Status Quo spezifischer Maßnahmen von Hostingdiensten zur Inhaltemoderation

Studie im Auftrag der Bundesnetzagentur

Aktenzeichen: Z25-3-2023-0011

Stand: 23.11.2023

Auftragnehmer

Goldmedia GmbH Strategy Consulting

Prof. Dr. Klaus Goldhammer | Dr. André Wiegand

Oranienburger Str. 27, 10117 Berlin

Tel. +49 30 2462660

www.goldmedia.com

Inhalt

Genderhinweis: Aus Gründen der besseren Lesbarkeit wird im Text auf die gleichzeitige Verwendung männlicher u. weiblicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen (z. B. „Nutzer" oder „Moderator") gelten gleichwohl für beiderlei Geschlecht.

Dies impliziert keine Benachteiligung des weiblichen Geschlechts, sondern soll im Sinne der sprachlichen Vereinfachung als geschlechtsneutral zu verstehen sein.

1 Situation und Auftrag

1.1 Situation

Durch die EU-Verordnung zur Bekämpfung terroristischer Online-Inhalte (EU 2021/784; TCO-VO) und dessen deutscher Begleitgesetzgebung (Terroristische-Online-Inhalte-Bekämpfungsgesetz; TerrOIBG) hat die Bundesnetzagentur (BNetzA) neue Aufgaben erhalten. In Zusammenarbeit mit dem Bundeskriminalamt (BKA) ist die BNetzA nunmehr dafür zuständig, Hostingdienste gem. Digital Service Act (DSA) bei bestimmten Verstößen gegen die TCO-VO dazu anzuhalten, effektiv gegen eine Verbreitung terroristischer Inhalte durch ihre Dienste vorzugehen.

Hierfür kann die BNetzA Hostingdiensten, die terroristischen Inhalten ausgesetzt sind, u. a. dazu verpflichten, spezifische Maßnahmen zur systematischen Unterbindung der Verbreitung terroristischer Inhalte zu ergreifen.

Diese spezifischen Maßnahmen müssen geeignet sein, damit Hostingdienste terroristische Inhalte selbst ermitteln und zeitnah entfernen können. Beispiele für spezifische Maßnahmen sind etwa eine angemessene Ausstattung mit Personal, geeignete technische Mittel oder benutzerfreundliche Meldemechanismen für die Nutzer der jeweiligen Dienste.

Solche spezifischen Maßnahmen werden umgangssprachlich als „Content Moderation" (zu Deutsch: Moderation von Inhalten oder Inhaltemoderation) bezeichnet. Unter Inhaltemoderation versteht man die Verfahren und organisierten Praktiken zur Überprüfung von nutzergenerierten Inhalten, die auf Online-Plattformen oder anderen Hostingdiensten veröffentlicht werden, um festzustellen, ob ein Inhalt für u. a. einen Dienst oder unter einer bestimmten Gerichtsbarkeit geeignet ist. Die Verfahren können dazu führen, dass nutzergenerierte Inhalte von Moderatoren entfernt werden, die im Auftrag bzw. als Stellvertreter eines Dienstes handelt. Online-Plattformen, insbesondere soziale Medien, generieren große Datenmengen an nutzergenerierten Inhalten verschiedener Arten (Text, Video, Audio, Live-Inhalte wie Streams, Chats etc.). Für diese Inhalte besteht für die veröffentlichenden Dienste die Notwendigkeit, ihre eigenen Regeln (im Folgenden: „Gemeinschaftsrichtlinien") und den geltenden Rechtsrahmen durchzusetzen, da die Veröffentlichung unangemessener Inhalte eine gewichtiges Haftungsrisiko für den Dienst darstellen kann.[1]

Dieser Rechtsrahmen besteht in Deutschland im Wesentlichen aus dem Digital Services Act (DSA) der EU, welcher das bislang nur national gültige Netzwerkdurchsetzungsgesetzes (NetzDG) ablöst, der hier im Fokus stehenden TCO-Verordnung, den Vorgaben des Strafgesetzbuches sowie den Vorgaben des Jugendschutzgesetztes und des Jugendmedienschutz-Staatsvertrages.

[1] Vgl. https://link.springer.com/referenceworkentry/10.1007/978-3-319-32001-4_44-1, abgerufen am 22.09.23

1.2 Studienauftrag

Ein Hostingdienst, der terroristischen Inhalten ausgesetzt ist, hat nach TerrOIBG zunächst selbst darüber zu befinden, welche „spezifischen Maßnahmen" er zur Eindämmung von deren Verbreitung ergreift. Hierbei steht es dem Hostingdienst frei, jedwede Maßnahme vorzusehen, die er für geeignet hält, um gegen die Verfügbarkeit terroristischer Inhalte in seinen Diensten vorzugehen.

Die BNetzA beurteilt anschließend die Wirksamkeit und Angemessenheit „spezifischen Maßnahmen" zur Eindämmung terroristischer Inhalte nach Art. 5 Abs. 3 TCO-VO unter Berücksichtigung von Größe und Finanzkraft des jeweiligen Hostingdienstes.[2]

Um verhältnismäßige Entscheidungen zu treffen, benötigt die BNetzA ein Grundverständnis der möglichen und sachgerechten Maßnahmen zur Inhaltemoderation, die ein Hostingdienst ergreifen kann, um der Verbreitung illegaler und insbesondere terroristischer Inhalte über seinen Dienst entgegenzuwirken.

Um dies sachgerecht und ggfs. gerichtsfest beurteilen zu können, wird in der vorliegenden Studie zunächst der Status quo der bestehenden, bislang einschlägigen Marktpraktiken von Hostingdiensten zur Inhaltemoderation dargestellt, um der Verbreitung illegaler Inhalte entgegenzuwirken. Zudem werden die bestehenden und sich in Entwicklung befindlichen technischen Verfahren der Inhaltemoderation, sowie die damit verbundenen Aufwände analysiert (Arbeitspaket 1).

Im Anschluss werden Mindeststandards spezifischer Maßnahmenumfänge entwickelt, die von Hostingdiensten im konkreten Einzelfall eingefordert werden können, insbes. wenn sie mehrfach terroristischen Inhalten ausgesetzt sind. Der hiermit verbundene Aufwand soll zielgerichtet und verhältnismäßig sein, insbesondere im Hinblick auf die individuelle Finanzkraft des Hostingdienstes (Arbeitspaket 2).

1.3 Methodik

Die Status-quo-Analyse basiert auf folgenden Arbeitsschritten:

Einleitend erfolgte eine einordnende normative Betrachtung des komplexen Zusammenspiels des EU-Rechts mit nationalen, öffentlichen und privaten, verpflichtenden und soft-law-basierten Normen, um zu prüfen, welche Anbietertypen auf Basis welcher rechtlichen Vorgaben auch unabhängig von der neuen TCO-Regulierung Maßnahmen der Content-Moderation durchführen müssen.

In einem nächsten Schritt erfolgte eine Auswertung der vorliegenden NetzDG-Berichte sowie erster DSA-Reportings, um die bereits vorliegenden Informationen zu den eingesetzten Maßnahmen zum Umfang des Aufkommens an gemeldeten/identifizierten und behobenen Verstößen sowie den eingesetzten Maßnahmen der Content-Moderation der berichtspflichtigen Unternehmen zusammenzufassen.

Parallel dazu wurde ein Desk-Research zu den verfügbaren und eingesetzten Maßnahmen zur Inhaltemoderation durchgeführt.

---

[2] Falls die Beurteilung negativ ausfällt, ist der Hostingdienst zur Ergreifung weiterer Maßnahmen aufzufordern. Auch in diesem Fall konkretisiert der Anbieter wiederum selbst, welche weiteren spezifischen Maßnahmen er ergreifen möchte (Art. 5 Abs. 6 UAbs. 2 TCO-VO).

Darauf aufbauend wurden Expertengespräche sowohl mit Hostingdiensten als auch mit externen Anbietern manueller und automatisierter Verfahren der Content-Moderation und Meldestellen geführt. Hierfür wurde zuerst ein Überblick des deutschen Hostingdienstemarkts erstellt. Die zu untersuchenden Hostingdienste wurden mit dem Auftraggeber abgestimmt. Anschließend wurden folgende Expertengespräche geführt:

Hostingdienste

- 8 Hostingdienste-Anbieter,
- darunter 2 sehr große Online-Plattformen

Moderations-Dienstleistungen

- 6 Anbieter von Moderations-Dienstleistungen

Meldestellen

- eco – Verband der Internetwirtschaft e. V.
- Landesanstalt für Medien Nordrhein-Westfalen (LfM)

Folgende Fragekomplexe wurden mit den verschiedenen Gesprächspartnern diskutiert, wobei die Schwerpunktsetzung variabel war und vom Erkenntnisinteresse des jeweiligen Gespräches bestimmt wurde:

- Unternehmensgröße
- Technische und personelle Dimension
- Effektivität der Maßnahmen
- Abgrenzung / Zusatzaufwand durch TCO-VO
- Schweregrad der Betroffenheit
- Berücksichtigung von Art. 5 Abs. 3 lit. c TCO-VO
- Ausblick

In Kombination mit den Ergebnissen der NetzDG-Berichtsauswertung wurden Unterschiede, Gemeinsamkeiten und konsensuale Marktstandards (Best-Practice-Varianten) der Content-Moderation für verschiedene Hostingdienste-Typen und -Größen herausgearbeitet.

Im Anschluss wurde für kleinere, mittlere und große Unternehmen abgeleitet, welche Maßnahmen jedwedem Unternehmen möglich sind und welche Maßnahmen in welcher Weise von der jeweiligen Unternehmensgröße abhängen.

Zugleich erfolgte auch eine Bewertung, ob die identifizierten Maßnahmen und Systeme zur Inhaltemoderation in Art und Umfang ausreichend sind, insb. terroristische Online-Inhalte gem. Art. 5 TCO-VO zu entfernen.

Auf dieser Basis wurde abschließend beurteilt,

a) welche Anbieter welche Maßnahmen für die Umsetzung der Vorgaben aus Art. 5 Abs. 2 und 3 TCO-VO unter Berücksichtigung des Grundsatzes der Verhältnismäßigkeit gem. Art. 5 Abs. 3 lit. b und c TCO-VO umsetzen sollten und
b) welche Maßnahmen durch andere regulatorische Vorgaben grundsätzlich auch erforderlich sind und für eine Berücksichtigung der TCO-VO angepasst bzw. erweitert werden können.

2 Rahmenbedingungen

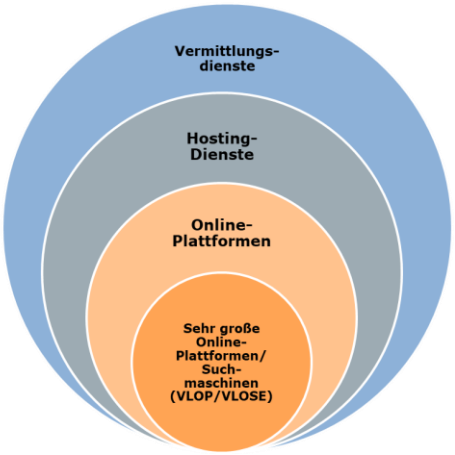
In diesem einleitenden Kapitel werden die rechtlichen Rahmenbedingungen der Content-Moderation zusammengefasst und dargestellt, welche Anbieter in Deutschland von der TCO-Verordnung umfasst sind. Diese Darstellung bildet die Grundlage für die Analyse der eingesetzten bzw. zur prüfenden Maßnahmen der Content-Moderation im Rahmen der TCO-Verordnung.

