10 classes are used for this purpose. Depending on the initial classification of the reporting reason by the reporter, different moderation processes ("cues") can be triggered.

Fig. 4YouTube reporting functions for videos and comments



Source: YouTube app version 18.29.33, screenshots from September 2023

Larger platforms in particular have different processes depending on the type of report, which are then processed by different, specialised teams. The reports are usually categorised in more detail within the teams. The taxonomies differ from provider to provider, but a distinction between around 30-50 main categories and up to 100 subcategories was confirmed as standard in the industry during the interviews for this study.

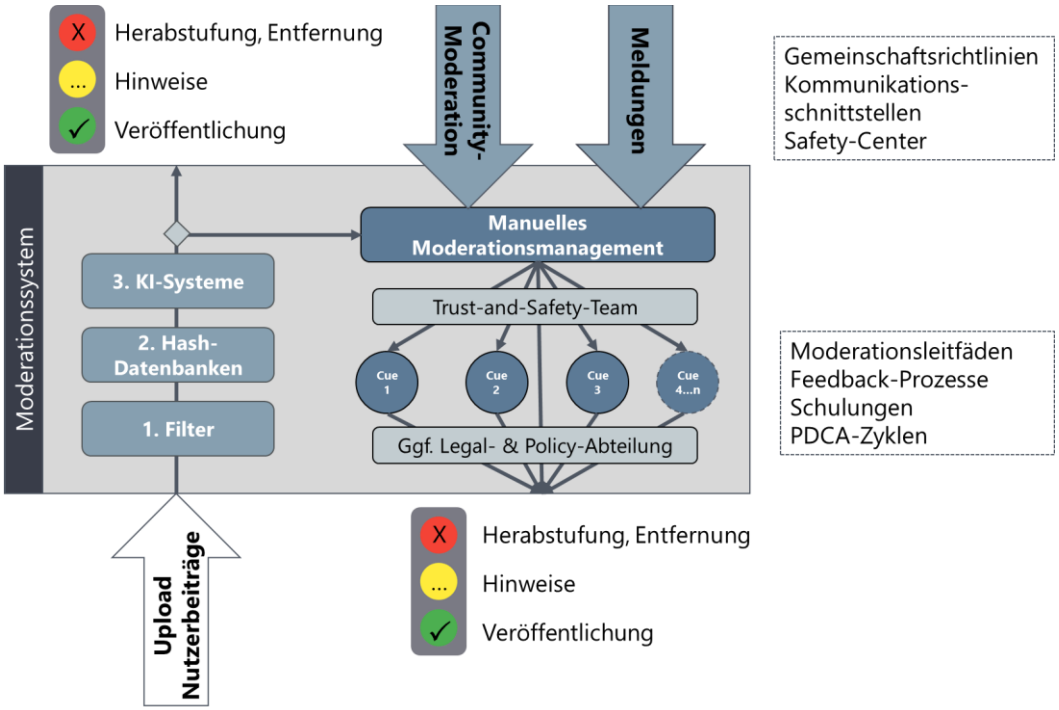
The following describes a simplified, generalised moderation architecture that is structurally applicable to all online platforms (see Fig. 5). It is important to note that the development of community guidelines must be distinguished from their application. The development of community guidelines must be carried out by the service; there is usually a separate department for this, which may be called Community, Community Management or Legal and Policy.

The application of community guidelines, on the other hand, is usually carried out in the Trust and Safety department, where all content moderation procedures are operationally controlled. Automated and manual procedures are used to fulfil the trust and safety tasks. It is common for both internal solutions and employees as well as external services to be used. However, at least in the case of large to very large online platforms, the use of internal solutions predominates for technical procedures and the use of external service providers predominates for manual procedures. In organisational terms, the moderators of larger services are divided into teams that fulfil a specific moderation task ("cues"). In the case of smaller providers, there is only one moderation team and a lower degree of specialisation among the moderators.

The starting point for content moderation is all content that is saved by users on an online service. The moderation process is controlled by a special moderation or moderation management platform. These can be comprehensive enterprise software solutions that are cloud-based, especially in the case of large online platforms, and enable the integration of technical subsystems (e.g. filter systems) as well as providing the graphical

user interfaces for the human moderators. The entire process, including various escalation stages and revision procedures, is mapped and logged by the central platform. Depending on the design of the process, particularly critical moderation decisions may sometimes be escalated beyond the safety team to the legal and policy team for clarification, for example in the case of particularly far-reaching moderation decisions with policy implications for the service.

Fig. 5 Exemplary moderation architecture of an online platform



Source: Goldmedia Analysis 2023

As can be seen in the schematic diagram, the central moderation system can be expanded modularly with any number of technical filter and auxiliary systems. It is irrelevant whether this is an in-house development or a purchased solution from a third-party provider. The integration of additional (sub-)systems may be necessary for various reasons, for example to train an AI-based system based on your own content and moderation decisions or to integrate additional service providers for moderation services (manual moderation procedures).

With regard to the **technical systems used,** the majority of the online services interviewed for the study work with solutions developed in-house, at least in some areas.[40] The majority of these use filter systems without AI support or their own moderation management solutions. In individual cases, customised AI solutions are also being tested. The very large online platforms in particular work with self-developed moderation architectures that are customised to their own needs. External service providers are connected to these systems.

[40] For services that do not distribute user-generated content at the core of their business model and are not subject to the DSA, however, external solutions are primarily used for content moderation.

However, with the constantly growing volume of user-generated media content, the **need for external services** in the field of automated recognition of images, video and audio content is growing. Recognising these is much more complex than simply analysing text and the analysis tools required for this cannot be developed by internal teams (alone), even for larger platforms. Due to the increased regulatory requirements (most recently NetzDG, currently DSA), the moderation management systems must also be adapted in order to expand the labelling and make it compliant with regulations. For smaller providers in particular, these adjustments to internally developed solutions tie up a proportionately large amount of capacity. In future, external technical systems are therefore likely to be increasingly integrated into the moderation processes, also in order to free up in-house development resources for the core tasks of service provision.

With regard to **manual moderation,** the majority of smaller services focussed on Germany rely on their own permanent moderation teams, which are supported by freelancers and assistants in Germany. Here, the moderators usually work closely with the community management or the trust and safety team. The large to very large online platforms, on the other hand, only use external service providers for manual moderation (see Chap. 3.4.2), which generally operate at different locations in different countries and work separately from the service's community management.

Overall, it is clear that all providers - from small to very large services - pursue a multi-layered approach by using both automated and manual content moderation processes. The spectrum ranges from providers that rely heavily on automated recognition processes and proactive moderation to providers that have so far primarily relied on self-moderation by the user community (community-led moderation).

There is currently no single, fundamentally superior moderation method or moderation tool. Each method has specific advantages and disadvantages with regard to parameters such as recognition speed, recognition rates, costs and personnel requirements. The respective mix of methods also depends on the subject area, time of day, current situation or target market, meaning that the methods can also be weighted differently within a service and the use of moderation resources is subject to continuous change.

However, moderation decisions, such as the removal of content or the exclusion of users, are decided by all providers in manual moderation processes.

3.3 Content moderation process

Content moderation is either automated by machine algorithms or manually by human moderators. Automated moderation offers scalability and speed, which is why it is primarily used proactively. Manual moderation is better able to capture complex nuances. In practice, a combination of both approaches is used in most cases to optimise efficiency and accuracy.

3.3.1 Automated procedures

Services use various automated processes to ensure the quality, security and integrity of online content.

Control-based balancing and filter systems

Rule-based matching and filtering systems identify, flag, block or remove unwanted, harmful or unacceptable content before/while it is uploaded to a platform and, in the event of an offence, block it directly (upload filter) or flag the content for manual review.[41] They use predefined lists, rules, pattern recognition processes and algorithms to enable targeted monitoring and categorisation of content.

- **Word filters:** Word filters are one of the simplest methods of content moderation. They can be set up manually on any moderation system. Word filters delete, replace and flag words and entire expressions that violate community guidelines.[42] Pre-sets of word filters can be purchased or licenced.
- **Automated Content Recognition (ACR):** ACR technologies automatically analyse multimedia content such as images, videos[43] and audio. Unique features or patterns are recognised, which are then compared with a database of known problematic content. For example, works protected by copyright or unwanted material can be recognised.
- **Content hashing:** Content is converted into unique cryptographic hash values. These hash values serve as a digital fingerprint and enable a very quick comparison ("matching") with a database to recognise identical or similar content. [44]
- **Content digital fingerprinting:** Similar to content hashing, a digital fingerprint of content is created here, but often on the basis of more complex algorithms. This means that edits or changes to content can also be recognised.[45]

Compared to other content moderation systems, such as manual moderation processes or predictive, AI-supported systems with machine learning, filter systems are characterised by their automation, speed and efficiency.

41 See Policy Department for Citizens' Rights and Constitutional Affairs (2020): "The impact of algorithms for online content filtering or moderation", Chapter 3, p. 35, online at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf, accessed on 09.10.23

42 Cf. https://en.wikipedia.org/wiki/Wordfilter

43 In the case of video analysis, individual images of a video feed are usually captured and treated as images.

44 See Policy Department for Citizens' Rights and Constitutional Affairs (2020): "The impact of algorithms for online content filtering or moderation", Chapter 3, p. 35, online at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf, accessed on 09.10.23

45 Cf. ibid.

Tab. 8 Examples of equalisation and filter systems

Name	Field of application	Developer	Users
Audible Magic Identification	Copyright Audio Recognition	Audible Magic	Soundcloud, Sony, Disney etc.
Content ID	Copyright content identification	YouTube	YouTube
Content Levels	Blocking of adult content	TikTok	TikTok
PowerTrack API	Filter tweets	Twitter/X	Twitter/X
Twitch Audio Recognition	Copyright Audio Recognition	Twitch	Twitch
PhotoDNA	Preventing and addressing child victimisation, including abduction, abuse and exploitation	Microsoft	National Centre for Missing & Exploited Children (NCMEC)

Source: Goldmedia Analysis 2023

Such filter systems, which compare source material with existing databases, generally work well in practice and represent an important component of content moderation on large online platforms. Although such filtering systems offer many advantages, they are also associated with certain challenges and limitations, primarily because they can only be applied to content that is already known. For example, word filters that filter out offensive terms are circumvented by adapting formulations and spellings just enough so that they are no longer recognised by the word filter. Users also circumvent the blocked words by changing spellings or working with more or less subtle allusions that require the reader to make certain transfers. This circumvention of filters is sometimes one of the reasons for the emergence of online jargon ("leetspeek") and has long been part of internet culture. The effectiveness of rigid filter systems is therefore heavily dependent on the filter lists being continually adapted and further developed, as users consciously test the limits of what can be written.

The following table provides an overview of the advantages and disadvantages of such filter systems for content moderation.

Tab. 9 Advantages and disadvantages of rule-based filter systems for content moderation

Advantages	Disadvantages
Speed The automated processes allow content to be checked in real time, enabling a rapid response to problematic content.	**Misinterpretations** Sometimes filter systems can misinterpret content and incorrectly block or remove legal or harmless content.
Scalability Filtering systems can handle large volumes of content, which is particularly important as an immense amount of content and media is uploaded to online platforms every day.	**Grey areas** Not all problematic content is clearly identifiable, and there are cases that require human judgement.
Continuous improvement Filter rules can be adjusted and improved manually at any time.	**Creative handling of content** People who want to spread harmful content can try to circumvent filter systems by modifying or disguising it.
Customisability Platforms can adapt the parameters of the filter systems to their own usage guidelines and needs.	**Lack of contextualisation** Filter systems often work on the basis of keywords, patterns or rules. As a result, they cannot always correctly recognise the context of a post or comment.

Source: Socialays (2023): "Pros and Cons of AI vs Manual Content Moderation" , online at: https://socialays.com/blog/pros-cons-ai-manual-content-moderation/, accessed 07.09.23

Self-learning filter systems - Artificial intelligence

In addition to rule-based filter systems, pre-trained or self-trained filter systems using artificial intelligence (AI) are increasingly being used. These are able to recognise the tonality of longer texts or references to critical topics or views, or to detect violations of community guidelines or illegal content in images or video contributions.

This makes them an important building block for scaling moderation performance and supporting and relieving manual moderation against the backdrop of a constantly growing volume of user-generated content. At the same time, the integration of AI can reduce the psychological strain on human moderators if incriminating or disturbing content is recognised by the AI and made unrecognisable to protect the moderators.

Two types of artificial intelligence are primarily used for content moderation:

- **Machine learning (ML)** is generally used to process large data sets in order to recognise patterns and correlations. Algorithms are used to learn from experience and make predictions or decisions. ML models can handle complex tasks by automatically identifying patterns in the data and drawing conclusions based on them.[46]
- **Natural Language Processing (NLP)** is concerned with converting written or spoken language into a form that can be understood by computers. This can be done using statistical or ML models. NLP enables complex texts to be analysed, interpreted and meaning to be extracted. NLP is used for translations, text processing, sentiment analyses and many other language and text-based tasks.[47]

The implementations of ML and NLP, which can be observed in the context of filter systems for content moderation, can be distinguished from each other in that pattern recognition using machine learning is primarily used for the identification of images or video content and NLP, with and without ML support, is explicitly used for the processing of language**.**

Many NLP-supported systems are already available **in the text area** that can recognise problematic content in context, beyond word filters, or even create sentiment analyses on the tonality of a text. Depending on the configuration, the creators can be informed that a text may be problematic while it is being written. These systems can also be found in moderation chatbots that can make and explain moderation decisions. These moderating chatbots are known as "AutoMods".

These AI systems are offered as part of a moderation system or as stand-alone modules that can be integrated into the moderation system via interfaces. In concrete terms, this means that user contributions are forwarded via the moderation system to the corresponding tool, which processes and checks the content and moderates it if necessary. Such systems are often capable of learning. The results of manual moderation processes are reported to the AI system. In this way, the AI modules used can be individually trained based on the data (user contributions and moderation decisions) of the respective customer platforms.

In the image sector, pattern recognition for pornographic or violent content, for example, has already been taking place on the major platforms for many years during upload.

In the video and audio sector, however, the detection of problematic or legally protected content is primarily based on hash databases (see YouTube Content ID or Twitch Audio Recognition in Table 8). In addition to a lack of platform-specific training sets, the challenge here is the insufficiently precise recognition of video and audio material, for example due to poor technical quality of the video content, overlapping audio sources or other technical reasons. Only a few external solution providers are already in a position to offer market-ready products.[48] It should also be noted that a complete content

[46] Cf. iodine (2020): "Machine Learning versus Natural Language Processing: What is the Difference?", online at: https://iodinesoftware.com/insights/blog-machine-learning-versus-natural-language-processing-what-is-the-difference/, accessed on 09.10.23

[47] Cf. ibid.

[48] These include the solutions from Amazon (see page 53), which are still quite expensive and do not have real-time capabilities.

analysis of video data requires a great deal of computing power due to the volume of data (this applies in particular to real-time analyses on streaming platforms) and is associated with high costs per analysed video contribution (see section 3.4.3, AWS Amazon Recognition). For this reason, such systems will probably continue to be used by platforms primarily on an ad hoc basis.

The translation and AI-based analysis of audio tracks on video clips or podcasts is also still under development. The first step here is to convert sound into text using tools such as Assembly AI or other text-to-speech applications. The next step is to analyse the text. Spotify acquired the audio content moderator Kinzen in 2022 after a long collaboration to control the many podcasts on its platform. [49]

Context-sensitive analysis of connected platform content

A key future development in the field of AI-supported systems is the context-sensitive analysis of live or video content, in which video streams are analysed together with the accompanying comments and other metadata in order to assign them a risk profile. A lot of content does not violate community guidelines per se, but only in the context of the intention of the statement. This has so far made automated early detection based on systems available on the market very difficult.

In the context-sensitive analyses currently under development, verbal and textual statements are included in the analysis together with the user data and metadata stored on the platform (e.g. their avatar, images including metadata, contact networks, connections to problematic communities). Both large social media platforms and service providers are developing such **predictive methods for early detection,** as these could make a significant difference in early detection, especially in the case of particularly drastic events (homicides, terrorist attacks, etc.), if they were ready for the market. On this basis, even supposedly unremarkable streams can achieve a high risk profile based on an AI prediction and thus be flagged much earlier and displayed to a human moderator for review.

[49] Cf. https://inside.com/podcasting/posts/spotify-acquired-content-moderation-company-kinzen-318967, retrieved on 09.10.23

3.3.2 Manual procedures

Manual procedures are all decision-making processes in which moderation decisions are made by one or more people. Depending on the service, these decision-making processes can be organised in a variety of ways and take place at different hierarchical levels. The main levels of manual moderation are described below; the implementation of the main levels described here can extend across different Group companies in the case of very large online platforms.

3.3.2.1 Community moderation by the users of a service

Community moderation by the users of a service ("community-led moderation") is one of the oldest approaches to content moderation. Even in early non-commercial and largely idealistically driven Usenet forums and bulletin boards, it was common for users to be hierarchised and have different permissions. "Forum admins" were usually given the task of moderating, deleting, moving content, etc. The corresponding authorisations could be acquired on the basis of criteria such as regular voluntary work, expertise and moderation skills.

This type of volunteer moderation is still used by numerous services today and is sometimes an integral part of a service's philosophy, particularly in the case of non-commercial and interest-driven services.

The more specifically a particular community is addressed by a service or part of a service, the more it makes sense from a commercial perspective to retain users (or particularly active parts of the user base) in the long term in a participatory manner and to win them over for the further development of their community. For example, users who have reliably reported violations over a longer period of time can be appointed as user moderators, possibly with access to the moderation system including authorisation to hide content.

Community moderation by the users of a service is not limited to non-commercial or niche offerings. Large online platforms, such as Twitch, also attach considerable importance to community moderation by their own users and incorporate the possibility of community moderation within certain limits.

For online platforms with highly live-driven content (e.g. live video streaming) and an active user community, community moderation is often the most effective tool for reacting quickly to unwanted content. Automated systems, especially AI-supported systems, do not yet work with sufficiently low latency to be able to monitor live content in real time.

However, pure community moderation by users themselves is not practicable for commercial platforms. And the subsequent removal/blocking of content is carried out by the moderators working on behalf of the service.