2.1 Rechtliche Rahmenbedingungen

2.1.1 Begriffssystematik des Digital Service Acts

Zur Klassifizierung verschiedener digitaler Dienste kategorisiert die EU-Definition im Digital Services Act (DSA) unter dem Oberbegriff „Vermittlungsdienste" verschiedene Dienstarten, darunter Online-Plattformen. Eine besondere Form der Online-Plattformen ist dabei die sehr großen Online-Plattform (Very Large Online Platforms, VLOP) bzw. sehr großen Suchmaschine (Very Large Online Search Engine, VLOSE). Vermittlungsdienste unterfallen hingegen in drei Unterkategorien: reine Durchleitungs-, Caching-, und Hostingdienste. Das folgende Schaubild dient der Illustration der Kategorien.

Abb. 1 Unterscheidung digitaler Dienste nach DSA-Kategorien



Quelle: BNetzA 2023

Dabei ist wesentlich, dass die in der Grafik dargestellten Kategorien aufeinander aufbauen und sich gegenseitig beinhalten: Eine sehr große Online-Plattform erfüllt damit immer auch die Kriterien für eine Online-Plattform, sowie die eines Hostingdienstes sowie die eines Vermittlungsdienstes. Eine Hostingdienst ist z. B. immer eine Teilmenge der Vermittlungsdienste, jedoch ist ein Hostingdienst nicht zwingend eine Online-Plattform oder eine sehr große Online-Plattform.

Im Unterschied hierzu bauen die drei Unterkategorien für **Vermittlungsdienste** nicht aufeinander auf. Ein Dienst ist ein Vermittlungsdienst, wenn er entweder die Kriterien für einen reinen Durchleitungs-, oder für einen Caching-, oder einen Hostingdienst erfüllt. Reine Durchleitungs- und Caching-Dienste können für sich Haftungsausschlüsse in Anspruch nehmen, solange sie die Integrität der übermittelten oder bereitgestellten Informationen nicht verändern.

Ein Hostingdienst speichert von Nutzern bereitgestellte Informationen in deren Auftrag. Nutzer können in diesem Zusammenhang jede natürliche oder juristische Person sein, die Hostingdienste mit dem Ziel in Anspruch nehmen, Informationen zugänglich zu machen. Für Hostingdienste[3] gelten Haftungsausschlüsse, solange sie zügig Inhalte entfernen oder den Zugang dazu sperren, sobald sie tatsächliche Kenntnis über rechtswidrige Tätigkeiten oder rechtswidrige Inhalte erhalten.[4] Diese Verfahren zum Erhalt der Haftungsausschlüsse wurden in der EU-E-Commerce-Richtlinie für Hostdingdienste unter dem Begriff „Notice and Takedown" eingeführt[5] und im DSA als „Notice and Action"-Mechanismus begrifflich erweitert.[6]

Ein Hostingdienst ist auch **eine Online-Plattform**, wenn das Kerngeschäft die Speicherung und öffentliche Verbreitung von Informationen im Auftrag eines Nutzers umfasst. Ausgenommen von dieser Definition sind Hostingdienste, bei denen die Verbreitung von nutzergenerierten Inhalten nur eine Nebenfunktion eines anderen Dienstes oder um eine unbedeutende Hilfsfunktion des Hauptdienstes darstellt.[7] Bei einer Online-Plattform erfolgt die öffentliche Verbreitung zwingend durch den Dienst, während bei einem Hostingdienst, der keine Online-Plattform ist (z. B. reine Cloud-Dienste) die Nutzer (oder Dritte) über die öffentliche Verbreitung von Daten entscheiden.

Als **sehr große Online-Plattformen** (very large online platforms, VLOPs) und sehr große Suchmaschinen (very large online search engines, VLOSEs) gelten laut DSA Dienste, die eine durchschnittliche monatliche Zahl von mindestens als 45 Millionen aktiven Nutzern in der EU haben (und somit mehr als 10 Prozent der 450 Millionen Verbraucher in der EU erreichen). Bislang wurden folgende 19 Unternehmen durch die EU-Kommission als sehr große Online-Plattformen klassifizier:

[3] Ausnahmen hierzu gelten vor allem für Online-Plattformen, die eCommerce betreiben.
[4] Vgl. Erwägungsgründe 21, 22 DSA
[5] Vgl. Richtlinie 2000/31/EG „Richtlinie über den elektronischen Geschäftsverkehr" Art. 14
[6] Vgl. Art. 14 DSA
[7] Vgl. Art. 3 Abs. 1 lit. i DSA

Tab. 1 Sehr große Online-Plattformen und -Suchmaschinen in der EU

Sehr große Online-Plattformen	Durchschnittliche Nutzer pro Monat in der EU (Unternehmensangaben)
Alibaba AliExpress	>45 Mio.
Amazon Store	>45 Mio.
Apple AppStore	>45 Mio.
Booking.com	k. A.
Facebook	255 Mio.
Google Play	274,6 Mio.
Google Maps	278,6 Mio.
Google Shopping	74,9 Mio.
Instagram	250 Mio.
LinkedIn	k. A.
Pinterest	k. A.
Snapchat	k. A.
TikTok	100,9 Mio.
Twitter/X	100,9 Mio.
Wikipedia	k. A.
YouTube	401,7 Mio.
Zalando*	27.449 million (76.247 million for retail service and platform service)
Sehr große Online-Suchmaschinen	**Durchschnittliche Nutzer pro Monat in der EU (Unternehmensangaben)**
Bing	107 Mio.
Google Search	278,6 Mio.

Quellen: Europäische Kommission 2023; Nutzerangaben nach Reuters, vgl. https://www.reuters.com/technology/google-twitter-meta-face-tougher-eu-online-content-rules-2023-02-17/, abgerufen am 31.08.2023
**Zalando: https://en.zalando.de/legal-notice/*
Hinweis: Zalando wehrt sich aktuell gegen die Einstufung als VLOP
Vgl.: https://www.handelsblatt.com/unternehmen/handel-konsumgueter/online-modehaendler-zalando-geht-mit-klage-gegen-strengere-eu-regulierung-vor/29227268.html, abgerufen am 09.10.2023

2.1.2 Anforderungen an Moderation von Inhalten gem. DSA

Der DSA gilt für sämtliche Vermittlungsdienste ab dem 17. Februar 2024. Für bereits von der EU-Kommission identifizierte sehr große Online-Plattformen und – Suchmaschinen gilt er bereits seit dem 25.08.2023.[8] Zukünftig neu benannten VLOPs wird jeweils eine Frist von 4 Monaten eingeräumt, um den Vorgaben des DSA zu entsprechen.[9]

[8] In diesem Zuge werden finden derzeit Anpassungen an den Nutzungsbedingungen der sehr großen Online-Plattformen statt, um dem Regulierungsrahmen des DSA zu entsprechen.
Vgl. https://www.facebook.com/legal/terms_preview_DSA, abgerufen am 20.09.23

[9] Vgl. https://digital-strategy.ec.europa.eu/de/policies/dsa-vlops, abgerufen am 22.09.23

Sämtliche Vermittlungsdienste müssen Anordnung von den zuständigen nationalen Justiz- oder Verwaltungsbehörden zum Vorgehen gegen „rechtswidrige Inhalte“ [10] Folge leisten. Diese „Anordnungen zum Vorgehen gegen rechtswidrige Inhalte“ müssen Informationen darüber enthalten, aus welchem Grund die betroffenen Inhalte illegal sind und den räumlichen Geltungsbereich der Anordnung darstellen. Je nach räumlichem Geltungsbereich muss der Vermittlungsdienst den Inhalt dann entweder vollständig löschen oder den Zugang dazu in bestimmten Ländern sperren. Der Vermittlungsdienst muss der Behörde mitteilen, wann der Anordnung Folge geleistet wurde. Direkte zeitliche Fristen sind mit der Anordnung jedoch nicht verbunden.[11] Eine allgemeine Verpflichtung zur proaktiven Überwachung von Nutzerbeiträgen oder aktiven Nachforschung besteht explizit nicht.[12]

Alle Vermittlungsdienste müssen in allgemeinen Geschäftsbedingungen den Nutzern darlegen, welche Beschränkungen in Bezug auf die von den Nutzern bereitgestellten Informationen gelten. Die Leitlinien, Verfahren, Maßnahmen und Werkzeuge, die zur Moderation von Inhalten eingesetzt werden, einschließlich der algorithmischen Entscheidungsfindung und der menschlichen Überprüfung, sowie die Verfahrensregeln des Dienste-internen Beschwerdemanagementsystems müssen darin transparent gemacht werden.[13]

Um ein angemessenes Maß an Transparenz und Rechenschaft zu gewährleisten, sind alle Vermittlungsdienste oberhalb der Schwelle von Klein- und Kleinstunternehmen im Sinne der Empfehlung 2003/361/EG der EU-Kommission dazu verpflichtet, jährlich einen öffentlichen Bericht über die von ihnen betriebene Moderation von Inhalten zur erstellen.[14] Hierbei muss in aggregierter Form dargestellt werden, auf welchem Weg welche Art von Verstößen (rechtswidrige Inhalte, Verstöße gegen Geschäftsbedingungen/Gemeinschaftsregeln) auf Basis welcher (Rechts-)Grundlage identifiziert wurden und wie diese Verstöße in welchen Median-Zeiträumen bearbeitet wurden.[15]

Dieser „Transparenzbericht“ soll auch eine qualifizierte Beschreibung automatisierter Mittel zur Inhaltemoderation enthalten, mit Angabe der genauen Zwecke, mit Indikatoren für die Genauigkeit und mit der Fehlerquote dieser automatisierten Mittel.[16]

Sämtliche Hostingdienste, ungeachtet ihrer Größe, müssen darüber hinaus leicht zugängliche und benutzerfreundliche Melde- und Abhilfeverfahren bereitstellen, die es erleichtern, dem Hostingdienst bestimmte Informationen zu melden, die eine meldende Partei als rechtswidrige Inhalte ansieht.[17]

---

[10] Als „rechtswidrige Inhalte“ werden alle Inhalte definiert, die „… nicht im Einklang mit dem Unionsrecht oder dem Recht eines Mitgliedstaats stehen“. Vgl. Art. 3 lit. h DSA

[11] Vgl. Art. 9 DSA

[12] Vgl. Art. 8 DSA

[13] Vgl. Art. 14 Abs. 1 DSA

[14] Vgl. Erwägungsgrund 49 DSA

[15] Zur Eindämmung von unerwünschten Inhalten können durch den Dienst, je nachdem ob es sich um einen „rechtswidrigen Inhalt“ oder „nur“ um einen Verstoß gegen eigene Gemeinschaftsregeln handelt, verschiedene Moderationsmaßnahmen ergriffen werden, z. B. geringe Verbreitung, Demonetisierung, Sperrung des Zugangs oder Entfernung eines Inhalts. Nutzerkonten können geschlossen, oder eine Verwarnung ausgesprochen werden („Strike“). Vgl. Art. 3 lit. t DSA

[16] Vgl. Art. 15 Abs. 1 lit. e DSA

[17] Vgl. Erwägungsgrund 50 DSA

Erfüllt ein Hostingdienst auch die Kriterien einer Online-Plattform, muss dieser darüber hinaus ein Beschwerdemanagement einrichten, um eine interne Streitbeilegung zu ermöglichen.[18] Zusätzlich müssen alle Online-Plattformen jeden einzelnen Fall einer Moderationsentscheidung die zu Einschränkungen für einen Nutzer führen (Herabstufung der Sichtbarkeit von Inhalten, räumliche Einschränkung der Sichtbarkeit, Entfernung von Inhalten, Sperrung von Nutzern) direkt nach Abschluss des Verfahrens an die neu eingerichtete DSA-Online-Transparency-Datenbank übermitteln.[19] Die nachfolgende Grafik zeigt bespielhaft, wie diese Entscheidungen an die Datenbank zu übermitteln sind:

Abb. 2 Beispielhafte Meldung von TikTok an die 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

Quelle: https://transparency.dsa.ec.europa.eu/statement/626604f4-ce5f-4cbd-945d-0a1eef5fac88, abgerufen am 22.09.23

Sehr große Online-Plattformen bergen besondere (systemische) Risiken für die Verbreitung illegaler Inhalte und für Schäden in der Gesellschaft.[20] Aus diesem Grund gelten für diese Unternehmen besonders strenge Vorgaben. Ihnen wurde die Pflicht auferlegt, ihre systemischen Risiken zu bewerten, einfache Meldewege zu etablieren sowie robuste Instrumente zur Moderation von Inhalten bereitzustellen. Sie müssen Risikominderungsmaßnahmen ergreifen – beispielsweise um der Verbreitung von Desinformation

[18] Vgl. Art. 20 Abs. 3 DSA

[19] Vgl: Art. 24 Abs. 5 DSA

[20] Vgl. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_de, abgerufen am 22.09.23

entgegenzuwirken – und ihr Risikomanagement extern auditieren lassen.[21] Die nachfolgende Tabelle gibt einen Überblick über die abgestuften Anforderungen des DSA auf Ebene der Vermittlungsdienste, Hostingdienste, Online-Plattformen und VLOPs.

Tab. 2: Neue Verpflichtungen für Dienste nach DSA

Neue (kumulative) Verpflichtungen	Vermittlungsdienste	Hostingdienste	Online-Plattformen	Sehr große Plattformen
Transparenzberichtspflichten	●	●	●	●
Meldung DSA Transparency Database			●	●
Zusammenarbeit mit nationalen Behörden bei Anordnungen	●	●	●	●
Angabe von Kontaktstellen und ggf. gesetzliche Vertretung	●	●	●	●
Meldung und Abhilfe sowie Pflicht zur Unterrichtung der Nutzer/innen	-	●	●	●
Meldung von Straftaten die „Gefahr für das Leben oder die Sicherheit einer Person" bedeuten können	-	●	●	●
Beschwerde- und Rechtsbehelfsmechanismus, außergerichtliche Streitbeilegung	-	-	●	●
Vertrauenswürdige Hinweisgeber	-	-	●	●
Maßnahmen gegen missbräuchl. Meldungen sowie Gegendarstellungen	-	-	●	●
Spezielle Pflichten für E-Commerce-Plattformen, z. B. Überprüfung der Berechtigungen von Drittanbietern, Compliance by Design, stichprobenartige Kontrollen	-	-	●	●
Verbot von Werbung, die sich gezielt an Kinder richtet oder spezielle personenbezogene Daten nutzt	-	-	●	●
Transparenz der Empfehlungssysteme	-	-	●	●
Transparenz von Online-Werbung gegenüber Nutzern	-	-	●	●
Verpflichtung zu Risikomanagement und Krisenreaktion	-	-	-	●
Externe und unabhängige Prüfung, interne Compliance-Funktion und öffentliche Rechenschaftspflicht (Audits)	-	-	-	●
Datenzugang für Digitale-Dienste-Koordinator, EU-KOM, zugel. Forscher	-	-	-	●
Möglichkeit für Nutzer, Empfehlungen anhand von Profiling abzulehnen	-	-	-	●
Verhaltenskodizes	-	-	-	●
Zusammenarbeit im Krisenfall	-	-	-	●

Quelle: Europäische Kommission: „Gesetz über digitale Dienste: mehr Sicherheit und Verantwortung im Online-Umfeld" (Stand September 2023)

[21] Vgl. https://ec.europa.eu/commission/presscorner/detail/de/ip_23_2413, abgerufen am 22.09.23

2.1.3 Anforderung an die Moderation von Inhalten gem. TCO-VO

Inhalte, die zu terroristischen Straftaten direkt oder indirekt anstiften, indem sie diese befürworten oder verherrlichen oder anderweitig zu deren Begehung beitragen[22], werden als rechtswidrige Inhalte grundsätzlich vom DSA mit umfasst. Die TCO-VO erweitert jedoch das Instrumentarium zur Eindämmung terroristischer Inhalte und stellt höhere Anforderungen an Hostingdienste in Bezug auf die Moderation terroristischer Inhalte als der DSA.

Durch die TCO-VO wurde das neuartige Instrument der „Entfernungsanordnung" geschaffen. Im Unterschied zu einer „Anordnungen zum Vorgehen gegen rechtswidrige Inhalte" gem. Art. 9 DSA sind Hostingdienste bei einer Entfernungsanordnung dazu verpflichtet, innerhalb einer Stunde nach Eingang der Entfernungsanordnung terroristische Inhalte zu entfernen oder den Zugang zu terroristischen Inhalten zu sperren.[23]

Die Frist zur Entfernung terroristischer Inhalte liegt deutlich unter vergleichbaren, bisherigen Zeitspannen, wie sie etwa bei Urheberrechtsverletzungen, Hatespeech oder kriminellen Inhalten bspw. im Rahmen des Digital Millenium Content Acts (DMCA), dem Urheberrechts-Diensteanbieter-Gesetz (UrhDaG), dem Netzwerkdurchsetzungsgesetz (NetzDG) oder dem Digital Service Act (DSA) bislang für die Entfernung von Inhalten gelten.

Die TCO-VO gilt grundsätzlich für Hostingdienste im Sinne des Art. 1 Abs. 1 lit. b der EU-Richtlinie 2015/1535 für Dienste der Informationsgesellschaft, wobei dessen Formulierung „in der Regel gegen Entgelt" von der Bundesnetzagentur so ausgelegt wird, dass auch solche Dienste, die ohne Gegenleistung angeboten werden, darunter fallen, sofern sie eine große Ähnlichkeit zu Diensten aufweisen, die üblicherweise gegen Entgelt erbracht werden. Maßgeblich ist nach der Auffassung der Bundesnetzagentur, dass vergleichbare Dienste derselben Dienstekategorie gewöhnlich gegen ein Entgelt erbracht werden. Das Merkmal ist beispielsweise auch erfüllt, wenn ein Nutzer dem Diensteanbieter im Gegenzug für die Leistung die Nutzung seiner Daten (etwa für eine Vermarktung an Werbetreibende) gestattet. Erfasst sind auch Fälle indirekter Finanzierung, z. B. in denen der Diensteanbieter den Dienst ohne direktes monetäres Entgelt aus Marketingzwecken oder zum Zweck der Werbung anbietet und durch andere Produkte oder Dienste quersubventioniert.

Die TCO-VO gilt auch für Klein- und Kleinstunternehmen im Sinne der in der Kommissionsempfehlung 2003/361/EG[24]. Allerdings soll die Unternehmensgröße bei etwaigen Sanktionen, neben weiteren Faktoren, berücksichtigt werden (Art. 18 Abs. 2 lit. f) TCO-VO).

In Deutschland ist das BKA die zuständige Behörde, die solche Entfernungsanordnungen erlassen kann. Zur Übermittlung von Entfernungsanordnungen soll zwischen BKA und Hostingdiensten, zuständigen Behörden in anderen EU-Ländern sowie Europol das IT-System PERCI[25] (vgl. Anhang Kap. 10.2) zum Einsatz kommen.

---

22 Vgl. Art. 2 Abs. 7 TCO-VO

23 Vgl. Art. 3 Abs. 3 TCO-VO

24 Weniger als 250 Beschäftigte und ein max. Jahresumsatz von höchstens 50 Mio. EUR oder eine Jahresbilanzsumme, die sich auf höchstens 43 Mio. EUR beläuft.

25 Plateforme Européenne de Retraits de Contenus Illégaux sur Internet

Inwiefern Hostingdienste bzw. „ausgesetzte" Hostingdienste in der Lage sind, fristgerecht auf Entfernungsanordnungen zu reagieren, ist bislang empirisch nicht zu beantworten. Im Jahr 2022 wurden durch das Bundeskriminalamt keine Entfernungsanordnungen auf der Grundlage von Art. 5 TCO-VO erlassen. Es gingen auch keine Entfernungsanordnungen ausländischer Stellen zur Prüfung der Rechtmäßigkeit beim BKA ein. Das IT-System PERCI, das für die Übermittlung von Entfernungsanordnungen zwischen den zuständigen Behörden und Hostingdiensten genutzt wird, ging erst im Sommer 2023 in Betrieb. Öffentliche Informationen zum Meldeaufkommen liegen noch nicht vor.

Da 2022 keine Entfernungsanordnungen erlassen wurden, gab es folgerichtig 2022 auch keine Hostingdienste, die als „terroristischen Inhalten ausgesetzt" eingestuft wurden. Somit wurden 2022 auch keine deutschen Hostingdienste zu „spezifischen Maßnahmen" nach Art. 5 TCO-VO verpflichtet. [26]

Das Bundeskriminalamt hat im Jahr 2022 hingegen 10.472 Referrals, Löschersuchen zur Entfernung oder Sperrung von Inhalten, an Hostingdienste übermittelt. Das Referral genannte Löschersuchen ist im Gegensatz zur Entfernungsanordnung rechtlich nicht verpflichtend. Löschungen infolge von Referrals erfolgen durch Hostingdienste auf freiwilliger Basis. Im Jahr 2022 kamen Hostingdienste in 88 Prozent der Fälle (9.207) den Löschersuchen des BKA durch Entfernung oder Sperrung von Inhalten nach.[27]

Das Ausbleiben von Entfernungsanordnungen in den verbleibenden Fällen lag u. a. daran, dass sich Löschersuchen auf Inhalte bezogen, die zwar strafrechtlich relevant waren, aber keine terroristischen Inhalte im Sinne der TCO-VO darstellten, oder dass Hostingdienste ihren Geschäftssitz außerhalb der EU hatten.[28]

2.1.4 NetzDG - Rückblick

In Deutschland war mit dem Netzwerkdurchsetzungsgesetz (NetzDG) bereits seit 2018 eine gesetzliche Verpflichtung zur Moderation von Inhalten in Kraft. Dem NetzDG unterlagen soziale Netzwerke[29], die in Deutschland mindestens zwei Millionen registrierte Nutzer haben.

Dem NetzDG fehlte das Instrument der behördlichen Anordnungen. Jedoch verpflichtete es die sozialen Netzwerke zu halbjährlichen Transparenzberichten. Darin war für spezifische Straftatbestände des Strafgesetzbuches darzulegen, in welcher Häufigkeit Beschwerden eingegangen sind und wie mit diesen Beschwerden verfahren wurde.[30]

Die Straftatbestände, die vom NetzDG umfasst werden, sind in § 3a Abs. 2 NetzDG enumeriert. Die folgende Tabelle gewährt einen Überblick. Einige der enumerierten Straftatbestände hatten einen direkten oder indirekten Bezug zu terroristischen Handlungen und damit zum Geltungsbereich der TCO-VO.

---

[26] ebd.

[27] ebd.

[28] ebd.

[29] Die Definition eines sozialen Netzwerks im NetzDG entspricht im Wesentlichen einer Onlineplattform im DSA (Vgl. § 1 Abs. 1 NetzDG).