3.3.2.2 Self-moderation by content creator

Self-moderation by content creators typically exists on platforms where the relationship between content creators and content consumers is highly asymmetrical, such as live streaming platforms like YouTube or Twitch. The respective content creators (YouTubers, streamers or creators) can set their own moderation guidelines on their channels, which can be stricter than the community guidelines of the respective platform. In order to enforce these, the platforms provide content creators with special moderation tools. In addition to the option of manually deleting individual comments on their own posts, additional word filter systems are also offered. In some cases, content creators can also integrate moderation tools from freelance developers via the corresponding interfaces of the online platforms. In addition, members of their own community can also be given moderation privileges so that they can, for example, moderate chat and comment functions in real time while the content creator organises a live stream. Self-moderation by content creators does not replace the moderators working on behalf of the service, but complements them in certain environments.

The implementation of these moderation tools depends on the nature and focus of the platform: On Twitch, for example, moderators ensure that the chat follows the etiquette and content standards set by the streamer by removing offensive content and spam. On YouTube, moderators help to review and manage comments that people leave on a video or messages that participants send during the live chat of a stream. YouTube distinguishes between standard moderators and lead moderators, who have additional content moderation options. The channel operator determines the moderation privileges individually. The following table shows a selection of moderation tools from Twitch and YouTube.

Tab. 10 Selection of moderation tools on Twitch and YouTube

HD*	Instrument	Self-description of the instrument on the platform
Twitch	AutoMod	Automated method for identifying potentially risky chat messages
	Chat rules	Channel operators can create their own set of rules for their channel to inform new viewers about what behaviour is appropriate in the chat. Twitch can also use chatbots to inform users that their contribution has been blocked or that they have been banned.
	Moderator tools in the chat	Moderators defined by the streamer can view the chat and blocking history of chatters and leave and display comments on users.
	Block links	Setting that prevents links from being posted on the channel
	Delay in non-mod chat	Setting that causes messages to appear in the chat with a slight delay
	Email & SMS verification	Setting that prevents users from writing to the chat without a verified email address and mobile phone number
	Follower-only & subscriber-only mode	Options with which you can specify whether users must follow you or be subscribed so that they can write in the chat
	Blocked chatters	List of users who have been permanently blocked from chatting in the channel.
YouTube	Call up the channel	If you notice a message in the live chat, you can go directly to the channel of a live chat participant and find out more about them first

HD*	Instrument	Self-description of the instrument on the platform
	Remove content	You can remove any inappropriate or potentially abusive or offensive content. If you delete a message, it will be permanently removed from the live chat along with all replies.
	Temporarily block users	You can prevent someone from sending messages in the live chat for a period of five minutes.
	Hide users on this channel	The chat messages and comments of this person are then no longer visible to other viewers. The hidden user will not be notified of this.
	Check potentially inappropriate messages	You can show or hide comments or messages that have been withheld for review based on your community settings.
	Select community default settings	You can activate functions that use technology to automatically recognise spam, self-promotion, nonsensical and other inappropriate content in comments.
	Activate/deactivate live chat	You can activate or deactivate the live chat at any time, even after the event has started.
	Change participation mode	You can customise the participation mode in the live chat and only allow contributions by subscribers, contributions by members or live comments.
	Switch on message delay	You can restrict how often a user can send a chat message by setting a limit for the time between comments.
	Recognise blocked words	You can block messages in the live chat that contain certain terms or similar words.
	Hint	As a lead moderator, you do not have access to the Live Control Room or YouTube Studio. Lead moderators cannot appoint other moderators.

**HD: Hosting services*
Source: Twitch "Setting up moderation settings for your Twitch channel" (as of September 2023), online at: https://help.twitch.tv/s/article/setting-up-moderation-for-your-twitch-channel?language=de, and https://dev.twitch.tv/docs/irc/, accessed on 07.09.23
YouTube "Using moderation tools" (as of September 2023), online at: https://support.google.com/youtube/answer/10888907?hl=de&co=GENIE.Platform%3DAndroid, retrieved on 07.09.23

For example, Facebook offers its creators the option of specifying what kind of text, photo or video posts visitors can post on their own page. Posts from other people can always be allowed or deactivated. There is also the option to publish photo and video posts only after a manual review. Comments on posts on a page cannot be deactivated, but individual comments can be hidden or deleted. In addition to the automatic word filters, creators can also activate a word filter for "vulgar language".[50] Individual users can also be excluded from the site or ratings for the site in general can be deactivated.[51]

TikTok also offers special moderation tools for content creators and live channels, for which up to one hundred user moderators can be appointed. There is also the option for users to hide keywords from their own user experience or to personally block unwanted content or users via a button.

[50] Cf. https://www.facebook.com/help/1017549069082358, accessed on 22/09/23
[51] Cf. https://de-de.facebook.com/business/help/1323914937703529, accessed on 22/09/23

3.3.2.3 Manual moderation on behalf of a service

Manual moderation on behalf of a service is still the most important component within the moderation process. The majority of reports from automated procedures[52] and all reports received manually via the platform's reporting channels are checked and processed manually by human moderators. This can result in considerable personnel costs for large online platforms (see section 4.1), which are usually outsourced to service providers for manual moderation as part of business process outsourcing. [53]

For smaller providers, however, it is also common in Germany for manual moderation to be realised entirely within the company by its own employees.

Specialised outsourcing service providers provide the staff for manual moderation and operate the centres in which the moderators work, equipping them and offering psychological support for the employed moderators. Due to the sensitive working environment and the psychological stress that sometimes arises, content moderation predominantly takes place on the premises of a moderation service provider. Remote work is unusual for reasons of protecting the moderators and their families.

As specialised service providers in a process with a strong division of labour, service providers do not develop the guidelines on which the moderation is based, but only apply them. Despite their moderation experience, outsourcing partners are merely service providers for online platforms. Strategic advice on content moderation issues is not a primary area of business for service providers. However, manual moderation service providers in the market are endeavouring to expand their offering in order to become full-service providers for content moderation. As they have experience in dealing with many community guidelines and moderation processes and apply the same legal framework for several clients, they also act as a transfer centre for best practices.

The market for outsourcing service providers is characterised by the fact that the leading companies operate globally and run moderation centres in many countries on different continents. Providers spoken to for this study maintain up to 20 locations in all regions of the world and employ up to 80,000 people. Such globally operating service providers are preferably commissioned by very large online platforms, as these service providers can offer content moderation close to the language or cultural area of many regional markets from a single source. To spread the risk, however, large online platforms generally work with several moderation service providers (even within the same regional market or language area).

Content in German is also moderated by some outsourcing service providers in Germany, although moderation centres in other EU countries are also used due to the cost advantages, as long as sufficient German-speaking workers can be recruited on the local labour markets. The moderation of German-language content in the major moderation

[52] This does not include simple filter systems that make decisions based on binary logic, but rather more sophisticated, predictive filter systems that usually generate a need for manual moderation from a detection probability of 80 % to 85 %.

[53] Large or very large online platforms also often moderate a sample of their content in-house, although this should be considered primarily in the context of the platform's internal quality control and the further development of the community guidelines in the legal and policy team's range of tasks.

centres in Asia, Africa or India, on the other hand, is not common due to the lack of German language skills.

Service providers for manual moderation are used by all large and very large online platforms, regardless of how the platform positions itself on the spectrum between rigid and liberal content moderation.[54] This also includes hosting services that are not subject to the DSA or the TCO Regulation, such as messaging platforms.

For small platforms, there is hardly a sufficient need for moderation to justify outsourcing. The existing moderation teams of small platforms are too small for them to be outsourced with a significant reduction in costs (see section 4.3). In addition, the moderation processes must be sufficiently formalised to allow the manual moderation process to be carried out by a service provider.

According to the providers, a minimally staffed moderation team of a service provider that continuously moderates content requires a minimum team size of around 10-20 people who are deployed in a fixed team for a specific customer.[55] However, smaller services do not operate a continuous moderation effort. The advantages of outsourcing only outweigh the disadvantages when the moderation effort is well into double figures.

Future significance of manual moderation

The increasing use of the internet for video applications will also continue in the area of online platforms in the coming years. Social video platforms such as TikTok (with short vertical video clips/reels) or Twitch (with long live streams) are the drivers of this development. However, more established social networks such as Facebook, Instagram and YouTube are also increasingly focussing on the distribution of short video content for smartphone use (vertical video clips/reels).

This results in increasing demands on content moderation, as video and live content pose particular challenges. Proactive technical systems are already in use in the video sector. However, apart from the hash value-based detection of copyright infringements (music, images, video clips), only a few offences against community guidelines or laws (e.g. nudity) can be reliably detected by machine. Compared to the moderation of texts, where word filters and short text analyses take effect almost in real time, the moderation of video content remains more dependent on manual procedures.

Compared to uploaded video content, live video content poses significantly higher moderation requirements, as upload and distribution coincide and therefore no classic filter systems can be used.

To meet the challenges, YouTube is linking the mobile live streaming function to certain conditions. These include:

- Verified channel
- At least 50 subscribers (or 1,000 subscribers for young people)
- No channel restrictions for live streaming in the last 90 days

[54] One service provider for moderation services expressed this with "No provider has zero content moderation" and thus included providers who explicitly invite free expression of opinion on their platforms.

[55] For reasons of confidentiality and the necessary knowledge of the respective community guidelines, individual moderators are not deployed for different clients at the same time.

- Up to 24 hours waiting period for initial activation

Furthermore, when it comes to live content, all video platforms continue to focus primarily on involving their users in the moderation process (user messages/community moderation).

In addition, the constantly changing and newly emerging problematic (politically radical) networks and associated keywords and terminology pose a challenge for automated content moderation. Proactively monitoring and analysing the activities of radical networks in order to identify new types of challenges at an early stage is still primarily the domain of manual content moderation. Like law enforcement agencies, individual service providers have already specialised in monitoring areas of the internet that are difficult to access publicly (darknet) in order to detect changing communication patterns and new developments in niche communities in the area of terrorist threats at an early stage.[56]

Despite the developments, particularly in the area of AI-supported context-sensitive analyses (see section 3.3.1) and the predictive use of automated processes, it was pointed out several times in the expert interviews that this is a support for manual moderation. This means that manual moderation decisions remain predominant for more complex content. With the increased use of predictive procedures, which are intended to recognise events and situations in advance, a further increase in manual moderation work is therefore to be expected.

Ultimately, AI can improve the overall quality of predictions due to the processing of larger and more unstructured data volumes. However, it is a statistical model that makes predictions within the framework of given parameters and will not be able to make decisions autonomously for many moderation tasks in the future. Automated predictive detection of activities that are carried out from the outset with the intention of fraud by highly motivated offenders, such as scammers, haters or paedophiles, also remains difficult. Their behaviour includes the use of camouflage profiles and the use of age-specific vocabulary and diction, meaning that an NLP model can hardly be used effectively.

Automated systems are also being challenged by organised crime. There are anecdotal reports from service provider circles that even companies are working specifically to outwit AI-based moderation systems and are deploying considerable resources to develop "sophisticated fraud methods". AI-based detection systems also reveal their weaknesses when it comes to completely new fraud schemes: for example, the fraudulent scams in the medical mask trade that emerged during the coronavirus pandemic were not sufficiently recognised by automated systems as they were not adequately trained for this. In this respect, the AI solution providers themselves have also made it clear that AI cannot or will not solve all tasks in the moderation environment.

[56] To this end, specialists monitor relevant forums with specific threats, e.g. jihadist groups in the Middle East.

3.4 Service provider for content moderation

3.4.1 Market structure

There is a large market of providers for the moderation of user-generated content, whose services consist of different combinations of the following services:

a) **Moderation (management) systems** (SaaS platforms with stored escalation workflows, preparation of content for moderators, statistical recording, etc.)

b) **Technical systems for automated content recognition** (filter systems, AI-based moderation support, etc.)

c) **Manual moderation services** (moderation teams, moderation centre operations, staff support, training, etc.)

The business models of service providers differ depending on the service sector:

- Moderation systems - the software used by the moderators - are generally charged at a fixed price, although the price may include variable price components such as the number of connected moderators.
- Technical systems for content recognition are predominantly charged on an effort/volume basis, for example according to the number of API calls or the number of elements checked, whereby the price may include a fixed base component. [57]
- Moderation services are generally billed on the basis of the personnel deployed or on the basis of the agreed working hours. Performance or success-based remuneration models, such as those widely used for telecommunications services in call centres, are unusual in the area of content moderation.

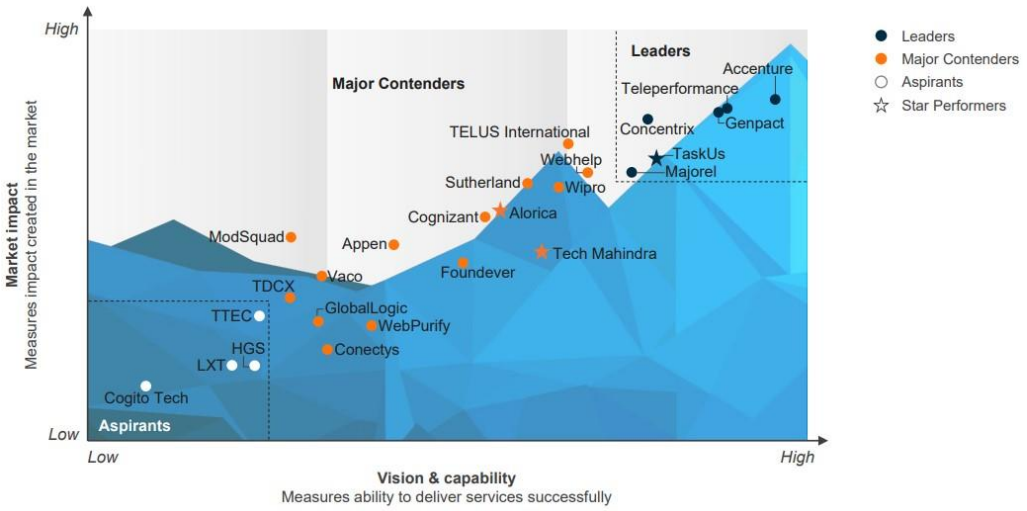
The market for content moderation can essentially be divided into the following segments:

1. **Specialised solution providers** in the field of **moderation management systems**
2. **Specialised solution providers** in the field of technical systems for **content recognition**
3. **Very large IT groups** that offer technical systems for content moderation as part of their cloud solutions and
4. **Service providers for moderation services** or full-service providers, many of which originate from the call centre-based outsourcing of telecommunications services (product consulting, customer support, etc.), have been offering manual content moderation for a long time and are now increasingly also offering technical content recognition systems (in cooperation or as an additional purchase).

The following figure provides an exemplary overview of the provider landscape in the US content moderation market:

[57] The definition of the elements can depend heavily on the area of application, e.g. a word, an image, a video, a text contribution, a post including accompanying content, etc.

Fig. 6 Trust and Safety Services Assessment 2023 for the US market



Source: TELUS International (2023): "Trust and Safety Services PEAK Matrix Assessment 2023 | TELUS International positioned as Major Contender", online at: https://assets.ctfassets.net/3viuren4us1n/56ae497eLEMIUYGWnvS26h/b1526f10dd3b79ec4283f19015ee8f60/Everest_Group_PEAK_Matrix_for_Trust_and_Safety_Services_Provider_2023_-_Focus_on_TELUS_international.pdf, accessed on 07.09.23

Providers in the area of moderation systems often already offer simple filter list systems as a licence component. More complex, in particular AI-supported systems for content recognition, both in the area of text and especially in the area of images and video, have so far been the domain of specialised providers.

As a result, there are numerous systems on the market, some of which are very specialised. At the same time, the use and connection of many different subsystems to their own moderation system is very challenging for online platform operators.

Larger providers in the field of moderation services, which primarily offer manual moderation and operate their own moderation centres, are striving to position themselves as fully-fledged full-service providers by purchasing or entering into strategic partnerships with specialised technical solution providers of AI-supported moderation tools. On the one hand, the aim is to achieve greater added value, including with major customers who currently pass on reported or self-detected violations to the service providers for manual moderation. In addition, the aim is to further reduce costs for customers in competition with other moderation providers by increasingly automating cases that are currently handled manually, thereby achieving differentiation. The company also wants to reach the market of smaller online platforms with primarily automated services.

3.4.2 Service provider

The following section presents a number of content moderation service providers that offer their services on the German market. This is an exemplary presentation of some services and their relevant services in the area of content moderation.

Tab. 11 Providers for exclusively automated moderation procedures and -solutions (as at: July 2023)

Product	Use of AI	Explanation
ActiveFence	Yes	Offer a generic API with filter systems, but also develop personalised solutions for their customers
Amazon Comprehend/ Recognition	Yes	Comprehend: Uses natural language processing to discover connections Recognition: Automates and rationalises image and video moderation workflows
Azure AI Content Safety	Yes	Classifies harmful content and gives it a risk-based rating
Azure AI Computer Vision	Yes	Uses visual data processing to label content
Bodyguard.ai	Yes	Pure software solution for moderation
Community Sift	Yes	Chat filter and content moderation system for social networks
CleanSpeak	Yes	Provides software solutions that protect customers from inappropriate content
Coral	Yes	Offers moderation of comments, especially in the area of online press
Disqus	Yes	End-to-end platform for the integration of users of the respective service. Offers moderation of comments and discussions
Ferret Go Conversario	Yes	Provides moderation AI models for dialogue managers and community builders
Ferret Go Engagently	Yes	Social platform with solutions for comments and community management
Hive Moderation	Yes	Offers various automated content moderation solutions for text, image and video content; moderation dashboard; recognition of AI-generated content
Moderate content	Yes	Programming interface for moderating image content in real time
Jigsaw Perspective	Yes	Is able to use natural language processing to mimic human understanding of words
Respondology	Yes	Management and customisation of moderation on clients' social media platforms
Thorn Safer	Yes	Provides solutions for platforms to investigate, remove and report child sexual abuse material
Sightengine	Yes	Programming interface (API) for image and video content; anonymisation of image and video content
WebPurify	Yes	Offers various automated content moderation solutions for text, image and video content as well as for metaverse applications

Source: Goldmedia Analysis 2023

In background discussions, technical service providers have emphasised that different source languages are not a particular hurdle for content moderation. Most providers work with language models that can interpret over 100 languages. In particular, text content suspected of terrorism can be easily detected technically. In this respect, further technical service providers that are not yet active in Germany can be expected to enter the market in the future.