[30] Vgl. § 2 NetzDG

Tab. 3 Enumerierte Straftatbestände im NetzDG

Paragraf	Straftatbestand	Bezug zu Terrorismus
§ 86 StGB	Verbreiten von Propagandamitteln verfassungswidriger Organisationen	indirekt
§ 86a StGB	Verwenden von Kennzeichen verfassungswidriger Organisationen	indirekt
§ 89a StGB	Vorbereitung einer schweren staatsgefährdenden Gewalttat	direkt
§ 91 StGB	Anleitung zur Begehung einer schweren staatsgefährdenden Gewalttat	direkt
§ 100a StGB	Landesverräterische Fälschung	indirekt
§ 111 StGB	Öffentliche Aufforderung zu Straftaten	indirekt
§ 126 StGB	Störung des öffentlichen Friedens durch Androhung von Straftaten	indirekt
§ 129 StGB	Bildung krimineller Vereinigungen	nein
§ 129a StGB	Bildung terroristischer Vereinigungen	direkt
§ 129b StGB	Kriminelle und terroristische Vereinigungen im Ausland; Einziehung	direkt
§ 130 StGB	Volksverhetzung	indirekt
§ 131 StGB	Gewaltdarstellung	indirekt
§ 140 StGB	Belohnung und Billigung von Straftaten	indirekt
§ 166 StGB	Beschimpfung von Bekenntnissen, Religionsgesellschaften und Weltanschauungs-vereinigungen	nein
§ 184b i. V. m. § 184d StGB	Verbreitung, Erwerb und Besitz kinderpornographischer Inhalte / Zugänglichmachen pornographischer Inhalte mittels Rundfunk oder Telemedien; Abruf kinder- und jugendpornographischer Inhalte mittels Telemedien	nein
§ 185 StGB	Beleidigung	nein
§ 186 StGB	Üble Nachrede	nein
§ 187 StGB	Verleumdung	nein
§ 201a StGB	Verletzung des höchstpersönlichen Lebensbereichs und von Persönlichkeitsrechten durch Bildaufnahmen	nein
§ 241 StGB	Bedrohung	nein
§ 269 StGB	Fälschung beweiserheblicher Daten	nein

Quelle: Goldmedia auf Basis § 3a Abs. 2 NetzDG

Während das NetzDG nur eine Moderation im Hinblick auf bestimmte Strafgesetze vorsieht, wird der DSA die Pflicht einführen, sämtliche rechtswidrigen Inhalte zu moderieren. Dies umfasst grundsätzlich die gesamte Rechtsordnung.

Durch den Regelungsentwurf zum Digitale Dienste Gesetz (DSA) werden das Netzwerkdurchsetzungsgesetz (NetzDG) und das Telemediengesetz (TMG) außer Kraft gesetzt (vgl. Artikel 37 DDG-Ermächtigungsgesetz). Für sehr große Online-Plattformen ist das NetzDG aktuell bereits nicht mehr gültig.

2.2 Hostingdienste mit wesentlicher Verbindung zu Deutschland

Hostingdienste, die in der Europäischen Union anbieten und Informationen öffentlichen verbreiten, unterfallen der TCO-VO.[31] Besteht hierbei eine wesentliche Verbindung zu Deutschland, z. B. aufgrund seiner Niederlassung, einer erheblichen Zahl von Nutzern in Deutschland oder einer spezifischen Ausrichtung auf den deutschen Markt, sind grundsätzlich deutsche Behörden wie das BKA und die BNetzA[32] zuständig.[33]

2.2.1 Arten von Hostingdiensten

Im Folgenden wird ein grober Überblick über Hosting Dienste gegeben, die eine wesentliche Verbindung zu Deutschland haben. Tab. 4 bietet einen Überblick über die 50 reichweitenstärksten Hostingdienste in Deutschland aus dem Jahr 2022. Hierbei wird bereits die Themenvielfalt unter den reichweitenstärksten Hostingdiensten in Deutschland deutlich.

Tab. 4 Top-50 der meistaufgerufenen Hostingdienste in Deutschland 2022

Domain	Nutzer pro Monat in Mio., Summe Januar - Dezember 2022	Durchschnittliche Nutzer in Mio. pro Monat 2022
google.com	890,6	74,2
youtube.com	300,2	25,0
facebook.com	281,3	23,4
amazon.de	246,2	20,5
wikipedia.org	232,8	19,4
google.de	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
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
wetter.com	41,7	3,5
pornhub.com	40,4	3,4
n-tv.de	40,4	3,4
sport1.de	33,8	2,8

[31] Vgl. Art. 1 Abs. 2 TCO-VO
[32] Vgl. § 1 TerrOIBG
[33] Vgl. Art. 2 Abs. 5 TCO-VO

Domain	Nutzer pro Monat in Mio., Summe Januar - Dezember 2022	Durchschnittliche Nutzer in Mio. pro Monat 2022
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.de	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
pressekompass.net	23,3	1,9
kaufland.de	23,2	1,9
tz.de	23,2	1,9

Quelle: Semrush 2022: Die meistbesuchten und meistaufgerufenen Websites in Deutschland 2022

Um die Vielfalt und Komplexität dieser Hostingdienste näher zu erfassen, wurden diese im Folgenden in unterschiedliche Kategorien unterteilt, die jeweils eine spezifische Funktion und Ausrichtung abbilden. Die Kategorien sind in DSA-Oberkategorien und Dienstekategorien unterteilt, wobei Oberkategorien Hostingdienste, Online-Suchmaschinen und Online-Plattformen umfassen. Dienstekategorien enthalten in der Regel mehrere Hostingdienste bzw. Online-Plattformen ähnlicher Funktion, die sich in Größe, Nutzerbasis und Funktionalitäten bisweilen stark voneinander unterscheiden können.

Tab. 5 Übersicht über Hostingdienste nach DSA-Systematik

DSA-Oberkategorie	Dienstekategorien
Hostingdienste	
Hostingdienste, die keine Online-Plattformen darstellen	Web-Hosting (inkl. File-Hosting und Website-Baukästen)
	E-Mail-Dienste, Kurznachrichtendienste und andere interpersonelle Kommunikationsdienste
	Cloud Computing (Hardware-, Computing-Power- und Security-as-a-Service)
	Cloud Storage
	Kollaborationsplattformen (Software-as-a-Service)
Online-Suchmaschinen	Online-Suchmaschinen
Online-Plattformen, Hosting-Dienste die Inhalte der Kunden/Nutzer veröffentlichen	
Online-Plattformen für Fernhandels-absatz	Online-Handelsplätze
	VoD-Plattformen (inkl. Transaktionsbasiertes VoD)
	Audiostreaming-Plattformen (Spotify)
	E-Book-/E-Journal-Plattformen
	App-Stores
	Plattformen kollaborativer Wirtschaft: Marktplätze (Unterkünfte, Kleinanz., Automobile, Immobilien, Ärzte, Handwerker, Crowdfun-ding …) und Sharing Economy
	Partnerschaftsbörsen/Dating-Plattformen
	Datenmarktplätze
	…
Online-Plattformen (soziale Medien)	Generische soziale Netzwerke (Facebook, Instagram, TikTok …)
	Social Business Networks (LinkedIn, XING …)
	Social-Video-Plattformen inkl. Live-Streaming (YouTube, Twitch …)
	Social-Audio-Plattformen (Soundcloud, …)
	Kurznachrichtendienste/Micro-Blogging (X …)
	Special-Interest-Communities/Foren/Image Boards (u. a. zu Gaming, Q&A, DIY, Bücher, Musik, Rezepte, Sport, Kunst …)
	Social Gaming (Roblox, Steam …)
	Dating
	Social Inspiration (Pinterest, Flickr …)
	Online-Enzyklopädien
	Location Based Services (Google Maps, Yelp, Tripadvisor …)
	Dokumentenportale (Scribd, Doc-Player, Slide-Player …)
	…

Quelle: Goldmedia-Analyse 2023

Je nach Ausrichtung der Plattform unterscheiden sich die Anforderungen an die Moderation von Inhalten. Plattformen, auf denen Politik oder kontroverse gesellschaftliche Themen diskutiert werden, sind in der Regel moderationsaufwendiger als etwa Reise-, Rezept-, oder Karriereportale. Auch Online-Plattformen mit einem Schwerpunkt auf Gaming können moderationsintensiv sein, allerdings hängt es dies stark vom jeweiligen Spieltitel und dessen Spielkultur ab. Plattformen, die sich an Kinder und Jugendliche richten, sind ebenfalls moderationsintensiv, da neben den üblichen moderativen Aufgaben ein altersgerechtes Umfeld aufrecht herhalten werden muss. Bei Online-Plattformen

für Fernhandelsabsatz besteht der Moderationsaufwand hingegen vor allem in der Betrugsvorbeugung und -bekämpfung und weniger in der Moderation von gewalttätigen oder hasserfüllten Inhalten.

Tab. 6 Monatliche Reichweite ausgewählter Domains (Desktop und online) nach Similarweb im August 2023, in Mio. Visits

Dienstekategorie	Online-Plattform	Mio. Visits pro Monat
Social-Video-Plattform	YouTube.com	33.900,0
Soziales Netzwerk	Facebook.com	17.400,0
Soziales Netzwerk	Instagram.com	6.700,0
Micro-Blogging-Plattform	Twitter.com/X	6.400,0
Social-Commerce-Plattform	Amazon.com	2.500,0
Social-Video-Plattform	TikTok.com	2.300,0
Micro-Blogging-Plattform	Reddit.com	1.900,0
Social-Video-Plattform	Twitch.tv	1.100,0
Social-Gaming-Plattform	Roblox.com	866,2
Q&A-Plattform	Quora.com	712,6
Social-Commerce-Plattform	Amazon.de*	445,8
Micro-Blogging-Plattform	Tumblr.com	217,0
Soziales Netzwerk	Snapchat.com	187,6
Social-Gaming-Plattform	Steampowered.com	161,2
Musikplattform	SoundCloud.com	126,0
Social-Video-Plattform	Vimeo.com	73,3
Q&A-Plattform	Gutefrage.net	40,5
Social-Video-Plattform	Bitchute.com	17,9
Special-Interest-Forum	moviepilot.de	17,4
Special-Interest-Forum	Gamestar.de	12,3
Special-Interest-Forum	Computerbase.de	9,6
Micro-Blogging-Plattform	Mastodon.social	4,6
Soziales Netzwerk	Nebenan.de	4,5
Soziales Netzwerk	Jappy.com	2,4
Soziales Netzwerk	StayFriends.de	2,4
Special-Interest-Forum	Boersennews.de	1,9
Q&A-Plattform	wer-weiss-was.de	1,1
Social-Video-Plattform	YouNow.com	0,8
Special-Interest-Forum	Musiker-board.de	0,7
Special-Interest-Forum	Worldofplayers.de	0,7
Special-Interest-Forum	Bvb-forum.de	0,6

** zusätzliche Angabe der DE-Top-Level-Domain zur COM-Top-Level-Domain*
Quelle: https:www.similarweb.com, abgerufen am 15.09.23

2.2.2 Weitere Hostingdienste

Hostingdienste, die im Auftrag von Online-Plattformen das technische Hosting realisieren, sind von der Haftung der von ihnen bereitgestellten Inhalte weitgehend ausgeschlossen, solange sie die Integrität der bereitgestellten Informationen nicht verändern. Hierzu zählen etwa Web- und Filehoster, Rechenzentren, Cloud-Storage-Provider, CDN-Dienste und vergleichbare Vermittlungsdienste. Sie müssen selbst keine Inhalte moderieren.