In addition to smaller, specialised providers of technical systems, the **very large IT groups Microsoft, Google and Amazon** now also offer automated moderation tools as an external service.[58] These are described in more detail in the following section. The advantage of these offerings is that they can be used "off-the-shelf" for a variety of languages and content types (text, image, video) and also offer transparent, scalable pricing models (billing according to data volume or items) without fixed costs. The disadvantage of these offers is that they can only be individualised to a limited extent (no specific training of AI models based on the corpora of own hosting content) or only solve specific moderation tasks without being a full-service solution. As a rule, the use of other cloud and hosting services from the respective very large IT group is also a prerequisite for the use of automated moderation processes. Additional work is still required for services in the context of the interface connection and in the area of manual moderation.

As a result, these tools have so far tended to be used by larger platforms such as Reddit or the New York Times[59] or by larger customers who already use other cloud services from the provider (e.g. AWS, Azure, etc.).

Recruitment agencies that moderate in German use moderation locations in Germany or in other EU countries such as Ireland or Malta, where moderators who speak German work. Machine translations are not used as the basis for human moderation decisions. According to the providers, the local (cultural) context is too decisive for this. This applies in particular to recognising hate speech and preventing terrorist content.[60] There is also a specialisation of individual moderators within the service providers according to subject area or legal field.

[58] Microsoft's product has not even been officially released yet, but is still in the "preview" phase (as of September 2023).

[59] Cf. https://www.perspectiveapi.com/case-studies, retrieved on 20.09.23

[60] Local context is primarily defined linguistically or via the topic area, as it is not possible to determine the exact country of origin of a statement on online platforms or would not be expedient due to the global nature of the large platforms. In the case of counter-terrorism, domain expertise, which generally includes a regional component, is more important than determining the exact country of origin of a statement.

Tab. 12 Providers of moderation processes and solutions that also operate moderation centres (as at: July 2023)

Provider	Moderation process	Notes
Besedo	AI and manual moderation	Offer an all-in-one solution (training of an algorithm and human moderation) or individual component solutions.
Genpact	Manual Moderation	Moderate content from over 30 countries.
Majorel	Manual Moderation	Working internationally from 20 locations with 80,000 employees and manual content moderation.
Pexly	Manual Moderation	Moderate content from over 45 international locations.
Telus International	AI and manual moderation	Offer an all-in-one solution (training of an algorithm and human moderation) or individual component solutions.
Webhelp	AI and manual moderation	Offer an all-in-one solution through automated AI and manual moderation.

Source: Goldmedia Analysis 2023

3.4.3 Automated moderation procedures of the very large IT groups

The automated moderation tools of the very large IT groups Microsoft (Azure), Alphabet (Jigsaw) and Amazon (AWS) are analysed in more detail below.

Microsoft: Azure AI Content Safety

Azure AI Content Safety[61] was introduced in 2023 and is currently only available in a limited "preview" phase (as of September 2023). Azure AI offers an AI-based, multi-modal solution for content moderation, primarily developed for the areas of gaming, social media, e-commerce, e-learning, media and advertising. Azure AI checks chats, images, user names and avatars in social networks, among other things. Azure AI Content Safety is used for ChatGPT, among other things.

Azure AI is particularly suitable for gaming environments, as live streams of multi-player games and associated chat histories can also be monitored by the system. The specialised service provider Two Hat Security, which already has many years of experience in content moderation on the Microsoft platform Xbox Live with its moderation solution Community Sift, was fully acquired by Microsoft in 2021.[62] The code base of Community Sift is also part of the new moderation tool Azure AI Content Safety.

Azure AI Content Safety uses natural language processing to semantically understand and check the meaning and context of language. This works independently of language. Azure AI Content Safety currently supports eight languages. Image recognition is based on Project Florence of the Microsoft AI Cognitive Services Initiative. [63]

The AI recognises harmful content such as hate, sexual content, self-harm and violence. Each category is assigned a severity level from 0 to 6. Based on this severity level, a company can review the flagged content, prioritise it and take appropriate action. This can be configured by the customer with a certain degree of freedom. Customers can integrate the selected APIs into their moderation management system and also use the Content Safety Studio to explore and test the functions of the service.

Azure AI Content Safety is charged variably depending on the amount of text and images to be moderated. The product is not yet available in Germany; in Switzerland, the costs for a small online platform for analysing user comments with an average length of 1,000 characters and a volume of 100,000 comments with the "Language Understanding (LUIS) Standard" product amount to EUR 138.60 per month. [64]

The recognition of images for moderation purposes (with the product "Content Moderator S0") costs around 92.40 euros for 100,000 images with Azure AI Content Safety (as of September 2023).[65]

[61] Cf. https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety, accessed on 22/09/23
[62] Cf. https://www.crunchbase.com/organization/community-sift, retrieved on 21/09/23
[63] Cf. https://www.microsoft.com/en-us/research/project/projectflorence/, accessed online on 18/09/23
[64] See https://azure.microsoft.com/en-us/pricing/details/cognitive-services/, accessed online on 20.09.23
[65] See https://azure.microsoft.com/en-us/pricing/details/cognitive-services/, accessed online on 20.09.23

Alphabet: Perspective by Jigsaw

Perspective is a project of Jigsaw, an altruistic subsidiary of Alphabet that focuses on developing technologies to combat online abuse and online disinformation.

Perspective is a moderation tool that uses machine learning and artificial intelligence (AI) to assess the tone and quality of comments in online discussions. Perspective analyses text comments and gives them a "toxicity score", which indicates how likely a comment is to be offensive or inappropriate.[66] It was developed in 2017 based on the requirements of the Google "Counter Abuse Technology Team"[67] and was initially trained and used at the New York Times. The tool is also used by other publishers, but is also part of the community management platform Coral and Disqus[68] (cf. Tab. 11), meaning that Perspective is also used in many other online communities. In Germany, for example, the magazine Der Spiegel uses the Coral community management platform.[69]

The idea behind Perspective is to make online communication safer and more respectful by providing platforms and users with tools to improve the tone of discussions and curb the abuse and spread of hate speech. For example, Perspective enables real-time feedback to posting users. Some platforms therefore use Perspective to provide users with immediate feedback on the potential toxicity of their posts. An additional feature of Perspective API is the individual control over the experienced toxicity from the user's perspective. The API therefore also provides a tool to determine the level of toxicity of comments one wishes to encounter online.

The tool runs on Jigsaw's servers and is free to use for the general public. Services can integrate Perspective's API to implement automated moderation or to help human moderators review comments more efficiently without having to pay for it. There are currently no plans to commercialise the service.

Amazon: Amazon Recognition and Amazon Comprehend

Amazon Comprehend is also a moderation tool that uses machine learning (NLP). It offers an extensive range of features that include various analysis functions, including customised classification, key extraction, sentiment analysis, event recognition and entity recognition. The service is offered for a variety of source languages. The use of AWS servers is a prerequisite for using the service.

Comprehend's recognition function makes it possible to extract event structures from unstructured data. This makes it possible to filter relevant information from large texts and prepare it for further use in AI applications.

The service recognises and names entities such as people, places, companies and much more that appear in the given text. This automated identification of entities enables efficient classification of texts.

Comprehend's Targeted Sentiment Analysis also provides detailed insights into the sentiment in texts. It not only recognises rough sentiment indicators such as positive, neutral or negative, but also precisely identifies the sentiment in relation to specific entities in the text.

---

66 Cf. https://www.perspectiveapi.com/#/home, accessed on 22/09/23

67 Cf. https://jigsaw.google.com/the-current/toxicity/countermeasures/, retrieved on 20.09.23

68 Cf. https://www.perspectiveapi.com/case-studies/, retrieved on 20.09.23

69 Cf. ibid.

Amazon Recognition identifies images and videos with inappropriate or sensitive content, such as offensive images. The probability of each image or video being problematic content is calculated.

Another feature of Recognition is the ability to train and use your own moderation models. On this basis, companies can develop and implement specific models that are tailored to their individual requirements and community guidelines.

The price for using the service is variable and is calculated based on the number of characters (for text) or the number of images that are checked.

Analysing user comments with an average length of 1,000 characters and a volume of 100,000 comments per month costs around USD 100 per month with Amazon Comprehend. Recognising images (image label recognition and image properties) with Amazon Recognition costs around USD 175 for 100,000 images (as of September 2023). [70]

Content moderation for video content is charged on a per-minute basis. For automated content moderation with Amazon Recognition, USD 12,000 is charged for the moderation of 100,000 videos with an average length of 1 minute at the current minute price of USD 0.12/min.[71] (as of September 2023). [72]

3.4.4 Integration of technical solutions and service providers

In addition to internally developed moderation systems, purchased moderation systems are also customised to the respective online platform. System changes are therefore rather unusual. As a rule, the online platform and provider of the moderation solution work together for many years and jointly develop the systems used.

However, individual components, such as filter systems, can be integrated relatively easily into existing configurations. The technical connection of a third-party system to an existing moderation management platform poses no particular challenge, as these platforms are designed for modular expandability. The integration of an additional filter system via interfaces into an existing moderation solution means - depending on the system - a development or customisation effort of a few days to a few weeks. If a complex AI-based system is to be integrated, it can take several months to train the AI until optimal results are achieved. However, the basic availability of a system based on a generic training set should be available in just a few weeks.

In the case of personnel services, the technical integration effort can be accomplished just as quickly. However, the hiring and training of moderators can be significantly more time-consuming. Larger service providers may be able to dispense with new hires due to their staffing levels and cover the staff required for a new team with the help of existing personnel resources. However, it can be assumed that the training process will take at least 4-6 weeks.

---

70 See https://aws.amazon.com/de/rekognition/pricing/, accessed online on 21.09.23

71 Cf. https://aws.amazon.com/de/comprehend/pricing/, accessed online on 21/09/23

72 For the online platform YouTube, to which, according to the company, around 500 hours of videos are uploaded per minute, this meant a calculated cost for Amazon Recognition of USD 62.2 million per month.

4 The practice of content moderation in Germany

The following chapter describes the practice of content moderation in Germany in more detail. In doing so, we will first look at large social networks that have so far been subject to the NetzDG. The first step is to analyse the information in the NetzDG transparency reports. This is followed by a more detailed portrait of some selected large online platforms by analysing information from interviews with the providers in addition to the information from the public transparency reports.

Similarly, small online platforms and providers are portrayed in the following chapter. The presentation in this chapter is based to a greater extent on the information from the interviews, as these providers are not yet subject to any reporting obligations comparable to the NetzDG.

The following chapter briefly summarises the findings on the cost of content moderation, before the final sub-chapter evaluates the significance of terrorist content on online platforms using the available TCO-VO transparency reports from selected providers.

4.1 Large social networks

Since 2018, the Netzwerkdurchsetzungsgesetz (NetzDG) has applied to "social networks" that have more than two million registered users in Germany. The law will be replaced by the DSA in February 2024.

Service providers subject to the NetzDG are obliged to report on their content moderation measures every six months. These transparency reports provide an insight into the volume of moderation and the moderation processes of large online platforms ("social media") operating in Germany. However, the transparency reports only cover the area of content moderation that can be prosecuted under criminal law, which is explicitly regulated by the NetzDG. Other large areas of content moderation by online platforms, such as in the area of fraud prevention and protection, are not covered by the transparency obligations of the NetzDG.

In order to comply with the legal requirements of the NetzDG to provide an effective and transparent procedure for dealing with complaints about illegal content[73], the major online platforms have set up separate moderation processes (cues) that are specially trained to deal with NetzDG enquiries. Users of a service can usually report a criminally relevant post directly to a NetzDG team, often bypassing the regular moderation process (in accordance with the provider's community guidelines). As a rule, users decide at the time of reporting which process should be used for moderation, usually in accordance with community guidelines or the NetzDG.

[73] Cf. section 3 (1) NetzDG

For the provider, this distinction is generally immaterial, as the community standards are for the most part stricter than the legal standards of criminal law.[74] This results in a number of methodological challenges when interpreting the data reported by the online platforms. With regard to the information in the transparency reports, it cannot be generally assumed that all content moderation that is relevant to the NetzDG is recorded, but only whether it was moderated by the NetzDG teams.[75] The information in the transparency reports should therefore be interpreted with caution, as they generally do not allow any conclusions to be drawn about the entire moderation process on the respective platform.

Despite these limitations for the quantitative data within the reports, the transparency reports provide important qualitative information on the platforms' moderation processes. Selected information from the NetzDG transparency reports of leading online platforms is summarised in tabular form below.

[74] Only in some areas are the requirements of German criminal law stricter than the community guidelines of large internationally operating social networks, such as the display of symbols of unconstitutional organisations.

[75] The processes differ from provider to provider, so there are certainly providers who have German content checked by two different teams (Community Guidelines and NetzDG), but this is not the rule.

Tab. 13 Key statements in the NetzDG reports of selected online platforms, July-December 2022

Platform	Users daily in Germany *	No. of reports according to NetzDG	Automated detection	Organisation	Personnel Equipment
Facebook	14.11 million	- 34,806 removed or blocked content - of which 33,700 violate the Community standards - of which 1,106 offences against the NetzDG	- Rate limits (to prevent bots) - Matching (content hashing) - Machine learning: automatic removal if AI is sufficiently safe	Review takes place in two stages: trained teams and lawyers	178 people in three teams for processing NetzDG complaints
YouTube	25.4 million **	- 233,440 reports according to NetzDG - of which 5 ,166 reports were unlawful - thereof 109 terror distances	Automated machine matching: use of hashes (digital fingerprints) Automated machine messages to submit content for human manual review	NetzDG team in two shifts, around the clock. The two shifts consist of clerks, senior content reviewers, the legal department of YT and Google as well as external law firms with criminal law expertise.	Special team for NetzDG complaints, external service provider in Germany employs 77 people
Twitter ***	2.82 million	- Number of NetzDG complaints received: 947,994 Number of NetzDG complaints with measures: 153,416	Use of heuristics (keyword patterns) and machine learning methods to react to new forms of policy violations	Special team for NetzDG reports downstream: First check for policy violations, only then for NetzDG violations	Network DG team: 150 people, 7% employed directly by Twitter, all others by contractual partners

Platform	Users daily in Germany *	No. of reports according to NetzDG	Automated detection	Organisation	Personnel Equipment
Instagram	14.81 million	- 4,273 removed contents - of which 4,155 violations of the Community directives, - 118 Violations of the NetzDG	- Rate limits (to prevent bots) - Matching (content hashing) - Machine learning: automatic removal if AI is sufficiently safe	Review takes place in two stages: trained teams and lawyers	178 people in three teams for processing NetzDG complaints
reddit	0.71 million	- 1,066 NetzDG complaints - Of these, 674 resulted in removal or blocking	- Use of automated tools on the platform for violations of content guidelines - No automated tools for NetzDG reports	Safety team for general content offences Community team for moderation offences Platform & Legal Policy team for the removal of illegal content	Platform & Legal Policy Team consists of 12 specialists, 4 of whom are specialised in handling NetzDG complaints
- **Twitch**	1.41 million	- A total of 37,607 reports - of which 929 reports according to NetzDG	- Use of machine learning models to combat offensive usernames, spam, fraud, offensive emotes and bot accounts emotes and bot accounts	- First check for policy violations, only then for NetzDG violations, illegal content is blocked within 24 hours	- At least 15 moderators at all times for NetzDG reports. If a more detailed review is required: escalation to internal team of specialists. In complex cases also to lawyers
Soundcloud	2.12 million per week	- A total of 114 NetzDG notifications - 47 other, not explicitly NetzDG-related messages	Do not use AI	Processing of NetzDG complaints via internal and external Trust and Safety Team	Internal Trust & Safety Team: 12 employees External Trust & Safety Team: 6 people, plus legal department for support

Platform	Nutzer daily in Germany *	No. of reports according to NetzDG	Automated detection	Organisation	Personnel Equipment
TikTok	5.64 million	- 226,479 NetzDG complaints - 24,534 Removals or blockings according to NetzDG - 20,051 Removals or closures under the Community Directives	Uploaded videos undergo automated checking for recognition and classification (e.g. terrorist symbols), automated removal or labelling of videos for manual moderation if necessary	Development of moderation guidelines by trust and safety team (SIC!); review of content by moderation teams	Special NetzDG team with 28 members, 11 of whom are employed by an external service provider
Pinterest	4.94 million	- 55 NetzDG complaints - 5,406 in connection with a breach of the Community Directives	- Automated tools flag content that violates the Community guidelines - Machine learning models that evaluate images and can already make enforcement decisions	Processing of complaints from NetzDG team; processing of complex cases by Trust & Safety Leads and legal department as well as external German legal advisors	6 people assess the complaints; 4 employees are responsible for moderation; Trust & Safety comprises 2 employees

** Goldmedia calculation based on: ARD-ZDF online study 2022. ** Goldmedia calculation based on: die medienanstalten 2022. intermediaries and opinion formation 2022-I, *** Data from Twitter before the takeover by Elon Musk*

Source: Goldmedia analysis 2023 based on the Federal Gazette

The table reveals some very clear differences in the staffing levels of the online platforms with regard to NetzDG moderation when compared to the size of the platform and the volume of reports. However, industry experts, service providers and moderation service providers were unable to confirm this observation in background discussions to the extent suggested by the information in the transparency reports. Although the moderation effort differs due to platform-specific factors such as orientation, target group and sensitivity of the content[76], the human moderation effort is essentially comparable, at least when platform-specific differences are taken into account. The major deviations suggested by the transparency reports according to the NetzDG appear to be primarily due to different collection metrics and definitional imprecision (see below), as the Community Guidelines generally have very large overlaps with the NetzDG. The extent of the

[76] Certain subject areas, such as politics or individual computer game titles, are generally considered to be moderation-intensive, as there are frequent reports in these areas, including in the area of hateful speech.

NetzDG-relevant moderation effort within the moderation teams that only moderate according to the community guidelines is difficult or almost impossible for many providers to determine. The following table shows the volume of NetzDG moderation, particularly in relation to terrorist content on selected platforms, in more detail.

Tab. 14 Key figures from the NetzDG reports of selected online platforms, July-December 2022

Period	07.-12.22	07.-12.22	07.-12.22	07.-12.22	01.-06.22
Platform	**Facebook**	**Instagram**	**YouTube**	**Twitter/ X**	**Twitch**
NetzDG complaints	**125.195**	**57.541**	**233.440**	**947.994**	**44.299**
of which "terrorism"* in %	21,1%	23,8%	k. A.	9,8%	k. A.
of which "terrorism and unconstitutional"** in %*	31,9%	35,4%	6,8%	15,2%	8,6%
Removed contents	**34.806**	**4.273**	**32.150**	**153.416**	**2.894**
of which "terrorism" in %	6,7%	6,0%	k. A.	1,9%	k. A.
of which "terrorism and unconstitutional" in %	28,0%	34,4%	6,3%	19,8%	7,7%
Distance <24h	**92,9%**	**93,2%**	**85,6%**	**97,5%***	**98,6%**
of which terrorism in %	k. A.		87%	k. A.	k. A.
Distance >7 days	**0,3%**	**0,6%**	**0,4%**	**0,01%**	**0,1%**
of which terrorism* in %	k. A.		0,2%	k. A	k. A
Justified objections	3.173	883	k. A.	k. A.	68
Proportion of authorised Contradictions of removed content in %	**9,1%**	**20,7%**	**k. A.**	**k. A.**	**2,3%**
Personnel NetzDG moderation according to reports	178		77[77]	150	45
Jobs (full-time, Goldmedia estimate)****	**118,7**		**51,3**	**100,0**	**30,0**
Complaints per full-time position in 6 months	1.540		4.548	9.480	1.477

** Terrorism" includes sections 86, 86a, 89a, 91, 100a, 129a and 129b StGB*

***"Terrorism and unconstitutional content" also includes sections 129, 140 and 269 StGB*

**** The figure is based on the processing of NetzDG complaints and not on the processing of content removed in accordance with the NetzDG.*

***** It is generally assumed that the information on moderation teams is made up of 50 per cent full-time employees and 50 per cent part-time employees with half a full-time position.*

Source: Goldmedia Analysis 2023

A fundamental challenge in interpreting the information lies in the fact that online platforms do not have a standardised approach to content that is not permitted under both their own Community Guidelines and the NetzDG. Some providers generally check content according to both standards, while other providers refrain from checking content according to the other standard if it is not permitted.

[77] According to press reports, YouTube uses 16,974 moderators across the EU, see https://ch.marketscreener.com/kurs/aktie/ALPHABET-INC-24203373/news/Musks-X-hat-nur-einen-Bruchteil-der-Moderatoren-der-Konkurrenz-sagt-die-EU-45299109, accessed online on 13/11/2023

The data in the transparency reports of various platforms relating to terrorist content are also only comparable to a limited extent. Although platforms such as Facebook or Instagram list the relevant offences of the Criminal Code per complaint, there is an unknown amount of double reporting. The percentage figures at complaint level are therefore likely to be heavily distorted upwards. Platforms such as YouTube and Twitch group the complaints into larger subject areas - however, the chosen standard of "terrorism and unconstitutional content" is significantly more extensive than the definition of terrorism according to the TCO Regulation, meaning that these figures are also heavily distorted upwards.

The data on removed content is therefore more meaningful. For the online platforms analysed, the proportion of terrorist or terrorist or unconstitutional content is between 2 percent and 7 percent of the content removed on the basis of the NetzDG. The time from notification to removal is between 85.6 per cent and 98.6 per cent for the platforms analysed. The proportion of counter-proposals that were complied with after re-examination was between 2.3 per cent and 20.7 per cent for the online platforms analysed.

The information on moderation staff is very vaguely defined in the transparency reports and is sometimes described as "available"[78], which leaves considerable room for interpretation and is not limited to the staff directly involved in content moderation ("active"). In addition, the Meta Group, for example, makes identical statements about the personnel deployed for both Facebook and Instagram, which strongly suggests that the same personnel are deployed for both platforms.[79] The statements on the average moderation volume per full-time position are therefore only suitable to a limited extent for deriving a staffing ratio for content moderation.

4.1.1 Very large online platform in the area of social networks

Proactive, automated moderation procedures are basically an industry standard, as the very large online platform TikTok emphasises. All automated moderation procedures and tools are developed internally at TikTok.

The use of AI in moderation is generally seen as a very interesting development that is rapidly evolving within the industry, with no end in sight. Fields of application lie both in the enforcement of community guidelines (content recognition), but also in user-orientated features to support the user experience.

The use of AI in automated audiovisual moderation is on the rise, but its implementation must be tested "extremely carefully". TikTok also relies on an advanced AI-supported process for recognising unwanted content, which is constantly being further trained based on the moderation decisions of human moderators. TikTok is also open to other new solutions and is currently implementing further NLP-based audiovisual recognition mechanisms. However, their results are monitored manually by their moderators.

The key question that TikTok is asking itself for the future is: "How much AI can we use safely"? In principle, the same standards apply to AI-supported processes as to manual

[78] Cf. Facebook: NetzDG Transparency Report January 2023, Section 5B

[79] However, no reference is made to this in the transparency reports of the two platforms. In Tab. 14 this is assumed and taken into account accordingly when calculating the quotient in the last row.

moderation processes. In general, automated AI-supported processes should be developed to such an extent that they can guarantee the same or an even higher level of security.

Manual moderation at TikTok usually takes place at various locations in the industry, including Dublin, London and Germany in addition to the USA. This is also because local nuances play a key role in recognising hate speech and terrorism. Moderators are specially trained in the specific cues to recognise such offences and threats.

The strategic aim of using the AI-supported process is to identify cases in order to support manual moderation. Machine-identified cases are always moderated manually in accordance with company guidelines. Human moderation will still be required for sensitive moderation tasks in particular, such as terrorism, in order to ensure the safety of users. Cooperation with industry reporting databases (see section 8.2), such as Tech Against Terrorism, is essential for sensitive moderation areas such as terrorism and child abuse in order to benefit from the experience of other online platforms.

Finally, TikTok emphasises the importance of transparency towards users regarding what content is undesirable and how content is moderated. Information about this (community guidelines, data protection settings, options for content creators to moderate content themselves, rules for cooperation with law enforcement authorities) should therefore be kept centrally in a "security centre" and be easy for users to find. A transparency report on the TCO Regulation is also published there ("EU Terrorist Content Online Regulation (EU) 2021/784 Transparency Report").[80] The following table shows key information from this report:

Tab. 15 Specific indicators for the prevention of terrorist content on TikTok

Parameter	Specification	Timing
Distances (total)	53.385	07.06.22- 31.12.22
Objections from users (% of total)	22,1 %	07.06.22- 31.12.22
Successful appeals where content was restored (% of total)	11,5 %	07.06.22- 31.12.22

Source: TikTok 2023: EU Terrorist Content Online Regulation (EU) 2021/784 Transparency Report

During the reporting period, TikTok deactivated, among others, a network of 6 user accounts and 368,644 followers that operated out of Germany and used inauthentic user accounts in an attempt to influence the discourse around the Egyptian government and was intended for a purely Egyptian audience. TikTok has not received any removal orders from the European authorities under the TCO Regulation in the period 07/06/22-31/12/22.[81]

The following table shows general key figures for content moderation at TikTok:

Tab. 16 General key figures for content moderation on TikTok

Parameter	Specification	Timing	Remarks

[80] Cf. https://www.tiktok.com/transparency/en/tco-report/, accessed on 22/09/23

[81] Cf. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-1/, retrieved on 14.09.23

Unique users per month in Germany	20.6 million monthly 5.64 million daily	2023 2022	Estimate[82] cf. Tab. 13
Average user contributions per month (total)	5.056 billion	1st quarter 2023	TikTok Community Guidelines Enforcement Report
Average number of user contributions removed (to total)	0.6 percentage points	1st quarter 2023	TikTok Community Guidelines Enforcement Report
Average user contributions removed in Germany (to total)	0.0066 percentage points	1st quarter 2023	TikTok Community Guidelines Enforcement Report
Average number of user posts removed in the area of violent extremism (to total)	0.0084 percentage points	1st quarter 2023	TikTok Community Guidelines Enforcement Report
Proactively removed	94,9 %	1st quarter 2023	TikTok Community Guidelines Enforcement Report
away from it before distribution	77,4 %	1st quarter 2023	TikTok Community Guidelines Enforcement Report
removed in less than 24 hours	85,9 %	1st quarter 2023	TikTok Community Guidelines Enforcement Report
Moderation decisions successfully objected to by users	6,8 %	1st quarter 2023	TikTok Community Guidelines Enforcement Report
Total employees	450 people (Germany only)	30.08. 2022	Verdi[83]
Employed moderators of the service	28 persons (NetzDG only)	2nd half of 2022	TikTok NetzDG transparency report
Moderators for Germany (to total moderators)	2,3 %	1st quarter 2023	TikTok Community Guidelines Enforcement Report

Source: Goldmedia Analysis 2023

4.1.2 Large online platform in the social video sector

With between 30 and 32 million active users worldwide, the social video platform Twitch is not one of the very large online platforms. Of the active users, around 7 to 8 million actively create content and distribute it via the service.

The social video platform uses both external and internally developed automated processes. For manual moderation, the service works together with an external service provider, as is customary in the industry.

[82] Cf. https://www.smart-home-fox.de/tiktok-nutzer-statistiken, retrieved on 14.09.23

[83] Vgl. https://www.verdi.de/themen/arbeit/++co++6fe4812a-23a2-11ed-87a9-001a4a160129#:~:text=TikTok%20geh%C3%B6rt%20zum%20chinesischen%20Mutterkonzern,%C3%BCber%20450%20Mitarbeiterinnen%20und%20Mitarbeiter.

The specific challenge for the social video platform is that it mainly offers live content. This means that all content that is uploaded to the platform is produced at the same moment it is distributed. The majority of content does not remain on the platform, but is only accessible for a certain period of time (for example, 48 hours). Only a small proportion of the content on the platform remains permanently available.

The current trend towards increased automation is also strongly felt by this service. Service providers have great ambitions and many systems are being developed, including within the service. For example, the machine learning application "Ally" from Spirit AI is used internally for its own moderation solutions.

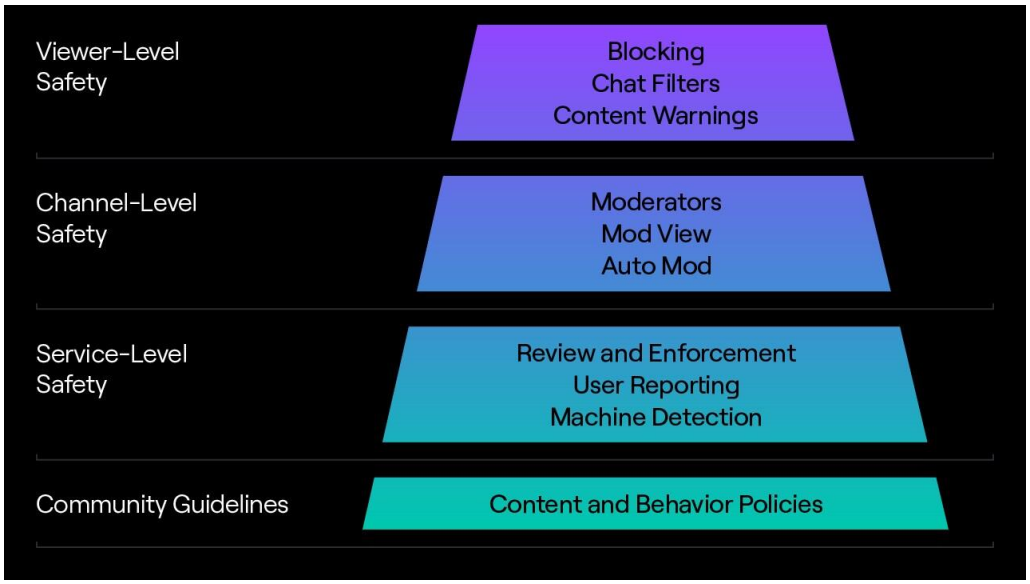
The results of the external service providers tested so far are still somewhat sobering. The product promises currently still exceed the actual performance. Recognition is often still not context-based enough to be able to moderate the specific content of the service. This means that reliability is not yet guaranteed to the required extent. AI-supported processes are not a miracle weapon.

Implementing automated processes for live content is much more difficult, which is why the security architecture differs greatly from other providers, where publication only takes place after the upload. The strategic focus is on community-led moderation (see Chap. 3.3.2.1). Tools (sometimes AI-supported) are developed and made available for this purpose in particular.

NLP-based systems can cope well with special application scenarios, such as chat moderation of particularly shrill content (e.g. insults and other "static" terms). The established hash matching tools also deliver good results.

In conclusion, the provider emphasises that security is a multi-layered concept and that the search for a single solution to the diverse challenges of content moderation is not sensible. People within the moderation process will remain the key to successful content moderation in the future.

Fig. 7 Multi-layered moderation concept of the social video platform



Source: Twitch, online at: https://safety.twitch.tv/s/article/Safety-at-Twitch?language=en_US, *accessed on 31.07.23*

The service publishes a transparency report on the TCO Regulation ("EU Terrorist Content Online Regulation 2022 Transparency Report").[84] During the reporting period, the service did not receive any removal orders from European authorities in accordance with the TCO Regulation.

The following table shows general key figures for content moderation of the service (cf. Tab. 17). Only moderation decisions (enforcement actions) by the moderators of the service are shown; moderation decisions within the framework of community-led moderation are not included here. Content removals only play a subordinate role on the platform, as the content is generally no longer distributed when a moderation decision is made[85], which is why the service does not report the number of removals, but rather the number of enforcement actions.

Tab. 17 General key figures for content moderation of the social video service

Parameter	Specification	Timing	Remarks
Unique users per month in Germany	3.58 million monthly 1.41 million daily	March 2020 2022	agof daily digital facts cf. Tab. 13
Average hours watched per month	1.847 billion	2nd half-year 2022	Transparency Report H2 2022
Average number of Enforcement Actions per month	194.000	2nd half-year 2022	Transparency Report H2 2022
Average enforcement actions per month in the area of terrorism	17	2nd HY 2022	Transparency Report H2 2022
Enforcement actions successfully objected to by users	852	2nd HY 2022	Transparency Report H2 2022

Source: Goldmedia Analysis 2023

4.1.3 Large online platform in the social gaming sector

With around 66 million daily active users worldwide (as at Q1 2023) and an average of 27.4 million monthly active users in the EU (as at August 2023)[86], the social gaming platform Roblox is not one of the very large online platforms.

On the online platform, users can create their own games in an intuitive development environment, discover games created by other users and share their gaming experiences. The online platform is particularly popular with younger players, as the games

[84] Cf. https://safety.twitch.tv/s/article/2022-EU-Terrorist-Content-Transparency-Report?language=en_US, accessed on 22/09/23

[85] Should this still be the case, the offending content will be removed by the moderators.

[86] Cf. https://en.help.roblox.com/hc/de/articles/13061336948244-Digital-Services-Act, accessed on 18/09/23

offer a strong social component where users can chat with others and play together. Almost 50 per cent of users are under 13 years old.

The service emphasises the importance of clear communication of community standards. This also includes communicating the potential consequences of posting undesirable content. In addition to warnings and the removal of such content, this also includes the permanent blocking of your own user account or a police report if there is an immediate threat.

The service uses both technological and manual processes (around the clock) to moderate content. Technically, content is first compared with known illegal content (e.g. terrorist content and material on the sexual abuse of children) using industry databases. The European Internet Forum (EUIF) and the Tech Against Terrorism initiative are particularly worthy of mention in this context.

In addition to these industry databases, content is checked against a separate database in which the content previously removed from the platform is stored as hash values. Due to the specific nature of the platform, chat filters play an important role; for example, the Community Sift solution from Two Hat Security is used[87] . However, NLP-based methods developed in-house are also increasingly being used.

The service states that up to 1,000 people manually moderate content on the platform based on the community guidelines. These are usually located at external service providers in the region of the country of origin, as the local context is crucial for manual moderation. Users deliberately change the spelling to such an extent that automated filter systems no longer work. Knowledge of local linguistic usage is therefore essential for the correct categorisation of such utterances ("leetspeek").

Among the content moderators, there is also a specialised (internal) team that deals exclusively with the prevention of terrorist content ("terrorism and violent extremism", TVE). The employees in the area of terrorism prevention have relevant professional experience, for example from working for intelligence services or the FBI, and manage terrorist topics, names, memes, iconography and more. Due to new threat situations that arise primarily in small splinter groups, it makes sense for the online platform to employ a unit specialising in terrorism that keeps a close eye on such dynamically developing situations and groups. These employees also develop the basic training in terrorism prevention for all moderators of the service.

The service distinguishes between security systems and signalling channels. The security systems include

- Automated and manual image verification
- Automated chat filters and rules
- Special chat restrictions for users under the age of 13
- Safety control centre for users and their parents
- User reports from the community.

The reporting channels include:

1. Monitoring (automated process)

[87] Cf. https://roblox.fandom.com/wiki/Moderation_system, retrieved on 20.09.23

2. User messages
3. Trusted flaggers

Reported content is checked manually by the moderators, regardless of the reporting channel. The majority of the moderation volume comes from user reports.

The service publishes a transparency report on the TCO Regulation ("TCO Annual Report") on its website.[88] During the reporting period, the service did not receive any removal orders from European authorities in accordance with the TCO Regulation. No special adjustments to its own infrastructure were necessary when the TCO Regulation came into force, as the current internal processes for the prevention of terrorism were already in place at that time.

Finally, in the interview conducted for this study, the service cautions that security is a journey, but not a destination that can be achieved. Particularly in the area of terrorism and violent extremism, the situation is characterised by new events and groups, or rebranding and new manifestations, meaning that security in this area remains a moving target.

4.2 Small online platforms and providers

The following shows how

a) small online platforms with fewer than two million registered users in Germany that are not subject to the NetzDG, and
b) small hosting services that are not considered online platforms, as the distribution of user content is only an insignificant and inseparable ancillary function of another service, as exemplified by the comment sections of online newspapers [89]

operate their content management.