Allerdings sind auch diese Hostingdienste, die im Auftrag anderer Hostingdienste tätig werden, dazu verpflichtet, zügig Inhalte zu entfernen oder den Zugang dazu zu sperren, sobald sie Kenntnis über rechtswidrige Tätigkeiten oder Inhalte erhalten.

3 Status Quo der Inhaltemoderation

3.1 Einführung

Basis der Inhaltemoderation sind die Allgemeinen Geschäftsbedingungen der Dienste und der darin verankerten Gemeinschaftsrichtlinien („Community Standards", „Netiquette", „Code of Conduct" etc.). In der Regel gehen die Regelungen, die in den Gemeinschaftsrichtlinien eines Dienstes verankert sind, deutlich über die gesetzlichen Anforderungen hinaus. Rechtswidrige Inhalte werden daher häufig bereits über die allgemeinen Prozesse moderiert, welche die Einhaltung der Gemeinschaftsrichtlinien eines Dienstes sicherstellen.[34]

Die Zulässigkeit eines Inhaltes kann sich dabei von Dienst zu Dienst stark unterscheiden, abhängig von der Ausrichtung und der Zielgruppe des Dienstes. Ein Dienst, der sich etwa vornehmlich an Kinder und Jugendliche richtet, gibt sich in der Regel engere Moderationsvorgaben, insbesondere in Bezug auf vulgäre Sprache oder sexuelle Inhalte, als etwa ein Dienst, der sich vornehmlich an Erwachsene richtet.

Spezifische Bereich bieten hierbei besondere moderative Herausforderungen. So kommt es in In-Game-Chats bestimmter Computerspiele, die stark kompetitiv sind, gehäuft zu Hass, und menschenverachtenden Äußerungen. Im Bereich der an Kinder und Jugendliche gerichteten Dienste ist Cybergrooming eine der größten Herausforderungen.

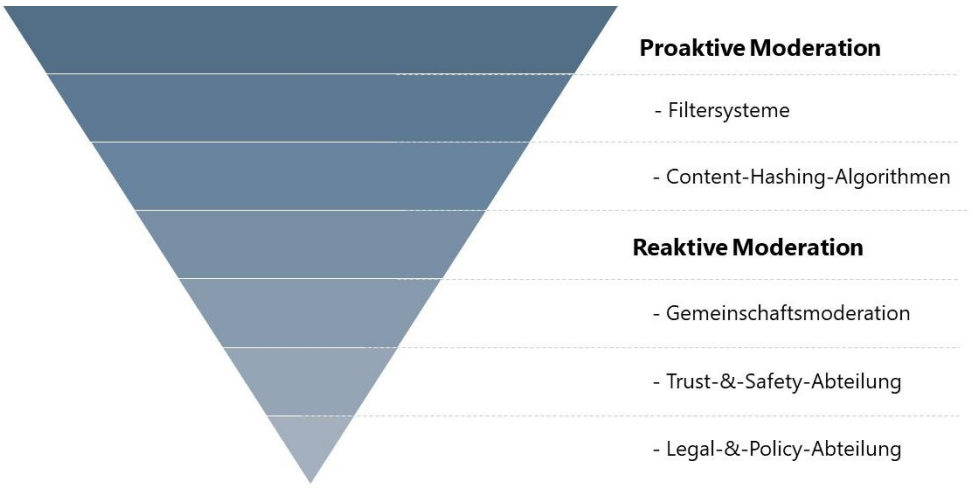
Während einfach identifizierbare Verletzungen der Gemeinschaftsrichtlinien in Text-, Bild- oder Videoform je nach Plattform bereits durch Upload-Filter erkannt und ohne manuelle Nachprüfung blockiert werden, erfolgt die Moderation komplexerer Inhalte (z. B. längere Textbeiträge zu tendenziell kritischen Themen), die durch die Moderationssoftware „geflaggt" oder von Nutzern gemeldet werden, teilweise über einen mehrstufigen Prozess.

Dieser Prozess verläuft über Einzelentscheidungen eines Moderators über Einholung von Zweit- und Drittmeinung weiterer Themenfeldexperten innerhalb des Trust-and-

[34] Hinzu kommen im Bedarfsfall spezialisiertere Moderationsteams, welche die Einhaltung (landes-)spezifischer Moderationsvorgaben sicherstellt. So werden Meldungen nach NetzDG von spezialisierten NetzDG-Teams moderiert, solange der Meldende innerhalb der Meldung den Bezug zum NetzDG herstellt.

Safety-Teams bis hin zu juristischen Prüfschritten, die zum Teil wiederum auf mehreren Ebenen erfolgen können.[35] So werden bei einigen Anbietern besonders sensible Entscheidungen nicht mehr durch das Trust-and-Safety-Team, sondern durch das übergeordnete Legal-and-Policy-Team getroffen, zum Teil erst, nach Rücksprache mit einer externen Rechtsberatung.

Abb. 3 Prozess der Inhaltemoderation (schematische Darstellung)



Quelle: Goldmedia Analyse 2023

Vorgaben zur Moderation müssen periodisch überarbeitet werden, da die Dienste ständig durch neue Formen von missbräuchlichen Inhalten, z. B. im Bereich Desinformation, herausgefordert werden. Bei größeren Diensten findet dies in strukturierten Aktualisierungsprozessen tlw. mit juristischer Beratung durch das Legal-and-Policy-Team statt. Wobei der Austausch über neuartige Herausforderungen auf den verschiedenen Ebenen der Content-Moderation auch horizontal stattfindet, um hinreichend agil auf neuartigen Moderationslagen reagieren zu können. Bei neuartigen Moderationslagen kann es daher am ehesten zu nicht-harmonisierten Moderationsentscheidungen kommen, insbesondere wenn verschiedene Teams an verschiedenen Standorten bzw. Ländern involviert sind. Die Sicherstellung einer einheitlichen Anwendung der Gemeinschaftsrichtlinien ist dann die Aufgabe des Legal-and-Policy-Teams.

3.1.1 Proaktive Moderationsverfahren

Im Bereich der proaktiven Moderationsverfahren werden schwerpunktmäßig automatisierte Systeme eingesetzt, wie sie auch in Kapitel 3.3.1 beschrieben sind. Hierbei kann unterschieden werden zwischen vollautomatischen Filter-Systemen, die eine Veröffentlichung eines Beitrags direkt unterbinden und Filtern, die Beiträge für eine nachträgliche manuelle Überprüfung „flaggen".

[35] Im Falle des Dienstes YouTube reicht der Eskalationsprozess von der Befassung der eigenen Rechtsabteilung über die Einbeziehung der übergeordneten Rechtsabteilung von Google Germany bis zur Hinzuziehung externer strafrechtlicher Rechtsberatung (Vgl. Google 2023: YouTube Transparenzbericht 2. Halbjahr 2022).

Auch wenn zahlreiche Moderationstools durch Dienstleister kommerziell angeboten werden, setzen die meisten Dienste entweder auch oder sogar schwerpunktmäßig selbst entwickelte Systeme ein. Die Gründe hierfür sind vielfältig: Zum Teil sind die Dienste bereits länger am Markt als Lösungen von Dienstleistern, zum Teil sind die Dienste auch zu spezifisch in ihren Inhalten, als das eine generische Moderationslösung von großem Nutzen wäre. Ein weiterer wesentlicher Punkt ist auch, dass Dienste, die Moderationslösungen benötigen, in der Regel über hinreichende interne IT-Entwicklungskapazitäten verfügen, um entsprechende Lösungen passgenau für ihre eigenen Zwecke zu entwickeln. Im besten Falle geschieht dies in einem Safety-by-Design-Ansatz auch bereits vor bzw. parallel zur Entwicklung von nutzerweisenden Dienstemerkmalen.

Manuelle Moderation durch Menschen findet teilweise ebenfalls als proaktives Moderationsverfahren Anwendung, auch wenn der Stellenwert hier gegenüber den automatisierten Systemen deutlich zurücksteht. Manuelle Moderation dient hierbei vor allem zur Beobachtung allgemeiner Trends und Entwicklung auf der Plattform, insbesondere in sensiblen, moderationsaufwendigen Bereichen, wie etwa Politik, oder bei besonderen Lagen.

3.1.2 Reaktive Moderationsverfahren

Im Bereich der reaktiven Moderation wird zum Zeitpunkt der Studienerstellung ausschließlich manuell, das heißt durch menschliche Entscheidungen der sog. „Content-Reviewer", moderiert.

Reaktive Moderationsverfahren setzen voraus, dass ein Inhalt

a) entweder von einem automatisierten System mit hinreichender Wahrscheinlichkeit als problematisch eingestuft wurde (vgl. Kap. 3.1.1),
b) oder dass ein Inhalt gemeldet wurde.

Meldungen erfolgen dabei entweder durch die Nutzer bzw. die Community, durch Behörden oder durch Meldestellen.

Nutzer melden problematische Inhalte über die eigens auf der Webseite des Hostingdienstes eingerichtete Kontaktadresse oder über eine dafür eingerichtete Eingabemaske. Dienste, die Meldungen über Eingabemasken vorsehen, nutzen diese zur Erstklassifizierung der Beschwerden, indem der Nutzer gebeten wird, den jeweilig betroffenen Rechtsbereich der Meldung anzugeben. In Abhängigkeit von der Erstklassifikation des Meldegrunds werden je nach Größe des Hostingdienstes unterschiedliche Moderationsprozesse („Cues") ausgelöst und bspw. Meldung direkt einem mit dem Rechtbereich betrauten/spezialisierten Moderatorenteam übergeben.

Aber auch der Meldeursprung definiert, wie mit einer Meldung umgegangen wird:

Behörden (Public Flagger) kommunizieren mit den Hostingdiensten in der Regel über eigene Schnittstellen oder Kontaktstellen (Beispiele: Kontaktstelle des BKA/Kontaktschnittstelle PERCI oder die BNetzA-Liste der Gesetzlichen Vertreter gem. TCO-VO).

Meldungen von Nutzern, die einen bestimmten Status oder eine Moderatorenrolle in einer Community haben oder die in der Vergangenheit vielfach verlässliche Meldungen erbracht haben, werden priorisiert behandelt. Teilweise verfügen diese Nutzer als Nutzermoderatoren auch über dedizierte Meldekanäle zum Hostingdienst oder haben sogar eingeschränkten Zugang zum Moderationssystem.

Neben diesen „Private Flaggern" weisen viele Plattformen auch den im Markt agierenden NGOs, die sich gegen Hatespeech oder Kindesmissbrauch engagieren, einen besonderen Status zu. Das „YouTube Priority Flagger Program" etwa priorisierte neben Meldungen von Behörden (Public Flagger) auch Meldungen von NGOs, die in der Vergangenheit bereits besonders effektiv darin waren, YouTube Inhalte zu melden, die gegen die Gemeinschaftsrichtlinien verstoßen.[36] Im Jahr 2020 waren in Deutschland rund 30 Organisationen Teil des YouTube Priority Flagger Program, weltweit waren es um die 180 Organisationen.[37]

Der DSA gibt nun in Art. 22 vor, dass die staatliche Stelle des Digitale Dienste Koordinators (DDK/DSC) „Trusted Flagger" benennen darf. Diesen öffentlich benannten Trusted Flaggern werden im DSA Transparenzpflichten auferlegt.