The descriptions in this section are essentially based on the statements of providers of smaller online platforms with whom background discussions were held at management and moderator level for this study.

4.2.1 Provider from the games sector -News

The online service in the area of games news is operated by a publisher in the entertainment sector, which operates editorial information and cross-brand community offerings (text-based offerings such as forums, comment functions, etc.) under various brand presences and also displays its content on its own websites and via social media. Its services are financed by subscriptions and advertising. Some of the content is freely available. As a predominantly editorial service, it is not subject to the DSA. The games news portal is supplemented by a large community platform with many different forums on the subject of games.

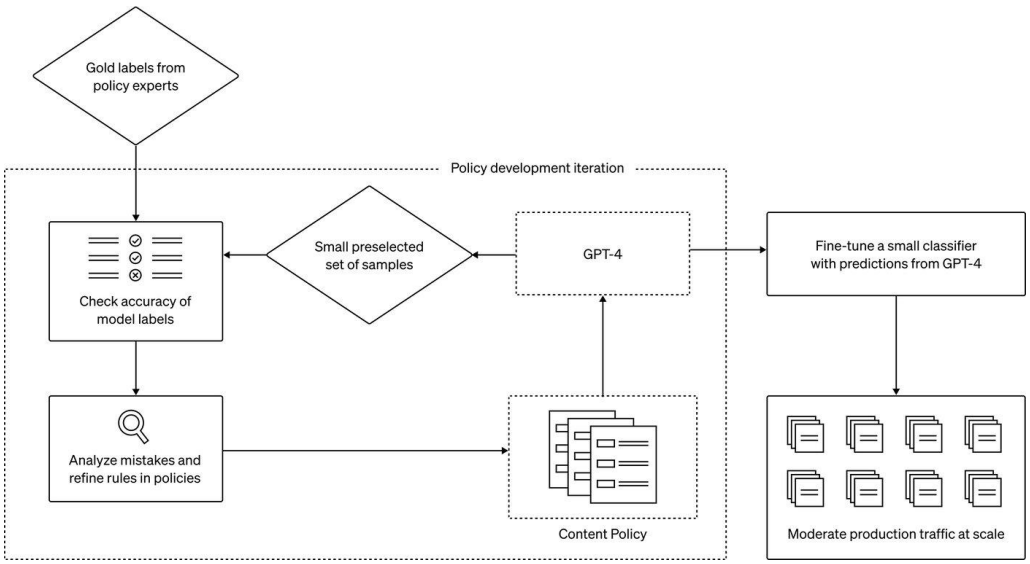
[88] Cf. https://corp.roblox.com/safety-civility-resources/?section=Tools&article=tco-annual-report
[89] Cf. recital 13 DSA

Own AI solution based on ChatGPT

To support the service's own moderators, an AI solution developed by the provider's internal development team has been tested since 2023, which is based on the Large Language Model ChatGPT from OpenAI in its current version GPT-4[90] . The core prompt here is that the model should behave like a forum moderator based on common community guidelines and flag the submitted content accordingly or also moderate it in automatic mode. No specific customisation to the company's own offerings was made.

Even though ChatGPT cannot be trained on the provider's data, it already delivers very satisfactory results for general text moderation tasks. OpenAI itself also uses ChatGPT to moderate the content generated and distributed by ChatGPT and to further develop its own content guidelines[91] .

Fig. 8 Process of using GPT-4 for moderating content and developing moderation guidelines



Source: OpenAI, online at: https://openai.com/blog/using-gpt-4-for-content-moderation, retrieved on 07.09.23

[90] This is due to the fact that even with a general large language model such as ChatGPT, (manual) moderation decisions are of paramount importance when training the AI model. For example, the same service providers that also moderate content for very large online platforms were sometimes used to train ChatGPT's AI. See https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai, accessed on 31.08.23

[91] Cf. https://openai.com/blog/using-gpt-4-for-content-moderation, retrieved on 31/08/23

Tab. 18 Key figures for content moderation in the games community

Parameter	Specification	Timing	Notes
Unique users per month	6.9 million	August-October 2022	AGOF, Company information
Visits per month	32.2 million*	April 2023	IVW
Average user contributions per month	1.03 million	June 2023	Expert discussion
Employed moderators of the service	4.5 (incl. freelancers)	June 2023	Expert discussion

** only main offer, other offers of the provider in the gaming sector not IVW-listed*
Source: Goldmedia Analysis 2023

The moderators work daily during normal office hours, while the AI solution moderates automatically at night.

In addition, the instrument of community moderation is used extensively to relieve the burden on the employed moderators. This involves around 100 active users of the service who have been appointed as user moderators by the service. The service's selection criteria for user moderators is a quota of at least 30 voluntarily reported contents per month with a reliability of correct reporting of at least 85 per cent. The user moderators are divided into three hierarchy levels, the highest of which also has removal rights. The quality of user moderation is monitored by the permanent moderation team.

4.2.2 Provider from the Q&A area platform

The service is a predominantly text-based question-and-answer platform that is not limited to any particular subject areas, but offers knowledge, experience and opinions on a wide range of topics and presents a broad diversity of opinions. The service is financed by advertising and through "Business Partners"[92] . As an online platform, the service is subject to the DSA.

Own AI solution

The service uses its own algorithm for proactive moderation, which was developed by the in-house data team. The decision to develop an in-house solution was made because the last tests (approx. 2019) of market-available solutions based on generic training sets were unsatisfactory. The algorithm was trained on the historical moderation decisions of the online platform. Particular attention was paid to filtering out hateful elements (hate speech, calls for violence), which should also reliably recognise terrorist content. The algorithm assigns a score for the probability of deletion between 0 and 1, and content with a score of 0.8 or higher is reviewed manually by moderators.

Role of the user moderators

To support the service's own moderators, the instrument of community moderation by the users of the service is used in particular. For this purpose, the service appoints certain users as user moderators by the community management if they fulfil certain criteria:

[92] You can communicate directly with business partners via the platform and ask them questions directly, similar to other company presences in social media.

- Proactive reporting of at least 30 contents per month
- Hit rate for problematic content in these self-initiated reports: at least 85 per cent

These user moderators in turn have their own hierarchy, which is characterised by different administrative rights:

- "Junior moderator"
- "Light moderator"
- "User moderator"

At the highest hierarchy level, the user moderators also have deletion rights and can remove content from the platform. The activities of the user moderators and compliance with the community guidelines are monitored by the service's community management. Community management is the level above the moderators, which also carries out moderator training. Community Management also deals with suspected false reports. There are currently around 100 such suspected false reports per day.

Tab. 19 Key figures for content moderation on the Q&A platform

Parameter	Specification	Timing	Remarks
Unique users per month	22.3 million	August 2023	Company information
Active users per month	1.85 million	August 2023	Company information
Average user contributions per month	1.98 million	July 2023	Expert discussion
Average number of user contributions removed	2,8 %	July 2023	Expert discussion
Moderation decisions objected to by users	5,6 %	July 2023	Expert discussion
Total employees	> 50	July 2023	Expert discussion
Employed moderators of the service	13 (incl. part-time employees)	July 2023	Expert discussion
Simultaneously active moderators	2-4, depending on revenue	July 2023	Expert discussion
Volunteer users moderators	72	April 2017	Press report
Employees in the community management of the service	9	July 2023	Expert discussion

Source: Goldmedia Analysis 2023

The moderators work from 8 a.m. to midnight every day. At night, the in-house AI solution moderates automatically and the user moderators manually. According to the provider, the interaction between technical moderation by the AI solution and user moderators works reliably: Even at night, reported content only remains online for a maximum of 3 to 6 minutes on average before it is deleted by a moderator.

4.2.3 Providers from the online news sector

The provider is a major daily newspaper media company that distributes news via print editions and online portals/apps in its core business and uses numerous additional social media channels. Its offerings are financed by subscriptions, individual print sales and advertising. Some of the content is freely available online. As a provider that primarily publishes news, it is not an online platform within the meaning of the DSA.

The online discussion forum shown here is specifically linked to an online daily newspaper offering and only allows text content. The discussion forum can only be used by subscribers to the newspaper brand. The further explanations refer exclusively to this brand presence and not to the other news offerings or undertakings of the media company. For the provider, the discussion forum is an important tool for acquiring subscribers and retaining customers. The publisher's user research shows that discussion participants build up a closer relationship with the product and subscribe to the service for longer than the average. Accordingly, the service will continue to be actively developed and expanded.

External service provider for automated moderation

The moderation tool used is the Engagently solution from the German service provider Ferret. According to the service provider itself, the moderation tool is used by many public and commercial German-language media providers to moderate discussions on social networks. The tool works on the basis of a predefined list of problematic key terms. The cost of using the tool for the provider is around 50,000 euros per year.

Manual moderation

The provider has an internal moderation team, which is primarily responsible for the manual moderation of content. In addition, there is a higher-level customer service team which, among other tasks, also serves as an escalation contact in the event of complaints about moderation decisions. There is also a social media team that prepares content for social networks and moderates the discussions there using the tools of the respective platforms. However, the activities of the social media team are not considered further below.

Manual moderation is carried out in-house by a team of 7 moderators every day of the year. On average, 2.8 moderators work in parallel, whereby the number is heavily dependent on the current news situation: For editorial reasons, the comments under some news items are only moderated manually, which greatly increases the manual moderation effort and is only feasible for a maximum of 2-3 news items per day.

Tab. 20 Key figures for content moderation in the community of an online newspaper

Parameter	Specification	Timing	Remarks
Active users per month	29.000	August 2023	Expert discussion
Average user contributions per month	617.000	August 2023	Expert discussion
Average number of user contributions removed	9,9 %	August 2023	Expert discussion
Employed moderators of the offer	7	August 2023	Expert discussion
Simultaneously active moderators	2,8	August 2023	Expert discussion
Employees in customer service of the offer	7	August 2023	Expert discussion

Source: Goldmedia Analysis 2023

The moderators work every day of the year. The automated moderation solution moderates at night. User comments therefore also go directly online at night as long as they have not been objected to by the automated preliminary check.

The volume of comments is constantly increasing, even though only subscribers to the service have access to this function. Due to the moderation-intensive environment, the provider does not expect to be able to significantly reduce the overall effort required for content moderation in future through the increased use of automated solutions. The background to this is that many articles work with allusions and references that involve the reader's transfer performance. This importance is not recognised by AI tools.

4.2.4 AI solution Zöe from Zeit Online

In 2016, Zeit Online developed an experimental artificial intelligence in-house to evaluate the extent to which it could support the Zeit Online moderation team. In a development period of seven days, an experimental prototype for a comment bot with a spam filter function was created by Zeit Online's "on-board mathematician", which only comprised 250 lines of code and already had a 75 per cent match rate with the human moderators.[93]

The development was carried out in the Python programming language on the basis of freely accessible software libraries. The Python library Keras was used, which is particularly suitable for neural networks. The AI is based on the Tensorflow software library from the Google Brain team, which is published under an open source licence.

According to the developer, the challenge during development was not the programming of the neural network, but its design: "What makes sense for text classification? How many layers do I stack on top of each other? How do I connect the neurons? What are the best types of layers to use?"[94]

[93] This project was documented in detail in a Zeit Online article, see Zeit Online "Mein Bot und ich", https://www.zeit.de/digital/2016-09/kuenstliche-intelligenz-kommentar-bot-zeit/komplettansicht, accessed on 20/09/2013

[94] Cf. ibid.

In 2018, when the volume of comments on Zeit Online totalled around 350,000 posts per month, the tool, which was initially developed experimentally in 2016, went live under the name "Zöe" and has been supporting the team of moderators ever since. [95]

4.2.5 Conclusion of the content moderation analysis for smaller platforms

In the following table, a personnel indicator for manual moderation is calculated on the basis of the three case studies presented.

Tab. 21 Personnel costs for manual moderation of smaller providers

	Reach per month (unique users)	Contributions per month	Moderators[96] (full-time equivalents)	Posts per moderator per month
Gaming provider	6.9 million*	1.03 million	3,38	305.000
Provider Q&A	1.85 million **	1.98 million	9,75	203.000
Provider News/Politics	29.000 **	617.000	5,25	118.000

** Total offer, ** Active users of the forum*
Source: Goldmedia 2023

The results show the differences in effort between the platforms, which result both from the moderation intensity of the topic environment and from the chosen moderation strategy.

The provider in the gaming sector operates in a rather moderation-intensive topic environment, albeit with a niche offering with a strong community component in which there are comparatively strong "self-regulatory" forces compared to large online platforms. One full-time moderator accounts for around 305,000 published posts per month.

The Q&A provider operates in an environment that is not overly moderation-intensive, but has few outwardly recognisable community elements, as the platform is designed more to provide a low-threshold offer for new active users without specific interests than to serve a "sworn" user community. The platform does not emphasise the community aspect as a particular incentive to use the platform and does not actively communicate its user moderators. One full-time moderator accounts for around 203,000 published posts per month.

The online news provider operates in a moderation-intensive environment. The platform also operates with a community concept, as only the community of subscribers has access to the service. However, the community is exclusively controlled hierarchically from above. The provider determines, for example, which news and reports the comment function is switched off or restricted for. This means that comments and some messages

[95] See Zeit Online "Wie wir Leserkommentare moderieren", https://blog.zeit.de/glashaus/2018/03/02/wie-wir-leserkommentare-moderieren/, accessed 20.09.23

[96] The information from the interviews reflects the headcount, which is made up of both permanent full-time employees and additional staff. In the following, it is generally assumed that the moderation teams are made up of 50 per cent full-time employees and 50 per cent part-time employees with half a full-time position.

are moderated exclusively by hand, which considerably increases the manual moderation effort. One full-time moderator is responsible for around 118,000 published posts per month.

The cost of technical moderation systems is lower than the personnel costs. In the case of the small provider that relies on a technical solution from a service provider, the personnel costs amount to around 80 per cent of the total costs for content moderation.

4.3 Terrorist content on online platforms

Since 2022, the importance of terrorist content in the area of content moderation can be seen not only in the general transparency reporting of the very large platforms, but also in the specific transparency reports on terrorist content that have been published by the hosting services since 2022 as a result of the transparency obligations under Art. 7 TCO Regulation. The following table shows the self-disclosures on terrorist content of selected very large online platforms

Tab. 22 Information on terrorist content in the EU based on TCO transparency reports, 2nd half of 2022

Article in TCO-VO		Face-book	Instagram	YouTube	Twitter/ X	TikTok
	Period	01.06.22-31.12.22	01.06.22-31.12.22	07.06.22-31.12.22	06.06.22-31.12.22	07.06.22-31.12.22
	Distances through own content mod.	4.1 million[97]	1.5 million	1.4 million[98]	k. A.	53.385
Art. 14(5)	Proactive reporting to law enforcement	k. A.	k. A.	k. A.	3	k. A.
Art. 7(3)(d)	Complaints	672.000	74.800	15.295	7.270	11.816
Art. 7(3)(g)	Justified complaints	90.800	11.900k	1.783	401	6.153
Art. 7(3)(g)	Proportion of justified objections to removals	2,2 %	0,8 %	0,1 %	k. A.	11,5 %
	Non-authorised removal order.	25	125	k. A.	k. A.	k. A.
Art. 7(3)(c)	Removal orders by qualified authorities	0	0	0	0	0
Art. 7(3)(e)	Official or judicial judicial review proceedings	0	0	0	0	0

Source: Goldmedia analysis according to TCO transparency reports of online platforms 2023

Of the very large online platforms analysed, the number of removals is highest for Facebook (4.1 million) and lowest for TikTok (53 thousand). The reason for the relatively large differences between the platforms is probably primarily due to a different definition of the content reported, as the definition of terrorism used by Facebook within the TCO-VO transparency report is relatively broad.

[97] Definition Facebook: Number corresponding to removals that violated Facebook's Community Guidelines on Dangerous Organisations and Individuals, Violence and Incitement, and Coordination of Harm and Promotion of Crime

[98] Definition YouTube: Number of items with terrorist content that have been removed due to violations of community guidelines or legal orders.

The number of justified complaints in the case of blocking due to terrorist content varies between 0.1 per cent (YouTube) and 11.5 per cent (TikTok) in the sample examined, whereby a rate of justified complaints in the double-digit percentage range is unusually high.

None of the very large online platforms analysed received a removal order from a qualified authority in the second half of 2022. In the discussions held with very large online platforms as part of the study, the moderation of terrorist content in general and the implementation of the TCO Regulation in particular were not considered to be particular challenges for the services. The basic infrastructure required for TCO-compliant behaviour is a fundamental component of the general content moderation infrastructure, even without a regulatory requirement.

Small services also shared this assessment across the board. The processes required for TCO-VO-compliant moderation are already part of the existing moderation processes. Added to this is the fact that terrorist content is generally a niche phenomenon for small services, as smaller services are not among the preferred distribution channels for terrorist content due to their limited reach. Removals due to terrorist content were described as isolated cases, or at least as occurring very rarely (less than 1 per cent of content).

4.4 Overall conclusion

The main findings on the status quo of content moderation, particularly in Germany, can be summarised as follows:

Community guidelines and transparency of moderation decisions

- The central lever for professional content moderation is the quality of the community guidelines. They must be concrete, operational and consistent. This determines the effectiveness of the other levels.
- The guidelines should be communicated via an easily accessible reporting platform (also for non-users).

Automated content moderation processes

- Automated moderation processes are used by all providers to support and prepare decisions. These include
  - Word filters (These are included in many external moderation systems. However, the lists must be expanded and maintained on a platform-specific basis).
  - Hash comparisons for image/sound (hash value platforms are generally available free of charge. However, interfaces must be programmed for use).
  - AI sentiment analyses (these have been available for text content from many external service providers at low cost since around 2017).
  - Automated full-service solutions are currently available from as little as €4,000/month.
- Automated solutions work very convincingly, especially in the area of text analysis.

- Solutions for analysing other media genres such as video or audio content are much more limited in terms of market availability and the results are less reliable. Live content (streaming) in particular still presents automated moderation processes with major moderation hurdles.
- Automated systems hardly support moderation-intensive content (certain political fields, gaming titles).
- Predictive, context-sensitive AI-supported processes are currently being developed, but are not yet being used on a large scale.

Manual moderation

- Manual moderation by human decision-makers remains the foundation of all content moderation.
- Manual moderation by personnel with sufficient language skills is of great importance for understanding the cultural context of utterances.
- Manual moderation, by employed moderators of the service, but also by content creators and users, is therefore even more important for video or live content than for the moderation of text-based content.
- The result of manual moderation depends on the quality of the community guidelines (see above), the moderation processes and the staffing.
- Guide value: Depending on the complexity, a full-time position for manual moderation must be calculated for 100,000 to 300,000 user contributions per month.
- The personnel costs for manual moderation can be mapped in relation to the volume of comments, even for the smallest providers.
- The moderation effort will increase in the medium term due to the increasing use of the services. Online platform operators are currently expanding their moderation teams or have recently expanded them. No external service provider is planning to reduce its teams or anticipates that improved automated processes could reduce staffing requirements. The opposite is more likely to be the case here: context-sensitive AI-based systems will further increase the amount of manual moderation.

Terrorist content and the TCO Regulation

- Notice-and-takedown procedures, in particular through trusted flaggers, are established throughout the industry and are very successful as a self-regulatory instrument. The procedures work quickly and effectively across borders/internationally.
- TCO removal orders are currently still the exception in the context of the notices that hosting services or online platforms receive from authorities.
- In the vast majority of cases, TCO specifications do not require any additional manual/automated monitoring and testing.
- The time limit for responding to a removal order within one hour in accordance with Art. 3 (3) TCO Regulation can be met by automated deletion and subsequent review.
- In particular, providers whose moderation process has weaknesses in key areas will also have problems complying with the TCO Regulation.

5 Minimum standards for content moderation

5.1 Derivation of abstract minimum standards

Hosting services that are "exposed" to terrorist content in accordance with Art. 5 Para. 4 TCO Regulation must take "specific measures" to curb this content in accordance with Art. 5 Para. 2 TCO Regulation. It is the task of the BNetzA to assess the effectiveness and appropriateness of these specific measures in accordance with Art. 5 Para. 3 TCO Regulation.

An assessment of the effectiveness of implemented specific measures is equivalent to an assessment of the quality of the overall content moderation of a hosting service. In particular, providers whose moderation processes show weaknesses in key areas will also have problems complying with the TCO Regulation.

Based on the previous analysis of the practice of content moderation in Germany, possible and appropriate content moderation measures that a hosting service can take to counteract the dissemination of illegal and, in particular, terrorist content via its platform are presented below. With regard to the requirement of (economic) appropriateness of the measures for hosting services of different sizes pursuant to Art. 5(3)(b) TCO Regulation, the costs and effort of the measures are discussed on the basis of the previous analysis.

All of the specific measures described are relevant market practices used by hosting service providers to moderate content. The measures are divided into the following areas:

a) Community guidelines
b) Moderation process
c) Manual moderation procedures
d) Automated moderation procedures
e) Cooperation with third parties

The measures are presented in tabular form in a matrix that differentiates between "All services" and "Large services/VLOPs" as well as between "Good industry practice" and "Extended measures".

Measures that are generally recognised and practised even for small hosting service providers with a focus on Germany can be found in the "All services" row. Larger services that operate on different markets and in different languages are shown separately. These include VLOPs/VLOSEs and other social networks that are subject to the NetzDG.

At the same time, the matrix classifies generally recognised measures that are usually practised by all relevant hosting services in the category "good industry practices". Additional measures are referred to as "extended measures", which are used by some services for structural reasons. These structural reasons include a potentially greater risk due to the target group addressed (e.g. minors) or the thematic focus (e.g. discussion

of political topics) as well as technical necessities due to the predominant type of signal (e.g. live content, video content).

The generally recognised "good practices of the industry" correspond to the specific minimum requirements of Art. 5 Para. 3 TCO Regulation and are generally feasible and appropriate for all hosting services and only indirectly dependent on their performance (size, financial strength, staffing levels, etc.). However, extended measures may be necessary in the event of a specific threat from or increased exposure to illegal content.

Especially for hosting services that are officially considered to be "exposed to terrorist content", it may be necessary to take extended measures in individual cases if the "good practices of the industry" have already been implemented and the dissemination of illegal content, especially terrorist content, cannot be satisfactorily curbed.

Tab. 23 Sample table "Possible specific measures, by size of service"

Size	Good practices in the industry	Extended measures
All services	▪ Industry consensus ▪ Generally recognised and practised ▪ Measures for all services realisable	▪ Measures that go beyond the industry standard ▪ may not make sense for some services or may not be feasible in terms of cost
Large services / VLOPs	▪ Consensus among VLOPs ▪ Often not feasible for smaller services due to the effort involved	▪ Measures that go beyond the VLOP standard ▪ often not useful for some services

Source: Goldmedia Analysis 2023

The type and scope of extended measures may depend more on the individual performance of the hosting service provider. However, the majority of the extended measures relate to organisational aspects that do not necessarily have to lead to substantially increased financial expenditure for content moderation.

5.2 Specific measures to achieve the minimum standards of content moderation

5.2.1 Community guidelines

The design of the community guidelines is a key pillar of the entire content moderation process. Beyond legal regulations, the service determines what content is permitted on its platform and how violations of the community guidelines are dealt with.

The provider thus defines its own benchmark against which its content moderation efforts can be measured. A concrete, operationalisable definition of undesirable content and sanction mechanisms is essential for the application of these guidelines within the moderation processes and procedures.

Tab. 24 Possible specific measures, by size of service

Size	Good practices in the industry	Extended measures
All services	▪ adequate, operationalisable Community directives ▪ Transparent standards of content moderation ▪ Communication of the Community Directives	▪ Clarification of the Community Directives
Large services/ VLOP	-	-

Source: Goldmedia Analysis 2023

In this respect, there is an objective and easy-to-check standard that can be used to initially check whether a provider has sufficiently dealt with aspects of content moderation, in particular with regard to illegal content.

The structuring and user guidance of the service also indicate the importance a provider places on communicating its own standards to its users. In addition to the mere findability of the community guidelines, the focus here is also on active communication of the standards and the associated sensitisation of users in the content upload environment

A review of the community guidelines of a service is also easy to realise in comparison to other services (good practice in the industry), as the community guidelines of all providers are transparent. Weaknesses in the specific design of community guidelines (lack of definitions, lack of operationalisation/concretisation, unaddressed areas of criminal law) can be identified even without knowledge of the technical or organisational details of moderation processes.

If the review of community guidelines reveals any ambiguities or shortcomings, the service can be requested to make changes in the sense of clarification. Compared to other measures, changes that only affect the policy level of a service can be implemented very quickly. This generally does not require the involvement of external service providers or the creation of additional internal resources. All hosting services should be able to react quickly to such a change recommendation.

5.2.2 Transparency of content moderation

Internal statistics of a content moderation service are of crucial importance for assessing the effectiveness of a provider's moderation practices. Due to legal requirements, key information is subject to publication. The transparency reporting obligations under Article 15 DSA and the previously existing transparency obligations under Section 2 NetzDG already provide a good overview of a service's content moderation practices. In future, the obligation to report to the DSA online transparency database in accordance with Article 20(3) DSA will also provide a further source of information to obtain an up-to-date overview of the scope of moderated posts, the occurrence of illegal content and the detection methods.

However, quantitative benchmarking of a service with other services on the basis of the transparency reports or the online transparency database is only possible to a limited extent, as the moderation and associated documentation processes can differ greatly. A longitudinal comparison of the transparency reports/online transparency database of a

service over different reporting periods is therefore more helpful. However, this will only be possible in the medium term, as the transparency reporting obligations will apply comprehensively to hosting services from 2024 (with the exception of micro and small enterprises).

Tab. 25 Possible specific measures, by size of service

Size	Good practices in the industry	Extended measures
All services	▪ Transparency reports in accordance with Art. 15 DSA (or comparable)	▪ Detailed key figures extended transparency reports ▪ More closely meshed (e.g. monthly) non-public reporting obligations
Large services/ VLOP	▪ Transparency reports additionally based on the NetzDG metrics	-

Source: Goldmedia Analysis 2023

The effectiveness of the moderation of illegal content is primarily measured by the processing time from receipt of the report and the reliability of the recognition of illegal content ("false negatives"/"false positives").

The envisaged reporting obligation under Art. 15 DSA, according to which only the media time required by intermediary service providers to reach a decision must be reported, is unfortunately not sufficiently meaningful for the assessment. The reporting standard of Section 2 para. 2 no. 9 NetzDG (within 24 hours, within 48 hours, within one week, etc.) would be better suited for a detailed assessment. If services only report at the statutory minimum level in their public transparency obligations, it will therefore still be necessary in future to ask providers for more detailed analyses of the moderation periods.

However, the reliability of the detection of illegal content should already be sufficiently recognisable from the information in Article 15(1)(d) DSA.

Furthermore, if necessary, a separate presentation of the processing times and the reliability of detection for the subset of terrorist online content in relation to the total amount of illegal content, as well as the definition of terrorist content that the service uses as the basis for its moderation practice, may be desirable. Although this information is often provided in transparency reports on a voluntary basis, if necessary, a provider potentially exposed to terrorist content would have to ensure that this information is available for an official review.

In addition, the statutory annual reporting obligation for a potentially suspended provider does not appear to be sufficient to allow acute problems arising from special situations or specific operational restrictions to be recognised. If necessary, reporting at shorter intervals could therefore be agreed with suspended providers as part of a probationary period in order to be able to conduct a targeted dialogue on concrete specific measures.

5.2.3 Content moderation process

The content moderation process regulates all organisational aspects of content moderation, from the receipt of messages to the assignment and prioritisation of messages,

the interlinking of technical systems with manual moderation procedures and the recording and monitoring of all moderation processes.

All process components must be adequately documented by the service for the functioning of its internal operating procedures. Particularly in the case of larger providers, where content moderation is based on a strong division of labour, task and role profiles as well as technical and personnel interfaces must be defined precisely and without contradiction. A process evaluation of content moderation can therefore be carried out on the basis of these internal process descriptions.

Tab. 26 Possible specific measures, by size of service

Size	Good practices in the industry	Extended measures
All services	▪ Moderation guide ▪ Process for training/further education ▪ Guidelines/assistance for community moderation ▪ Simple signalling functions ▪ Opposition proceedings ▪ Documentation of results	▪ More prominent reporting functions for (non-)users ▪ Continuous improvement (PDCA) and updating of guidelines ▪ Establishment of a central safety centre for users ▪ Greater sensitisation and involvement of users in moderation
Large services/ VLOP	▪ External service providers (BPO) ▪ Continuous improvement (PDCA) ▪ Central safety centre for users	▪ DSA specifications for VLOPs ▪ Risk management and crisis response (audited) ▪ Independent compliance department

Source: Goldmedia Analysis 2023

The central documents here are the entirety of moderation guidelines and aids, in which the community guidelines are operationalised in such a way that the employed moderators are enabled to make uniform moderation decisions. In addition, there are training and educational materials in which the specific standards for moderation decisions are usually illustrated and didactically conveyed using past moderation decisions.[99] In the case of services with elements of community moderation, there are also documents created specifically for the user moderators.

In the case of larger services, agreements with external service providers are also required, in which staff deployment, working hours, technical interface specifications, response times and service levels can be defined in addition to the moderation and training documents. Regulations for measuring quality and monitoring the moderation decisions made should also be defined for large platforms. Without structured mechanisms for quality checks, a service will not be able to evaluate its compliance with its own guidelines.

All services should be able to demonstrate how the further development process of the moderation guidelines is structured. As a rule, there are formalised feedback processes

[99] One example of this is Facebook's training material on dealing with terrorist content, which The Guardian published in 2017: https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content, accessed on 27/10/2013

and meetings at team level for this purpose. Due to the dynamic nature of content moderation, it can be assumed that adjustments can be made comparatively frequently and at short notice at a granular regulatory level. At the same time, however, there should also be further training events several times a year in which training is provided across teams and locations. If moderation guidelines and training documents are not up to date, moderating content in critical moderation environments such as terrorism in a way that is appropriate to the thematic dynamics is challenging.

In addition to the internal process documentation, the interface between users of the service and the internal moderation processes is an additional aspect that should be considered in the context of the process analysis. The following questions are the focus here:

- To what extent is it generally possible for users (and non-users) to report content?
- Are the reporting functions designed to be user-friendly and intuitive to find?
- Are users adequately informed about unwanted content and the consequences of reporting it?
- Is there a central location where users can obtain information on content moderation standards and, if necessary, make their own moderation settings (security centre)?

Very large online services should also have risk management and crisis response mechanisms involving higher levels.

The overall view of the process documents usually provides a good overview of the mechanisms that a service provides to deal with the dynamics in the area of content moderation. On this basis, specific recommendations for improving content moderation can already be made without having to refer to the organisation of a service's community guidelines.

5.2.4 Manual moderation

Manual moderation by employees of the service (safety and security team) is the foundation of all content moderation. Without manual moderation or employees of a service who deal with content moderation, adequate content moderation cannot be guaranteed. In this respect, the existence of a department responsible for content moderation within a service is mandatory for online platforms that publish content on behalf of their users.

The size of the service does not play a decisive role here. The personnel costs for manual moderation can be mapped even for small providers: For the small services spoken to as part of the study, an average of between 100 thousand and 300 thousand user contributions per month were accounted for by one full-time position in content moderation. Typical staffing levels in the moderation department ranged from 3-12 full-time positions, which usually moderate during office hours on weekdays and in the evening hours when usage is high.

As a "cost of doing business", the proportion of moderators in the total workforce is higher for smaller services than for large services, but moderation teams with up to a dozen employees do not represent an unreasonable burden even for small services.

Round-the-clock moderation is not feasible for small departments with fewer than 10 employees in content moderation.

From a significant double-digit number of employees in the area of content moderation, the advantages of (partially) outsourcing the safety and security team to specialised service providers outweigh the disadvantages. In particular, the initial outlay is relatively high, as external content moderation relies on significantly more extensive moderation guidelines and process descriptions than content moderation within a small moderation group that only moderates at a single location.

Tab. 27 Possible specific measures, by size of service

Size	Good practices in the industry	Extended measures
All services	yes, manual moderation is mandatory ▪ internal department	▪ Manual moderation 24/7/365 ▪ Own cue terrorism, if applicable
Large services/ VLOP	▪ 24/7/365 moderation ▪ Specialised cues (separate teams for different areas of the Community guidelines)	▪ Strengthening the cue for terrorism ▪ Establishment/strengthening of the internal analysis team for terrorist threats ▪ Involvement of a recognised, external moderation service provider

Source: Goldmedia Analysis 2023

Depending on the provider's moderation strategy, community moderation by users and self-moderation by content creators represent a central or supplementary element of manual moderation. Community moderation is a typical element of online forums. The large social media platforms, on the other hand, only allow users to moderate their own accounts.

While some services rely heavily on automated procedures for early detection, other services focus on the moderation of complaints by users. A mere comparison of staffing levels between different services is therefore hardly meaningful. When assessing the staffing ratio, for example, it is important to consider how relevant it is for the respective platform to maintain the relationship with the contributing users (e.g. free use vs. subscription-based use).

If a service is potentially exposed to terrorist content, it should first be investigated whether there are structural problems in the organisation of manual moderation:

- Are the moderation teams sufficiently integrated into the service's internal safety and security processes?
- Are the internal training and development processes sufficiently formalised?
- Can moderation processes be accelerated through greater specialisation of moderators in certain moderation areas?
- Is there a lack of expertise in recognising terrorist content?
- Was illegal content not recognised due to a lack of cultural context?
- Do problems accumulate at times when manual moderation is not carried out and does the period of manual moderation therefore need to be extended?

Depending on the outcome, it may be necessary to restructure the tasks within manual moderation, build up specific expertise in specialised teams or improve the general availability of manual moderation. Depending on the size and performance of the provider, this may mean not only expanding manual moderation but also incorporating external expertise from specialised service providers. External moderation service providers in particular usually have experience in the early detection and moderation of terrorist content. Provided that the service is of an appropriate size, the involvement of external expertise can provide a short-term remedy if a service that is potentially exposed to terrorist content does not yet have dedicated teams for the early detection or moderation of terrorist content.

5.2.5 Automated moderation procedures

The use of automated moderation processes is standard in the area of content moderation. All providers use automated processes to support content moderation, and automated processes play a key role in the early detection of problematic content, particularly in the case of text content.

All-in-one moderation solutions from external service providers also include (pre-trained) AI support for recognising problematic content. Some AI applications are also available as open source so that services can use them to develop their own applications (see Chap. 8.2).

AI-supported moderation tools from the very large IT groups are comparatively cheap to obtain in the area of text analysis and easy for services to implement in their existing moderation architecture: Marketable cloud-based moderation tools cost between around 90-180 euros for the moderation of 100,000 text comments or images.

Tab. 28 Possible specific measures, by size of service

Size	Good practices in the industry	Extended measures
All services	yes, automated moderation mandatory (for text)	▪ Synchronisation with hash databases (text, images, videos) ▪ External provider with TCO specialisation
Large services/ VLOP	▪ Synchronisation with hash databases (text, images, videos) ▪ AI training on your own data sets	-

Source: Goldmedia Analysis 2023

The use of automated systems in the area of content moderation is a minimum standard in the area of text moderation and can be expected from all providers. These systems do not require a high level of development effort, nor do they incur significant operating costs. The use of automated systems for image and video content is not yet widespread

among small services. As a rule, however, user-generated content for small services is primarily text-based. [100]

In this respect, a service that is potentially exposed to terrorist content is urgently required to use automated content moderation processes. On the basis of ongoing market analyses, it is important to check whether a service can optimise its existing moderation architecture with sensible, possibly newly available solutions for text moderation.

The landscape of external moderation providers is diverse and differentiated. Some providers focus in particular on their experience in the early detection of terrorist content. As the cost of automated solutions is generally low compared to the personnel costs of manual moderation, it seems reasonable for providers exposed to terrorist content to specifically request an offer from such providers. If no additional solutions are implemented, this would have to be justified.