Tab. 7 Kategorien von privaten und öffentliche vertrauenswürdige Hinweisgebern

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
Quelle: Appelman, N. & Leerssen., P. (2022) „On „Trusted" Flaggers". Yale-Wikimedia Initiative on Intermediaries & Information, online unter: https://law.yale.edu/sites/default/files/area/center/isp/documents/trustedflaggers_ispessayseries_2022.pdf, abgerufen am 07.09.23

Durch diese hervorgehobene Stellung im DSA wird die Arbeit vertrauenswürdiger Hinweisgeber weiter aufgewertet.

Die Liste der vom DDK benannten „Trusted Flagger" wird voraussichtlich hohe Schnittmenge mit den von den Hostingdiensten bereits identifizierten „Trusted Flaggern" aufweisen. Im weiteren Verlauf der Studie wird der Begriff „Trusted Flagger" jedoch nicht im Sinne des DSA, sondern im Sinne der durch die jeweiligen Dienste bereits als vertrauenswürdig und zuverlässig identifizierten Meldestellen und meldenden Einzelpersonen genutzt.

Der Anhang dieser Studie beschreibt alle für deutsche Hostingdienste relevanten Behörden und Meldestellen.

Die Sicherstellung einheitlicher Moderationsentscheidungen ist dabei eine wesentliche Herausforderung für die Dienste, insbesondere wenn es sich um sehr große Online-Plattformen handelt, die ihre Dienste global anbieten und aus verschiedenen Ländern moderieren lassen. Die Qualitätssicherung von Moderationsentscheidungen ist daher

[36] Vgl. YouTube-Hilfe „Das YouTube Priority Flagger Programm", online unter: https://support.google.com/youtube/answer/7554338?hl=de#:~:text=Das%20YouTube%20Priority%20Flagger%20Program%20bietet%20leistungsf%C3%A4hige%20Tools%20f%C3%BCr%20Beh%C3%B6rden,gegen%20die%20Community%2DRichtlinien%20versto%C3%9Fen, abgerufen am 18.09.23

[37] Vgl. HateAid „Trusted Flagger", online unter: https://hateaid.org/trusted-flagger/, abgerufen am 08.09.23

ein integraler Bestandteil des Moderationsprozesses, auch bei kleineren Anbietern. Die Qualitätssicherung wird in der Regel durch besondere Teams durchgeführt, welche sich aus erfahrenen Moderatoren zusammensetzen.[38]

3.1.3 Moderationsentscheidungen

Je nach Schwere eines festgestellten Verstoßes gegen die Gemeinschaftsrichtlinien können Moderatoren den Inhalt oder den Nutzer, der den Inhalt eingestellt hat, sanktionieren. Hierfür legt der Dienste-Anbieter fest, wie eine angemessene Reaktion auf einen Verstoß aussieht, der Ermessenspielraum eines Moderators ist hierbei gering.

Die Ausgestaltung des genauen Sanktionsregimes unterscheidet sich von Plattform zu Plattform. Einige grundlegende Prinzipien und Instrumentarien finden sich jedoch einen oder anderen Gestalt bei allen Diensten: Zunächst kann die algorithmische Verbreitung eines Inhaltes eingedämmt oder unterbunden werden, falls der Inhalt zwar unerwünscht ist, aber nicht direkt gegen die Gemeinschaftsrichtlinien verstößt. Solche Inhalte können dann weiterhin noch für bestimmte (private) Nutzergruppen zugänglich sein, werden aber nicht mehr kommerziell vermarktet. Inhalte können darüber hinaus auch von der Plattform gelöscht werden, falls der Inhalt gegen die Gemeinschaftsrichtlinien oder gesetzliche Vorgaben verstößt.

3.1.4 Umgang mit unzulässigen Inhalten

Wenn Inhalte gegen die Gemeinschaftsrichtlinien eines Dienstes verstoßen, werden sie entfernt und der postende Nutzer über die Entfernung in Kenntnis gesetzt. Hierbei ist es üblich, dass individuelle Verstöße eines Nutzers gezählt werden und ggf. weitergehende Sanktionen erlassen werden. Abhängig davon, gegen welche Richtlinien die Inhalte verstoßen und abhängig davon, welche und wie viele Verstöße bzw. Verwarnungen ein Nutzer bereits erhalten hat, kann das Konto in seiner Funktionalität eingeschränkt und vorübergehendend oder dauerhaft deaktiviert werden.

Inhalte, die nicht gegen Gemeinschaftsrichtlinien eines Dienstes verstoßen, aber dennoch problematisch oder anderweitig von geringer Qualität sind, können in ihrer Verbreitung eingeschränkt werden. Zudem nutzen Plattformen vorgeschaltete Hinweise auf potenziell sensible oder irreführende Inhalte, auch wenn sie nicht ausdrücklich gegen die Gemeinschaftsrichtlinien eines Dienstes verstoßen, um zusätzlichen Kontext zu solchen Inhalten zu liefern.

3.1.5 Streitbeilegung

Online-Plattformen unterhalten Prozesse zur Streitbeilegung, wenn Nutzer der Meinung sind, dass Inhalte ungerechtfertigt entfernt bzw. nicht zur Veröffentlichung freigegeben wurden oder sie ungerechtfertigt verwarnt wurden. Bei Diensten, die dem NetzDG unterliegen, behandeln gem. Gesetzeslage Beschwerden gegen Moderationsentscheidungen auf Basis von NetzDG anders als auf Basis der übrigen Gemeinschaftsrichtlinien. Support-Postfach. Bei Verstößen gegen eine im NetzDG aufgeführte Bestimmung erhalten Nutzer eine E-Mail, aus der die Rechtsnorm des deutschen Strafgesetzbuchs, gegen die verstoßen wurde, sowie die von der Plattform ergriffen Maßnahmen, hervorgeht.

[38] Beispiel YouTube: Ungefähr 30 Prozent der geprüften Inhalte werden durch Qualitätssicherungsteams überprüft. Vgl. Google 2023: YouTube Transparenzbericht 2. Halbjahr 2022

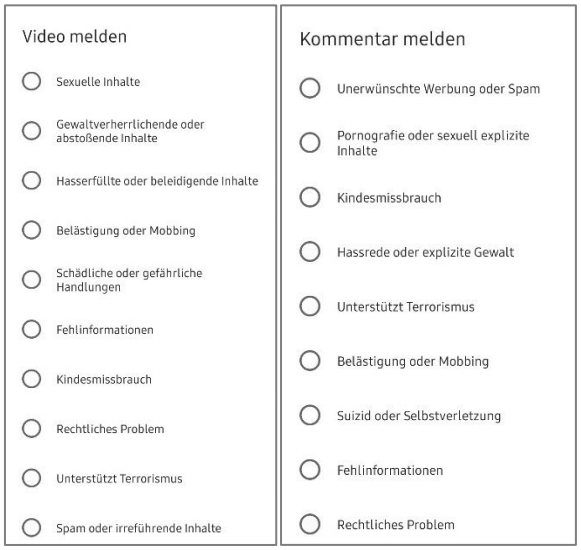
Nach Art. 17 DSA informiert Facebook neuerdings alle Nutzer, deren Inhalte als rechtswidrig oder als Verstoß gegen Gemeinschaftsrichtlinien eingestuft wurden, per Mail über die erfolgte Moderationsmaßnahme inkl. Begründung. Zudem gewährt Facebook gem. Art. 20 DSA den betroffenen Nutzern und bei gemeldeten Inhalten auch den meldenden Personen ab dem Eingang der Beschwerde für sechs Monate Zugang zum internen Beschwerdemanagementsystem. Über dieses können Beschwerden gegen die Entscheidung des Dienstes eingereicht werden.

3.2 Operativer Prozess der Inhaltemoderation

Durch die Inhaltemoderation werden von den Gemeinschaftsrichtlinien als unerwünscht definierte Inhalte sowie rechtswidrige Inhalte entfernt, bzw. nicht zur Veröffentlichung freigegeben, oder der Zugang wird eingegrenzt/herabgestuft und so die (weitere) Verbreitung verhindert. Eine echte „Löschung" erfolgt jedoch frühestens nach Ablauf der im DSA vorgegebenen Einspruchsfrist von sechs Monaten.[39]

Zur besseren Bearbeitung werden die Moderationstatbestände zunächst in Kategorien eingeteilt. Oft geschieht die erste Eingruppierung bereits im Prozess der Meldung durch den Meldenden. Hierfür werden kompakte, intuitiv verständliche Taxonomien von bis zu 10 Klassen verwendet. In Abhängigkeit von der Erstklassifikation des Meldegrunds durch den Meldenden können unterschiedliche Moderationsprozesse („Cues") ausgelöst werden.

Abb. 4 YouTube-Meldefunktionen für Videos und Kommentare



Quelle: YouTube App-Version 18.29.33, Screenshots aus dem September 2023

Insbesondere größere Plattformen haben unterschiedliche Prozesse in Abhängigkeit von der Art der Meldung, die anschließend von unterschiedlichen, spezialisierten Teams bearbeitet werden. Innerhalb der Teams findet in der Regel eine detailliertere Eingrup-

[39] Vgl. Art. 20 DSA

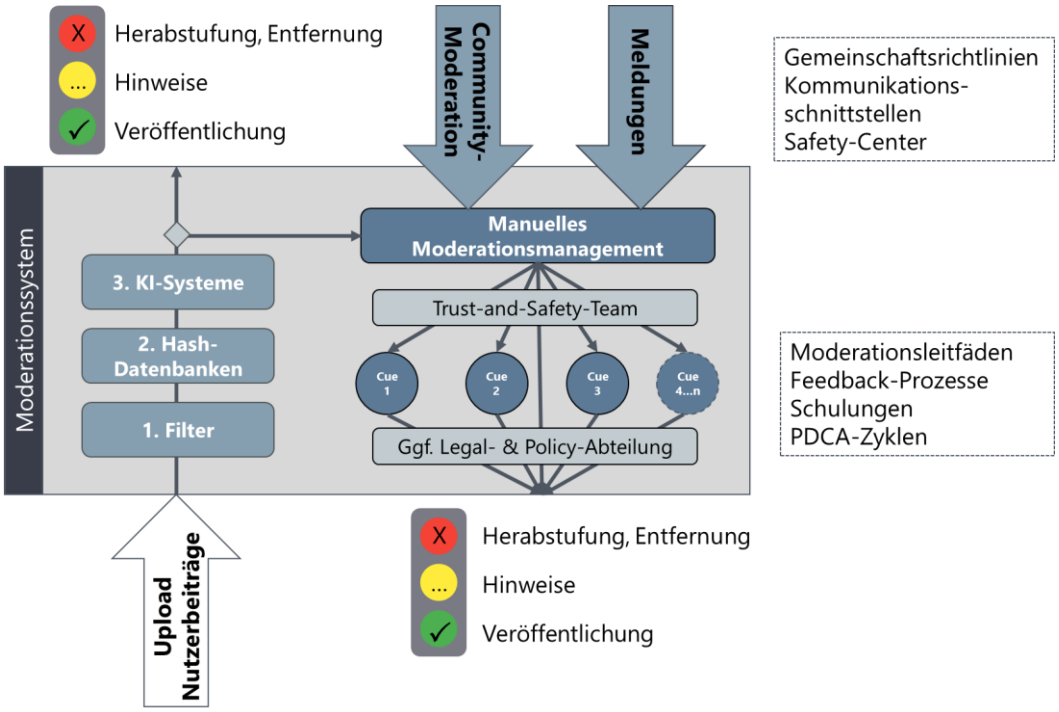
pierung der Meldung statt. Die Taxonomien unterscheiden sich von Anbieter zu Anbieter, aber eine Unterscheidung in rd. 30-50 Oberkategorien und bis zu 100 Unterkategorien wurde im Rahmen der Gespräche für diese Studie als branchenüblich bestätigt.