In addition to text content, only larger services currently also moderate image and video content using AI-supported automated processes, which are usually developed in-house and trained on their own data.[101] These are used in particular for the early detection of certain illegal content (misrepresentations, terrorist propaganda, etc.).

5.2.6 Co-operation with third parties (law enforcement and NGOs)

A specific point of transparency in content moderation is the description of a hosting service's cooperation with external bodies and organisations that can support a platform's moderation service.

Services should first have defined their own reporting channels to law enforcement authorities at process level, especially for terrorist content and other serious legal offences (e.g. child abuse) as well as current risk situations.

In addition to the mandatory cooperation with law enforcement authorities in accordance with Art. 9 DSA, it is possible to provide your own communication channels to trustworthy community members or reporting offices of NGOs, whose reports are prioritised (own trusted flagger programmes).

[100] Image and video content is primarily distributed by specialised large or very large online platforms, which not only distribute the content, but also simplify the creation of content by providing "creator tools" via their platform.

[101] External solutions available on the market that are trained on generic data sets have so far only been suitable to a limited extent, as their use is still comparatively expensive and comparatively unreliable.

Tab. 29 Possible specific measures, by size of service

Size	Good practices in the industry	Extended measures
All services	▪ Cooperation with law enforcement authorities	▪ Hash databases ▪ Appointment of Trusted Flagger
Large services/ VLOP	▪ Hash databases ▪ Trusted Flagger Programme ▪ Participation in (international) industry forums	-

Source: Goldmedia Analysis 2023

It is also possible to convert user content into hash values and compare them with various hash value databases (e.g. GIFCT Hash Sharing Database, TCAP Terrorist Content Analytics Platform, NCMEC "Take It Down" or the EU IRU's Check-the-Web application). These cross-industry hash databases (cf. chap. 8.2) are of particular importance, as they significantly reduce the classification effort for all providers. Once classified as illegal content, copies of content with the same content can be removed from other platforms without having to be manually reviewed again.

So far, it appears that it is mainly large services that are participating in the industry-wide hash databases, partly because membership of the Global Internet Forum to Counter Terrorism (GIFCT) is linked to an extensive admission process. However, if a service repeatedly attracts attention by disseminating content that is already indexed in the relevant GIFCT hash platforms, membership should be strongly recommended.

In addition, there are now several international forums that deal with the topic of content moderation of illegal content. Their publications and databases provide information for the further development of moderation processes. All EU-wide relevant reporting centres and databases on the topic of curbing illegal content on hosting services are presented in the appendix in chapters 7 and 8.

5.3 Summary and recommendation

Whether the number of moderators deployed and the technical systems implemented for the automated support of content moderation are sufficient to permanently curb terrorist content in particular cannot be assessed on the basis of fixed indicators due to the major differences between the various services in terms of content, media forms and communication styles.

However, transparency reports provide information on provider-specific features that can be helpful for an in-depth discussion with the service about moderation processes. The ability of a hosting service to provide an adequate moderation service can be determined in general terms from the publicly available community guidelines and transparency reports. On the basis of supplementary, internal documents of the service (moderation guidelines, process descriptions, etc.) and the specific circumstances that led to the classification as a "hosting service exposed to terrorist content", recommendations can then be formulated for the further specification of the self-imposed community guidelines, the community guidelines and the internal organisational processes.

Some anomalies, such as a very low number of moderators in relation to the monthly user volume, the non-utilisation of accessible sources for the identification of terrorist content (e.g. GIFCT, see appendix, section 8.2) or the lack of proactive, automated or AI-supported filter systems, can also be addressed directly. The findings of this study show

that the implementation or expansion of manual and, in particular, automated content moderation processes is technically feasible and financially reasonable.

However, the extent to which the process documentation presented is applied in moderation practice can only be verified by subsequently monitoring the success of the hosting services concerned. For a sustainable implementation of the additional specific measures, a continuous review of the moderation statistics of the respective service is therefore necessary. In the best case scenario, this is done in conjunction with a voluntary agreement with the affected service as to which measurable goals are to be achieved with the additional specific measures and over what period of time.

In order to monitor the moderation statistics, it would be expedient to agree on progress reports with the hosting services exposed to terrorist content within a probationary period in addition to the annually required transparency reports with shorter periods (e.g. 3-6 months), in which the qualitative further development of the moderation measures should also be outlined.[102] At the same time, the transmission of new versions of the moderation guidelines and training documents can also be agreed as soon as they are used in the moderation process.

Another future option for monitoring progress in the area of content moderation on online platforms more effectively is to use statistical analyses from the DSA Transparency Database. These can be analysed for the individual platforms at country level. However, organisational progress can only be recorded deductively on this basis.

In addition, the BNetzA may recommend the use of external advice on content moderation if hosting services exposed to terrorist content are unable to demonstrate significant progress in the further development of their moderation performance after an appropriate transition period.

Full-service providers in particular, who offer both manual moderation and their own moderation systems and content filters, can advise hosting services on the sensible use of the various moderation procedures. The auditing companies that perform audit tasks for VLOPs in the context of Art. 37 DSA are also positioning themselves as consultants in this area. It can be assumed that the market for advice will continue to grow, especially for small and medium-sized companies, due to the significant increase in the number of hosting services that are now required to submit a transparency report under the DSA compared to the NetzDG.

[102] If suspended hosting services fall into the category of "micro and small enterprises", the preparation of such reports is strongly recommended, as they are not subject to any legal transparency requirements under the DSA.

Appendix

6 National authorities involved in the content moderation process

6.1 Role of the Federal Criminal Police Office

The **Federal Criminal Police Office** (BKA) is the central office for combating crime on the Internet. The Central Reporting Office for Criminal Content on the Internet (ZMI BKA) brings together decentralised reporting structures that exist in the federal states to combat hate and agitation on the Internet. The ZMI BKA serves the effective prosecution of criminal offences on the Internet such as propaganda offences, incitement to hatred or threats. The cooperating reporting centres include: "HessenGegenHetze", "REspect!", "die medienanstalten" and "Justiz und Medien - konsequent gegen Hass" (see Table 30).[103]

The BKA also plays a leading role in combating terrorist content. The Police State Security Division, supported by the Islamist-motivated Terrorism/Extremism Division, assumes the central role at the BKA in implementing the TCO Regulation in Germany. It is the sole authorised authority for issuing and reviewing removal orders under the TCO Regulation and for processing dangerous content from hosting services.

The BKA is in close contact with other state institutions that also specialise in the prosecution of cybercrime, such as the Central and Contact Point Cybercrime NRW at the Cologne Public Prosecutor's Office (ZAC NRW) or the Central Office for Combating Internet and Computer Crime in Hesse (ZIT Hessen).

However, the BKA does not offer a general reporting portal for illegal Internet content within the meaning of the TCO Regulation. The BKA's contact point exists for communication with a hosting service affected by a removal order. Citizens can report criminally relevant internet-related content to their competent police authorities or other reporting centres.

[103] Cf. https://www.bka.de/DE/KontaktAufnehmen/HinweisGeben/MeldestelleHetzeImInternet/meldestelle_node.html, retrieved on 22/09/23

6.2 Other authorities in Germany

In addition to the BKA, other bodies are involved in the TCO-VO process.

- The **Federal Network Agency** is responsible for monitoring and, if necessary, penalising hosting service providers that do not comply with removal orders. It is responsible for assessing the specific measures pursuant to Art. 5 (4-8) of the TCO Regulation and for imposing further fines (see section 1).
- Online **police** stations **and state criminal investigation offices** also transmit information to the Federal Criminal Police Office to issue removal orders[104]

6.3 Role of the state media authorities

The 14 **state media authorities** in Germany, which operate together under the umbrella brand "die medienanstalten", are responsible for the licensing and supervision of private radio and television broadcasters in accordance with the Interstate Media Treaty. They check compliance with advertising regulations and are also committed to ensuring diversity in private broadcasting and on the internet. At the same time, they are responsible for ensuring compliance with the protection of minors in broadcasting and on the internet on the basis of the Interstate Treaty on the Protection of Minors in the Media. As a large proportion of illegal content appears on online platforms that are used by young people in particular, the state media authorities are strongly committed to curbing hate speech and violations of human dignity on the Internet. They promote projects to teach media skills and run their own programmes aimed at combating illegal content on the internet (see Table 30). Similar to EU organisations and various NGOs (see sections 7.1 and 8.1), they are active and professional reporting bodies and have the status of "trusted flaggers" for many online platforms. In accordance with the TCO Regulation, they are involved in the review process before removal orders are issued, if this appears necessary. [105]

Tab. 30 State media authorities and their projects against illegal content on the internet (as of August 2023)

Media organisation	Project against illegal content
Media Authority NRW	▪ Tracking instead of just deleting (reporting centre)
Bavarian State Centre for New Media (BLM)	▪ Justice and media - consistently against hate: reporting centre with secure online cloud for the preservation of evidence and interface to the Public Prosecutor General's Office
Baden-Württemberg Communications Authority (LFK)	▪ REspect! (registration office) ▪ mobile phone sector

[104] Cf. https://www.bka.de/DE/UnsereAufgaben/Deliktsbereiche/PMK/TCO-VO/TCO-VO_node.html#:~:text=The%20BKA%20does%20not%20offer%20any%20general%20reporting%20to%20the%20state%C3%police%authorities%C3%B6rden%20, retrieved on 22.09.23
Cf. https://www.bka.de/DE/KontaktAufnehmen/HinweisGeben/MeldestelleHetzeImInternet/ZMIProzess/zmiprozess_node.html, retrieved on 22/09/23

[105] Cf. § 2 TerrOIBG

Media organisation	Project against illegal content
Lower Saxony State Media Authority (NLM)	▪ Central Office for Combating Hate Crime on the Internet - Lower Saxony ▪ Cooperation agreement Hatespeech must not remain without consequences.
Media Authority Hesse	▪ HessenGegenHetze (reporting centre) ▪ #NoPowerToHate ▪ MeldeHelden App
Rhineland-Palatinate Media Authority	▪ Tracking and deleting
Saxon State Authority for Private Broadcasting and New Media (SLM)	
Media Authority Berlin-Brandenburg (mabb)	▪ Tracking instead of just deleting
Media Authority Hamburg / Schleswig-Holstein (MA HSH)	▪ Eye-catcher
Media Authority Saxony-Anhalt	-
Thuringian State Media Authority (TLM)	-
Mecklenburg-Vorpommern Media Authority (MMV)	-
Saarland State Media Authority (LMS)	▪ Courage on the net - Together against hate and hate speech ▪ Media Competence Centre of the Saarland State Media Authority
Bremen State Media Authority (brema)	▪ RIKO - Resignation is not an option

Source: Goldmedia Analysis 2023

AI instrument KIVI of the Media Authority NRW

In order to better fulfil its supervisory function for telemedia services[106], the Media Authority of North Rhine-Westphalia (Landesanstalt für Medien – LfM) uses the AI tool KIVI. The name KIVI stands for the fusion of the terms KI and vigilare (Latin for vigilant).[107]

The automated process checks social media platforms and websites for potential legal violations, identifies them and prepares them for review.

The tool was developed in 2020 on behalf of LfM by Condat AG in Berlin. The one-off development costs were between 150,000 and 200,000 euros. The AI was trained on the basis of image and text samples that have been assessed as offences by LfM in the past. The specific offence categories include depictions of violence, incitement to hatred, the use of anti-constitutional symbols and freely accessible pornography.

The tool searches seven online platforms, such as Twitter/X, YouTube, Telegram and VK (Russian social network). One online platform is scanned alternately every day for around 6 hours. More than 10,000 sites are scanned automatically. Platforms of the Meta Group are not currently scanned, as the Group does not permit this (as of September

[106] See Section 88 (4) of the North Rhine-Westphalia State Media Act

[107] Cf. https://www.medienanstalt-nrw.de/zum-nachlesen/recht-und-aufsicht/mit-kuenstlicher-intelligenz-zu-einer-modernen-medienaufsicht.html, accessed on 22/09/23

2023). It would be possible to expand KIVI monitoring, but this would require additional resources.

Suspected cases identified by the tool are investigated by around 5 LfM employees during regular office hours.[108] If the suspicion is confirmed, the case is passed on to in-house lawyers. In justiciable cases, the online platforms are informed; this happens around 25 times a month.

The LfM is extremely satisfied with the performance of the tool. According to its own information, the number of criminal reports has doubled compared to previous months.[109] Almost all media organisations in Germany are currently working with the AI solution developed by LfM. LfM has also already received requests from other European countries to use its AI tool KIVI.

7 Reporting and supporting Law enforcement authorities and EU organisations

7.1 Law enforcement authorities of the EU

The **EU Internet Referral Unit** (EU IRU) detects and analyses terrorist and violent extremist content on the internet and social media. The EU IRU was founded in 2015 and is based at Europol's European Counter-Terrorism Centre. Its work spans multiple language groups and jurisdictions. It provides strategic intelligence on jihadist terrorism and information that can be used in criminal investigations.

The EU IRU has the following core tasks:

- Supporting the relevant EU authorities by providing strategic and operational analyses;
- Labelling terrorist and violent extremist online content and passing it on to the relevant partners;
- Detect internet content used by smuggling networks to lure migrants and refugees and request their removal;
- Rapid implementation and support of the referral process in close cooperation with the industry[110]

The **European Union Internet Forum** (EUIF) **is** a public-private partnership of the EU to combat terrorist content, child sexual abuse and violent right-wing extremism

[108] Cf. https://www.bpb.de/lernen/digitale-bildung/werkstatt/513732/ki-in-der-medienaufsicht-was-leistet-das-tool-kivi, accessed on 22/09/23

[109] Cf. https://www.medienanstalt-nrw.de/zum-nachlesen/recht-und-aufsicht/mit-kuenstlicher-intelligenz-zu-einer-modernen-medienaufsicht.html, accessed on 22/09/23

[110] Cf. https://www.europol.europa.eu/about-europol/european-counter-terrorism-centre-ectc/eu-internet-referal-unit-eu-iru, accessed on 22/09/23

online. It was launched by the EU Commission shortly after the EU IRU was founded in 2015. The EU IRU has been a member of the EUIF ever since.

The priority areas in which the Forum is focussing its efforts are:

- the implementation of the EU Crisis Protocol (EUCP),
- Reactions to right-wing extremist and violent online content and
- Reactions to new challenges.[111]

The **European Counter-Terrorism Centre** (ECTC) has been bringing together Europol's law enforcement and counter-terrorism capacities since 2016. This includes, for example, analysis projects in the area of counter-terrorism. The establishment of the centre has significantly increased the exchange of information between the authorities.[112]

The European Union's **Digital Europe Programme** (DIGITAL) supports Safer Internet Centres in 27 European countries with the aim of promoting media literacy among children, parents and teachers, raising awareness of potential risks on the Internet and offering children and young people telephone advice on online problems. Reporting centres for illegal content are also financed. In Germany, the Safer Internet Centre is implemented by the **Safer Internet DE association.** In addition to the klicksafe Awareness Centre, this includes the internet hotlines internet-beschwerdestelle.de (operated by eco and FSM) and jugendschutz.net as well as the helpline for children and young people Nummer gegen Kummer.

7.2 EU law enforcement platforms and databases

The EU IRU operates **PERCI** (Plateforme Européenne de Retraits de Contenus Illégaux sur Internet), an **EU platform** developed by Europol **to combat illegal online content**. Through this platform, Member States can submit notifications and mandatory removal orders under the TCO Regulation to hosting services to remove all types of terrorist content and thus facilitate the implementation of the TCO Regulation. PERCI went live in July 2023. During the transition phase, the **Internet Referral Management Application (IRMA**) was used for referrals. Removal orders were issued via the **Secure Information Exchange Network Application (SIENA)**.[113] The EU IRU also operates the **Check the Web** portal - a reference library on jihadist online terror propaganda, including right-wing terrorist content. This is an operational tool to support EU Member States in recognising new content, trends and patterns in the context of terrorist propaganda.

[111] Cf. https://home-affairs.ec.europa.eu/networks/european-union-internet-forum-euif_en, accessed on 07.09.23

[112] Cf. https://www.bmi.bund.de/DE/themen/sicherheit/nationale-und-internationale-zusammenarbeit/internationale-terrorismusbekaempfung/internationale-terrorismusbekaempfung-textbaustein.html, accessed on 07.09.23

[113] Cf. https://www.europarl.europa.eu/doceo/document/E-9-2021-004182-ASW_DE.html, accessed on 07.09.23
Cf. https://www.europol.europa.eu/cms/sites/default/files/documents/PERCI%20TCO%20regulation_partial%20release.pdf, accessed on 07.09.23

7.3 Other EU counter-terrorism projects

The **EU** Intelligence Analysis Centre (**INTCEN) (**before March 2012 Joint Situation Centre, SitCen or JSC) is an organ of the European External Action Service and has intelligence tasks alongside the European Union Satellite Centre (SatCen) and the Intelligence Division. [114]

SIRIUS is an EU-funded project that facilitates access to cross-border electronic evidence for law enforcement and judicial authorities in the context of criminal investigations and proceedings.

Horizon 2020 was the EU research and innovation funding programme for the period 2014-2020 with a budget of almost 80 billion euros. The programme was replaced by Horizon Europe. For Horizon 2020, there was a call for proposals for the management of information and data flows to combat (cyber)crime and terrorism. The following table lists selected projects that were funded by the Horizon 2020 framework programme.

Tab. 31 Selection of EU-funded Horizon 2020 projects

Project	Mission Statement	URL
CONEXXIONs	Interconnected Next-Generation Immersive IoT Platform Of Crime And Terrorism Detection, Prediction, Investigation, And Prevention Services	https://www.connexions-project.eu/
COPKIT	Technology, training and knowledge for Early-Warning/Early-Action led policing in fighting organised crime and terrorism	https://copkit.eu/
CREST	Fighting crime and terrorism with an IoT-enabled autonomous platform based on an ecosystem of advanced intelligence, operations, and investigation technologies	https://project-crest.eu/
EXFILES	Europe fights against crime and terrorism	https://exfiles.eu/
GRACE	Global Response Against Child Exploitation	https://grace-fct.eu/
INSPECTr	Intelligence network and secure platform for evidence correlation and transfer	https://inspectr-project.eu/
ROXANNE	Real time network, text, and speaker analytics for combating organised crime.	https://roxanne-euproject.org/
SIRIUS	Cross-Border Access To Electronic Evidence	https://www.europol.europa.eu/operations-services-and-innovation/sirius-project
SPIRIT	Scalable privacy preserving intelligence analysis for resolving identities.	https://www.spirit-tools.com/
APPRAISE	Facilitating public & private security operators to mitigate terrorism scenarios against soft targets	https://appraise-h2020.eu/
Dante	Detecting and analysing terrorist-related online contents and financing activities	https://www.h2020-dante.eu/

[114] Cf. https://de.wikipedia.org/wiki/INTCEN, accessed on 07.09.23
Cf. https://op.europa.eu/de/web/who-is-who/organization/-/organization/EEAS/EEAS_CRF_237388, accessed on 07.09.23

PROTON	Modelling the processes leading to organized crime and terrorist networks	https://www.projectpro-ton.eu
GRACE	Global Response Against Child Exploitation GRACE aims to equip European law enforcement agencies with advanced analytical and investigative capabilities to respond to the spread of online child sexual exploitation material.	https://www.grace-fct.eu/
AIDA	Artificial Intelligence and Advanced Data Analytics for Law Enforcement Agencies Breakthrough techniques against cybercrime and terrorism	https://www.project-aida.eu/index.php
CTC Project	Cut The Cord (CTC) project aims to prevent and predict, while assisting Law Enforcement Agencies and other entities to fight financial crimes and "cut the cords" to non-traditional products for financing and supporting terrorist organisations.	https://ctc-project.eu/
NOTIONES	iNteracting network of intelligence and security practitiOners with iNdustry and academia actorS The vision of the NOTIONES network is to build and maintain a pan-European ecosystem of security and intelligence practitioners	https://www.notiones.eu/
RED ALERT	Real-Time Eary Detection and Alert System For online terrorist content based on natural language processing, social network analysis, artificial intelligence and complex event processing	http://redalertproject.eu/
Starlight	Enhancing the EU's strategic autonomy in the field of artificial intelligence (AI) for law enforcement agencies (LEAs).	https://www.starlight-h2020.eu/

Source: Goldmedia Analysis 2023

8 Non-governmental reporting offices and databases

Unlawful or otherwise problematic content that is distributed by hosting services can also be reported to other bodies in addition to the hosting services distributing it directly. These include official or public bodies, but also non-governmental bodies, such as those operated by Internet industry associations as an instrument of self-regulation. There are also a number of non-governmental organisations that deal with aspects of content moderation in civil society, particularly in the area of anti-Semitic or otherwise hateful content. As a rule, these organisations do not maintain their own reporting centres, but can cooperate with online platforms in the area of prevention.

8.1 Non-governmental reporting offices in Germany

Reporting or complaints offices are an effective tool for informing hosting services of potential violations of applicable law that have not been recognised by the provider's internal content moderation. The most relevant reporting centres in Germany are briefly presented below.

The terms "reports" and "complaints" are often used interchangeably to refer to communications submitted by Internet users when they encounter problematic or inappropriate content. The terms "report" and "complaint" are also used interchangeably below.

Internet complaints centre.de

Users can use online forms to report Internet content that they consider to be illegal to a complaints centre. An important online complaints centre in Germany can be reached at **Internet-Beschwerdestelle.de.** The service at Internet-Beschwerdestelle.de is a joint project of eco - Association of the Internet Industry (eco) and the Voluntary Self-Regulation Body for Multimedia Service Providers (FSM) (see below).

With Internet-Beschwerdestelle.de, eco and FSM have been offering a contact point for Internet users since 2004 to obtain information about safer use of the Internet and to submit complaints. They are also founding members of the **International Association of Internet Hotlines (INHOPE)**, the umbrella organisation for Internet hotlines, through which hotlines work together at an international level.

Complaints received there are legally examined by the respective institution and, if the reported content violates the relevant youth media protection laws or relevant criminal laws, further steps can be taken:

The content provider is asked directly to modify the content or the host provider is asked to arrange for the content to be removed. In serious cases, the complaint can also be forwarded directly to the responsible government agency in anonymised form. Complaints about illegal online content that is not hosted on a server in Germany are forwarded by eco and FSM to the relevant INHOPE hotline.

Complaints submitted via Internet-Beschwerdestelle.de are generally processed by eco and the FSM according to a division of tasks and responsibilities. FSM: World Wide Web (in cooperation with eco if related to an eco member), mobile content & apps and chat. eco: Newsgroups, spam / email, discussion forums and peer-to-peer.

eco - Association of the Internet Industry

eco - Association of the Internet Industry has been representing the interests of the Internet industry in Germany since 1995 and is the largest association of the Internet industry in Europe with around 900 members. Any Internet user who comes across content on the Internet that is relevant to the protection of minors or who wants to complain about unauthorised e-mail advertising can contact the complaints office, which has been in existence since 1996 - anonymously if they wish. Some services also use the complaints centre to obtain an external legal opinion in more complex cases of content moderation.

At eco, around 5 people are employed in the complaints centre every working day. All of them have successfully completed legal training (at least state examination). No automated procedures are used for the review and legal assessment of complaints.

In 2022, the complaints office received around 18,000 complaints, 49.2 per cent of which turned out to be justified. The majority of the justified reports (91.6 per cent) related to child pornography, while a very small proportion related to anti-constitutional content (0.4 per cent).

In 8,904 cases, content was removed by the hosting services through notice-and-takedown. The overall success rate was 97.7 per cent, even though only 29.6 per cent of the incriminated content was hosted by German services. The removal rate for child pornography from German hosting service providers is 100 per cent. On average, content is removed by the hosting services after 2.5 days.[115]

From eco's point of view, it is not only the very large online platforms where extremist content is disseminated. In its experience, small and sometimes private discussion forums are still very active. As these often lack easy-to-find reporting procedures, users often contact eco's complaints centre to report content.

In terms of criminal prosecution, the complaints office cooperates with the BKA, the Central and Contact Point Cybercrime NRW (ZAC) and normal police departments, among others. The INHOPE network plays a key role in international criminal prosecution.

FSM - Voluntary Self-Regulation of Multimedia Service Providers

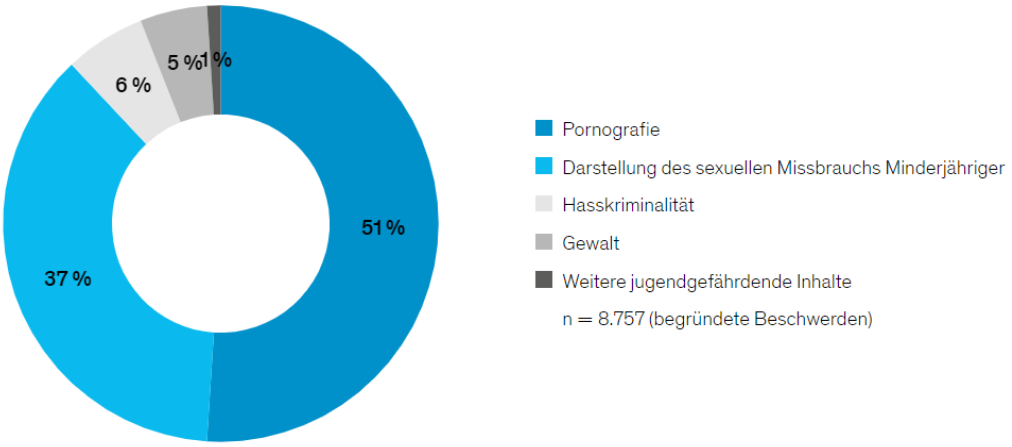
The Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM) is primarily dedicated to the protection of minors and combating illegal content in online media that is harmful to minors and adversely affects their development. The work of the FSM is primarily focussed on the self-regulation of service providers and the processing of complaints about illegal, youth-endangering and developmentally harmful content from member companies, but also from non-members. As a matter of principle, every complaint is reviewed manually by the FSM and, if necessary, submitted to an independent decision-making body, the Complaints Committee. [116]

In 2022, the FSM Complaints Office received a total of 12,956 complaints about illegal online content or content harmful to minors.

[115] Cf. https://www.eco.de/wp-content/uploads/2023/03/eco_beschwerdestelle_jahresbericht_2022.pdf, accessed on 22/09/23

[116] Cf. https://www.fsm.de/unternehmen/angebot/#beschwerdeausschuss, retrieved on 19.09.23

Fig. 9 Breakdown of substantiated complaints from the FSM Complaints Office by reason for complaint 2022



Source: FSM Annual Report 2022 "Complaints Office", online at: https://jahresbericht.fsm.de/2022/beschwerdestelle/, accessed on 19.09.23

In 68 per cent of cases (8,757 reports), the complaints were substantiated, i.e. content that violated German laws for the protection of minors in the media. Of the substantiated complaints, 51 per cent (4,455 cases) related to pornographic content and 37 per cent (3,224 cases) to depictions of the sexual abuse of minors.

A total of 55 per cent of the verified depictions of abuse of minors were hosted on German servers. The FSM immediately forwards these to the BKA and informs the hosting service in a notice-and-takedown procedure once evidence has been collected. The deletion of such content took an average of 1.5 days after receipt of the complaint, with the overall removal rate of such content being 100 per cent.

Jugendschutz.net

Jugendschutz.net was founded in 1997 and acts as a joint federal and state competence centre for the protection of children and young people on the Internet. Jugendschutz.net monitors online content for offences against the protection of minors and accepts complaints. The focus is on topics and services that are particularly relevant in terms of risks for children and young people. The jugendschutz.net team reviews the reported content, assesses it from a legal perspective and checks who is responsible for the content.

Jugendschutz.net handles complaints as follows:

- If a person responsible for a service is known, jugendschutz.net gets in touch and demands the removal of violations of youth protection regulations. If the responsible party is based in Germany and does not react or if their voluntary self-regulatory organisation remains inactive, the case is forwarded to the Commission for the Protection of Minors in the Media (KJM), which initiates media law proceedings.
- If the responsible party cannot be identified, the provider of the hosting service will be asked to rectify the offence. If the provider is a member of a voluntary self-regulatory organisation (e.g. eco), the infringement will be forwarded to them. In addition, jugendschutz.net forwards content to be indexed by the German Centre

for the Protection of Minors in the Media (Bundeszentrale für Kinder- und Jugendmedienschutz) so that it can be removed from search engines.

- In cases of foreign hosting services without contact in Germany, jugendschutz.net forwards the cases to cooperating complaint centres through international networks such as INHOPE (International Association of Internet Hotlines) and INACH (International Network Against Cyber Hate).

Help centres in Germany

The **Klicksafe** Awareness Centre is an EU initiative for greater online safety. Klicksafe aims to promote people's online skills and support them in using the Internet competently and critically with a wide range of programmes. It is coordinated by the Rhineland-Palatinate Media Authority and implemented jointly with the NRW Media Authority. Klicksafe has also been coordinating the Safer Internet Centre DE since 2008.[117]

HateAid is a non-profit organisation that campaigns for human rights in the digital space and is committed to combating digital violence and its consequences on a social and political level.[118]

As a networking centre against hate speech, **Das NETTZ** improves the framework conditions for engagement against hate online. They connect initiatives and activists from the digital civil courage community (currently around 138 players). To this end, they work closely with civil society organisations, political bodies and IT companies.[119]

Stark im Amt is a portal for local politics against hate and violence. It is the first central point of contact that provides local representatives with information and guidance on prosecuting hate speech.[120]

The **Amadeu Antonio Foundation is** committed to strengthening civil society in Germany against anti-Semitism, racism and right-wing extremism. To this end, it supports over 1,000 local initiatives and projects in youth culture, schools, victim protection, refugee initiatives and democracy projects financially, through education, public relations work and community networks.

Nummer gegen Kummer e. V. is the umbrella organisation for the largest free telephone counselling service for children, young people and parents in Germany.

Violence Prevention Network is a German non-governmental organisation active in the prevention of extremism and the deradicalisation of violent extremists. violent offenders. Its digital division (**Violence Prevention Network Digital) brings** together the findings and experience of the organisation's various offline and online projects, develops new digital projects that complement the existing structures of prevention practice in a targeted manner and tests new approaches to internet-based radicalisation prevention.

[117] Cf. https://www.klicksafe.de/die-initiative, accessed on 22/09/23

[118] Cf. https://hateaid.org/, accessed on 22/09/23

[119] Cf. https://www.das-nettz.de/, accessed on 22/09/23

[120] Cf. https://www.stark-im-amt.de/rat-und-tat/online-hetze/anzeigen-von-hassnachrichten/, accessed on 22/09/23

8.2 International organisations, committees and databases

Global Internet Forum to Counter Terrorism

The **Global Internet Forum to Counter Terrorism (GIFCT) is** a non-governmental organisation that aims to prevent terrorists and violent extremists from using digital platforms. The forum was founded in 2017 by Facebook, Microsoft, Twitter/X and YouTube to promote technical collaboration between member companies, drive relevant research and share knowledge with smaller platforms. In 2023, Meta took over the chairmanship of GIFCT, which now counts over 20 different online platforms as members that are committed to countering the spread of terrorist and violent extremist content online across all industries. GIFCT partners include Tech Against Terrorism (see below). [121]

The GIFCT operates a **hash-sharing database** in which the member companies provide terrorist, violent or extremist content that is shared via their platforms in various hash value variants.[122] Meta has also recently made a **hasher matcher actioner** (HMA) available on an open source basis, which each online platform can use to hash its own content and compare it with the content in the GIFCT database or other hash databases.[123]

The **Global Network on Extremism and Technology (GNET)** is the academic research arm of the GIFCT and aims to better understand the way terrorists use technology.

Tech Against Terrorism

Tech Against Terrorism (TATE) is an initiative launched and supported by the UN Counter Terrorism Executive Directorate (UN CTED) that works with the global tech industry to combat terrorist use of the internet while respecting human rights. Its action plan is based on outreach, knowledge sharing and practical support. Among other things, Tech Against Terrorism publishes an annual report on the prevailing terrorist threats on the Internet, which is aimed at the general public.[124]

Tech Against Terrorism has launched the **Knowledge Sharing Platform.** This is a collection of instruments (interactive tools and resources) with which start-ups and small tech companies can better protect themselves against terrorist exploitation of their services.

In 2018, they also founded the **Data Science Network**, the world's first network of experts working on the development and use of automated solutions to combat the use of small-scale technology platforms by terrorists while respecting human rights.

The **Terrorist Content Analytics Platform (TCAP)** developed by Tech Against Terrorism was launched in November 2020 with the support of Public Safety Canada. The TCAP is designed to prevent terrorists from using the internet by facilitating the quick and

[121] Cf. https://gifct.org/about/, retrieved on 22/10/23

[122] Cf. https://gifct.org/hsdb/, retrieved on 22/10/23

[123] Cf. https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/, accessed on 23.10.23

[124] Cf. https://www.techagainstterrorism.org/wp-content/uploads/2023/01/FINAL-State-of-Play-2022-TAT.pdf, accessed on 23.10.23

accurate removal of terrorist content. A team of in-house open source intelligence analysts tracks terrorist migration across a variety of technology platforms and reports URLs with terrorist content to the TCAP. They track the building of the world's largest database of verified terrorist content collected in real-time from verified terrorist channels on messaging platforms and apps (including hashes shared with GIFCT). They also support smaller technical platforms to improve content moderation.[125][126]

National Centre for Missing & Exploited Children

The **National Center for Missing & Exploited Children** (NCMEC) is the largest child protection organisation in the United States. They work to protect children and create important resources for them and the people who protect them. This includes taking action against Child Sexual Abuse Material (CSAM) on the Internet, which can be found in virtually every online space.[127] They run the **CyberTipline** - an online reporting system for all types of child sexual abuse online. They respond to millions of reports of child sexual abuse online every year. In addition, they offer numerous resources and support to victims and survivors of child sexual abuse. The NCMEC also offers a service called Take It Down, which helps remove nude, partially nude or sexually explicit photos and videos of minors by assigning a unique hash value to the images or videos. Online platforms can use these hash values to recognise these images or videos on their public or unencrypted services and take action to remove this content.

Lumen database

The **Lumen database** collects and analyses legal complaints and requests for removal of online material, helping internet users to know their rights and understand the law. Using this data, Lumen analyses the frequency of legal threats and shows internet users where the removal of content is coming from. [128]

9 Codes of conduct in the industry with reference to content moderation

A code of conduct is a regulatory instrument that is often developed in cooperation between the European Commission, civil society and companies. Such codes of conduct usually contain voluntary commitments for their signatories to take measures to achieve certain goals. Joining a code of conduct is voluntary.

These industry codes define desirable behaviours and can thus help online platforms to specify and achieve various otherwise vague due diligence obligations in connection with DSA.

The "Code of Conduct to combat illegal hate speech online" agreed between the European Commission and four large IT companies (Facebook, Microsoft, Twitter and YouTube) in May 2016 is particularly important. The code is intended to ensure that requests for the removal of online content are processed quickly by the companies. The

[125] Cf. https://www.terrorismanalytics.org/about, accessed on 23.10.23
[126] Cf. https://terrorismanalytics.org/about/how-it-works, accessed on 23.10.23
[127] Cf. https://www.missingkids.org/theissues/csam, accessed on 23.10.23
[128] Cf. https://www.lumendatabase.org/pages/about, accessed on 23.10.23

companies had committed to reviewing the majority of these requests within 24 hours and removing the content if necessary, while always upholding the principle of freedom of expression. So far, eight companies have signed up to the code: Facebook, YouTube, Twitter, Microsoft, Instagram, Dailymotion, Snapchat and jeuxvideos.com.[129]

Another important agreement is the "Code of Conduct against Disinformation", which has been in place since 2018 and was signed in a new version by 34 players in June 2022. The signatories have agreed on certain self-regulatory standards to combat disinformation. They committed to taking action in several areas, such as restricting the spread of disinformation, ensuring transparency in political advertising, improving cooperation with fact-checkers and improving access to their data for researchers.[130]

[129] Cf. https://ec.europa.eu/commission/presscorner/detail/de/qanda_20_1135, accessed on 06.09.23

[130] Cf. https://digital-strategy.ec.europa.eu/de/policies/code-practice-disinformation, accessed on 06.09.23