Im Folgenden wird eine vereinfachte, verallgemeinerte Moderationsarchitektur beschrieben, die strukturell auf sämtliche Online-Plattformen anwendbar ist (vgl. Abb. 5). Grundlegend ist hierbei zu berücksichtigen, dass die Entwicklung von Gemeinschaftsrichtlinien von deren Anwendung unterschieden werden muss. Die Entwicklung von Gemeinschaftsrichtlinien erfolgt zwingend durch den Dienst, in der Regel gibt es hierfür eine eigene Abteilung, die Community, Community-Leitung oder auch Legal-and-Policy heißen kann.

Die Anwendung der Gemeinschaftsrichtlinien erfolgt hingegen in der Regel in der Trust-and-Safety-Abteilung, hier werden sämtliche Verfahren der Inhaltemoderation operativ gesteuert. Zur Erfüllung der Trust-and-Safety-Aufgaben kommen automatisierte Verfahren und manuellen Verfahren zum Einsatz. Bei beiden ist es üblich, dass sowohl interne Lösungen bzw. Mitarbeiter als auch externe Dienstleistungen eingesetzt werden. Zumindest bei großen bis sehr großen Online-Plattformen überwiegt jedoch der Einsatz interner Lösungen bei technischen Verfahren und es überwiegt der Einsatz externer Dienstleister bei manuellen Verfahren. Organisatorisch gliedern sich die Moderatoren bei größeren Diensten in Teams, die eine bestimmte Moderationsaufgabe („Cues“) erfüllen. Bei kleineren Anbietern gibt es nur ein Moderationsteam und einen geringeren Spezialisierungsgrad der Moderatoren.

Ausgangspunkt der Inhaltemoderation sind sämtliche Inhalte, die durch Nutzer auf einen Online-Dienst gespeichert werden. Der Moderationsprozess wird hierbei von einer speziellen Moderations- bzw. Moderationsmanagement-Plattform gesteuert. Hierbei kann es sich um umfangreiche Enterprise-Softwarelösungen handeln, die insbesondere bei großen Online-Plattformen cloudbasiert arbeiten und sowohl die Integration technischer Subsysteme (z. B. Filter-Systeme) ermöglichen, als auch die grafischen Benutzeroberflächen für die menschlichen Moderatoren bereitstellt. Der gesamte Prozess, inklusive verschiedener Eskalationsstufen und Revisionsverfahren wird durch die zentrale Plattform abgebildet und protokolliert. Je nach Ausgestaltung des Prozesses kann es mitunter dazu kommen, dass besonders kritische Moderationsentscheidungen über das Saftey-Team hinaus bis zum Legal-and-Policy-Team zur Klärung eskaliert werden, etwa bei besonders weitreichenden Moderationsentscheidungen mit Policy-Implikationen für Dienst.

Abb. 5 Beispielhafte Moderationsarchitektur einer Online-Plattform



Quelle: Goldmedia Analyse 2023

Wie in der schematischen Darstellung erkennbar ist, lässt sich das zentrale Moderationssystem um beliebige technische Filter- und Hilfssysteme modular erweitern. Hierbei bleibt es unerheblich, ob es sich um eine Eigenentwicklung oder eine eingekaufte Lösung eines Drittanbieters handelt. Die Einbindung von weiteren (Sub-)Systemen kann aus verschiedenen Gründen notwendig werden, etwa um ein KI-basiertes System auf Basis der eigenen Inhalte und Moderationsentscheidungen zu trainieren oder um weitere Dienstleister für Moderationsdienstleistungen (manuelle Moderationsverfahren) einzubinden.

Mit Blick auf die **eingesetzten technischen Systeme** arbeitet die Mehrzahl der für die Studie gesprochenen Online-Dienste mit eigenentwickelten Lösungen, zumindest in Teilbereichen.[40] Hierbei kommen mehrheitlich Filtersysteme ohne KI-Unterstützung oder eigene Moderationsmanagementlösungen zum Einsatz. In Einzelfällen werden auch bereits selbst angepasste KI-Lösungen getestet. Insbesondere die sehr großen Online-Plattformen arbeiten mit selbst entwickelten Moderationsarchitekturen, die auf die eigenen Bedürfnisse ausgerichtet sind. Externe Dienstleister werden an diese Systeme angebunden.

Mit dem ständig wachsenden Aufkommen an nutzergenerierten Medieninhalten wächst jedoch der **Bedarf für externe Dienstleistungen** im Bereich der automatisierten Erkennung von Bildern sowie Video- und Audioinhalten. Deren Erfassung ist wesentlich komplexer ist als reine Textanalyse und die hierfür notwendigen Analysewerkzeuge können auch bei größeren Plattformen nicht (allein) durch interne Teams entwickelt werden. Durch die gesteigerten regulatorischen Anforderungen (zuletzt NetzDG, aktuell DSA)

[40] Bei Diensten, die nicht im Kern ihres Geschäftsmodell nutzergenerierte Inhalte verbreiten und nicht dem DSA unterfallen, werden hingegen vor allem externe Lösungen zur Inhaltemoderation eingesetzt.

müssen auch die Moderationsmanagementsysteme angepasst werden, um die Ausweisung zu erweitert und regulierungskonform zu gestalten. Gerade bei kleineren Anbietern, bindet diese Anpassungen eigenentwickelter Lösungen proportional viel Kapazität. Künftig dürften daher vermehrt externe technische Systeme in die Moderationsprozesse eingebunden werden, auch um eigene Entwicklungsressourcen für die Kernaufgaben der Dienste-Bereitstellung freizustellen.

Mit Blick auf die **manuelle Moderation** setzt die Mehrzahl der kleineren, auf Deutschland ausgerichteten Dienste auf eigene, festangestellte Moderationsteams, die durch freie Mitarbeiter und Hilfskräfte in Deutschland unterstützt werden. Hier arbeiten die Moderatoren in der Regel eng mit der Community-Leitung bzw. dem Trust-and-Safety-Team zusammen. Die großen bis sehr großen Online-Plattformen nehmen zur manuellen Moderation hingegen ausschließlich externe Dienstleister in Anspruch (vgl. Kap. 3.4.2), die in der Regel an verschiedenen Standorten in verschiedenen Ländern operieren und räumlich getrennt von der Community-Leitung des Dienstes arbeiten.

Insgesamt wird deutlich, dass sämtliche Anbieter – von kleinen bis zu sehr großen Diensten –einen mehrschichtigen Ansatz verfolgen, indem sowohl automatisierte als auch manuelle Verfahren der Inhaltemoderation zum Einsatz kommen. Die Spannbreite reicht von Anbietern, die sehr stark auf automatisierte Erkennungsverfahren und proaktive Moderation setzen bis zu Anbietern, die bislang vornehmlich auf die Eigenmoderation durch die Nutzergemeinschaft setzen (community-led moderation).

Derzeit gibt es kein einzelnes, grundsätzlich überlegenes Moderationsverfahren oder Moderationsinstrument. Jedes Verfahren hat spezifische Vor- und Nachteile mit Blick auf Parameter wie Erkennungsgeschwindigkeit, Erkennungsraten, Kosten, und Personalaufwand. Der jeweilige Verfahrensmix ist zudem abhängig vom Themenbereich, Tageszeit, aktueller Lage oder dem Zielmarkt, sodass auch innerhalb eines Dienstes die Verfahren unterschiedlich gewichtet werden können und der Einsatz der Moderationsressourcen einem kontinuierlichen Wandel unterliegt.

Moderationsentscheidungen, wie die Entfernung von Inhalten oder der Ausschluss von Nutzern, wird jedoch bei allen Anbietern in manuelle Moderationsprozessen entschieden.

3.3 Verfahren der Inhaltemoderation

Content-Moderation erfolgt entweder automatisiert durch maschinelle Algorithmen oder manuell durch menschliche Moderatoren. Automatisierte Moderation bietet Skalierbarkeit und Geschwindigkeit, weshalb sie primär proaktiv eingesetzt wird. Durch manuelle Moderation können komplexe Nuancen besser erfasst werden. In der Praxis findet in den meisten Fällen eine Kombination beider Ansätze Anwendung, um Effizienz und Genauigkeit zu optimieren.

3.3.1 Automatisierte Verfahren

Bei Diensten sind verschiedene automatisierte Verfahren im Einsatz, um die Qualität, Sicherheit und Integrität von Online-Inhalten zu gewährleisten.

Regelbasierte Abgleich- und Filtersysteme

Regelbasierte Abgleich- und Filtersysteme identifizieren, flaggen, blockieren oder entfernen unerwünschte, schädliche oder inakzeptable Inhalte bevor/während sie auf eine Plattform hochgeladen werden und im Falle eines Verstoßes blockieren diese direkt (Upload-Filter) oder flaggen den Inhalt für ein manuelles Reviewing.[41] Sie nutzen vordefinierte Listen, Regeln, Mustererkennungsverfahren und Algorithmen, um eine zielgenaue Überwachung und Kategorisierung von Inhalten zu ermöglichen.

- **Wortfilter:** Wortfilter zählen zu den einfachsten Verfahren der Content-Moderation. Sie können auf jedem Moderationssystem manuell eingerichtet werden. Wortfilter löschen, ersetzen und flaggen Wörter und ganze Ausdrücke, die gegen Gemeinschaftsrichtlinien verstoßen.[42] Pre-Sets an Wortfiltern können eingekauft bzw. lizensiert werden.
- **Automated-Content-Recognition (ACR):** ACR-Technologien analysieren multimediale Inhalte wie Bilder, Videos[43] und Audio automatisch. Dabei werden eindeutige Merkmale oder Muster erfasst, die dann mit einer Datenbank bekannter problematischer Inhalte abgeglichen werden. So können beispielsweise urheberrechtlich geschützte Werke oder unerwünschte Materialien erkannt werden.
- **Content-Hashing:** Hierbei werden Inhalte in eindeutige kryptografische Hash-Werte umgewandelt. Diese Hash-Werte dienen als digitaler Fingerabdruck und ermöglichen einen sehr schnellen Abgleich („matching") mit einer Datenbank, um identische oder ähnliche Inhalte zu erkennen.[44]
- **Content-Digital-Fingerprinting:** Ähnlich wie bei Content-Hashing wird hier ein digitaler Fingerabdruck eines Inhalts erstellt, jedoch oft auf der Basis von komplexeren Algorithmen. Dadurch können auch Bearbeitungen oder Veränderungen an Inhalten erkannt werden.[45]

Im Vergleich zu anderen Systemen zur Inhaltemoderation, wie manuellen Moderationsprozessen oder prädiktive, KI-gestützte Systeme mit maschinellem Lernen, zeichnen sich Filtersysteme durch ihre Automatisierung, Schnelligkeit und Effizienz aus.

---

41 Vgl. Policy Department for Citizens' Rights and Constitutional Affairs (2020): "The impact of algorithms for online content filtering or moderation", Kapitel 3, S. 35, online unter: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf, abgerufen am 09.10.23

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

43 Im Falle der Videoanalyse werden hierfür üblicherweise Einzelbilder eines Videofeeds erfasst und wie Bilder behandelt.

44 Vgl. Policy Department for Citizens' Rights and Constitutional Affairs (2020): "The impact of algorithms for online content filtering or moderation", Kapitel 3, S. 35, online unter: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf, abgerufen am 09.10.23

45 Vgl. ebd.

Tab. 8 Beispiele für Abgleich- und Filtersysteme

Name	Einsatzfeld	Entwickler	Nutzer
Audible Magic Identification	Copyright Audio Erkennung	Audible Magic	Soundcloud, Sony, Disney etc.
Content ID	Copyright Content Identifizierung	YouTube	YouTube
Content Levels	Sperrung von nicht-jugendfreien Inhalten	TikTok	TikTok
PowerTrack API	Filtern von Tweets	Twitter/X	Twitter/X
Twitch Audio Recognition	Copyright Audio Erkennung	Twitch	Twitch
PhotoDNA	Verhütung und Bewältigung von Viktimisierung von Kindern, einschließlich Entführung, Missbrauch und Ausbeutung	Microsoft	National Center for Missing & Exploited Children (NCMEC)

Quelle: Goldmedia Analyse 2023

Solche Filtersysteme, die Ausgangsmaterialien mit bestehenden Datenbanken abgleichen, funktionieren in der Praxis in der Regel gut und stellen einen wichtigen Baustein der Inhaltemoderation großer Online-Plattformen dar. Obwohl solche Filtersysteme viele Vorteile bieten, sind diese auch mit bestimmten Herausforderungen und Einschränkungen verbunden, vor allem, weil sie lediglich auf bereits bekannte Inhalte angewendet werden können. So werden Wortfilter, die anstößige Begriffe herausfiltern, etwa dadurch umgangen, dass Formulierungen und Schreibweisen gerade soweit angepasst, dass diese nicht mehr durch den Wortfilter erkannt werden. Nutzer umgehen auch die gesperrten Worte, indem Schreibweisen abgeändert werden, oder mit mehr oder weniger subtilen Anspielungen gearbeitet wird, die gewisse Transferleistungen der Leser verlangen. Diese Umgehung von Filtern ist mitunter einer der Ursachen für das Entstehen von Netzjargon („Leetspeek“) und seit Langem Bestandteil der Internetkultur. Die Effektivität von starren Filtersystemen ist daher stark davon abhängig, dass die Filterlisten fortwährend angepasst und weiterentwickelt werden, da Nutzer bewusst die Grenzen des Schreibbaren austesten.

Die nachfolgende Tabelle bietet eine Übersicht über die Vor- und Nachteile solcher Filtersysteme für die Inhaltemoderation.

Tab. 9 Vor- und Nachteile regelbasierter Filtersystemen für Inhaltemoderation

Vorteile	Nachteile
Schnelligkeit Durch die automatisierten Prozesse können Inhalte in Echtzeit überprüft werden, was eine rasche Reaktion auf problematische Inhalte ermöglicht.	**Fehlinterpretationen** Manchmal können Filtersysteme Inhalte falsch interpretieren und legale oder unschädliche Inhalte fälschlicherweise blockieren oder entfernen.
Skalierbarkeit Filtersysteme können große Mengen an Inhalten verarbeiten, was besonders wichtig ist, da täglich eine immense Menge an Beiträgen und Medien auf Online-Plattformen hochgeladen wird.	**Grauzonen** Nicht alle problematischen Inhalte sind eindeutig identifizierbar, und es gibt Fälle, die eine menschliche Bewertung erfordern.
Kontinuierliche Verbesserung Filterregeln können manuell jederzeit angepasst und verbessert werden.	**Kreativer Umgang mit Inhalten** Personen, die schädliche Inhalte verbreiten möchten, können versuchen, Filtersysteme durch Änderungen oder Verschleierungen zu umgehen.
Anpassbarkeit Plattformen können die Parameter der Filtersysteme an ihre eigenen Nutzungsrichtlinien und Bedürfnisse anpassen.	**Fehlende Kontextualisierung** Filtersysteme arbeiten oft auf Grundlage von Schlüsselwörtern, Mustern oder Regeln. Dadurch können sie den Kontext eines Beitrags oder Kommentars nicht immer richtig erfassen.

Quelle: Socialays (2023): „Pros and Cons of AI vs Manual Content Moderation", online unter: https://socialays.com/blog/pros-cons-ai-manual-content-moderation/, abgerufen am 07.09.23

Selbstlernende Filtersysteme – Künstliche Intelligenz

Neben regelbasierten Filtersystemen kommen mittlerweile auch vermehrt vortrainierte oder von Grund auf selbst trainierte Filtersystemen mithilfe von Künstlicher Intelligenz (KI) zum Einsatz. Diese sind in der Lage, die Tonalität längerer Texte oder Bezüge zu kritischen Themen oder Ansichten zu erfassen oder bei Bildern bzw. Videobeiträgen Verstöße gegen Gemeinschaftsrichtlinien oder auch rechtswidrige Inhalte zu erkennen.

Damit stellen sie einen wichtigen Baustein zur Skalierung der Moderationsleistung und Unterstützung und Entlastung der manuellen Moderation vor dem Hintergrund eines ständig wachsenden Aufkommens an nutzergenerierten Inhalten dar. Gleichzeitig kann die Einbindung von KI die psychische Belastung der menschlichen Moderatoren reduzieren, wenn belastende oder verstörende Inhalte von der KI erkannt und zum Schutz der Moderatoren unkenntlich gemacht werden.

Für die Moderation von Inhalten kommen primär zwei Varianten von künstlicher Intelligenz zum Einsatz:

- **Machine Learning (ML)** wird im Allgemeinen für die Verarbeitung großer Datensätze genutzt, um Muster und Zusammenhänge zu erkennen. Dabei werden Algorithmen verwendet, um aus Erfahrungen zu lernen und Vorhersagen oder Entscheidungen zu treffen. ML-Modelle können komplexe Aufgaben bewältigen, indem sie automatisch Muster in den Daten identifizieren und darauf basierend Schlussfolgerungen ziehen.[46]
- **Natural Language Processing (NLP)** befasst sich damit, geschriebene oder gesprochene Sprache in eine Form umzuwandeln, die von Computern verstanden werden kann. Dies kann u.a. mit statistischen oder ML-Modellen erfolgen. NLP ermöglicht, komplexe Texte zu analysieren, zu interpretieren und Bedeutung zu extrahieren. NLP wird für Übersetzungen, Textverarbeitung, Sentimentanalysen und viele andere Sprach- und Text-basierte Aufgaben eingesetzt.[47]

Die Umsetzungen von ML und NLP, die Im Kontext von Filtersystemen für Inhaltemoderation zu beobachten sind, lassen sich insofern voneinander abgrenzen, dass Mustererkennung mittels Machine Learning vornehmlich für die Identifikation von Bildern bzw. Videoinhalten verwendet wird und NLP, mit und ohne ML-Unterstützung, explizit für die Verarbeitung von Sprache eingesetzt wird.

Im Textbereich sind bereits viele NLP-gestützte Systeme verfügbar, die jenseits von Wortfiltern problematische Inhalte auch im Kontext erkennen oder sogar Sentimentanalysen zur Tonalität eines Textes erstellen können. Je nach Konfiguration kann den Erstellern noch beim Abfassen ein Hinweis gegeben werden, dass ein Text ggf. problematisch ist. Diese Systeme finden sich auch in Moderationschatbots, die Moderationsentscheidungen treffen und erklären können. Diese moderierenden Chatbots werden als „Auto-Mods" bezeichnet.

Diese KI-Systeme werden als Bestandteil eines Moderationssystems oder als eigenständige Module angeboten, die über Schnittstellen in das Moderationssystem integriert werden können. Konkret bedeutet dies, Nutzerbeiträge werden über das Moderationssystem an das entsprechende Tool weitergeleitet, das die Inhalte verarbeitet und prüft und ggf. moderiert. Oft sind solche Systeme lernfähig. Ergebnisse manuelle Moderationsprozesse werden hierbei an das KI-System weitergemeldet. So können die eingesetzten KI-Module anhand der Daten (Nutzerbeiträge und Moderationsentscheidungen) der jeweiligen Kundenplattformen individuell trainiert werden.

Im Bildbereich erfolgt die Mustererkennung bspw. für pornografische oder gewaltdarstellende Inhalte auf den großen Plattformen schon seit vielen Jahren bereits während des Uploads.

Im Video- und Audiobereich hingegen basiert die Erkennung problematischer oder rechtlich geschützter Inhalte vor allem auf Hash-Datenbanken (vgl. YouTube Content-ID oder Twitch Audio Recognition in Tab. 8). Herausforderung ist hierbei neben mangelnden plattformspezifischen Trainingssets die nicht hinreichend präzise Erkennung

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

[47] Vgl. ebd.

von Video- und Audiomaterial, etwa aufgrund von schlechter technischer Qualität des Videoinhalts, sich überlagernder Audioquellen oder anderen technischen Ursachen. Nur wenige externe Lösungsanbieter sind bereits in der Lage, marktreife Produkte anzubieten.[48] Zusätzlich ist zu berücksichtigen, dass eine vollständige inhaltliche Analyse von Videodaten aufgrund der Datenmengen sehr viel Rechenleistung erfordert (dies gilt insbes. für Echtzeit-Analysen bei Streaming-Plattformen) und mit hohen Kosten pro analysiertem Video-Beitrag verbunden ist (vgl. Kap. 3.4.3, AWS Amazon Recognition). Daher werden solche Systeme auch zukünftig von den Plattformen wohl vornehmlich anlassbezogen eingesetzt.

Auch die Übersetzung und KI-basierte Analyse von Tonspuren auf Videobeiträgen oder von Podcasts befindet sich noch in der Entwicklung. Hier erfolgt im ersten Schritt eine Überführung von Ton in Text mit Tools wie Assembly AI oder anderen Text-to-Speech-Anwendungen. Im nächsten Schritt folgt dann die Textanalyse. Spotify hat zur Kontrolle der vielen Podcasts auf seiner Plattform im Jahr 2022 nach bereits längerer Zusammenarbeit den Audio-Content-Moderator Kinzen erworben.[49]

Kontextsensitive Analyse verbundener Plattforminhalte

Eine wesentliche künftige Entwicklung im Bereich KI-gestützter Systeme ist die kontextsensitive Analyse von Live- bzw. Videoinhalten, bei denen Videostreams gemeinsam mit den ihn begleitenden Kommentaren und weiteren Metadaten analysiert werden, um ihnen ein Risikoprofil beizumessen. Zahlreiche Inhalte verstoßen nicht per se, sondern erst im Kontext mit der Äußerungsabsicht gegen Gemeinschaftsrichtlinien. Dies erschwert die automatisierte Früherkennung auf Basis marktverfügbarer Systeme bislang stark.

Bei den derzeit in Entwicklung befindlichen kontextsensitiven Analyse gehen die sprachlichen und textlichen Äußerungen gemeinsam mit den auf der Plattform gespeicherten Daten und Metadaten der Nutzer (z. B. deren Avatar, Bilder inkl. Metadaten, Kontaktnetzwerke, Verbindungen zu problematischen Communities) in die Analyse ein. Sowohl große Social-Media-Plattformen als auch Dienstleister entwickeln solche **prädiktiven Verfahren zur Früherkennung**, da diese gerade bei besonders drastischen Ereignissen (Tötungsdelikte, terroristischen Anschläge etc.) bei Marktreife einen wesentlichen Unterschied in der Früherkennung machen könnten. Auf dieser Basis können auch vermeintlich unauffällige Streams auf Basis einer KI-Prognose ein hohes Risikoprofil erreichen und hierdurch wesentlich früher geflaggt und einem menschlichen Moderator zur Überprüfung angezeigt werden.

---

[48] Hierzu zählen etwa die Lösungen von Amazon (Vgl. Seite 53), die noch recht kostenaufwendig sind und nicht über Echtzeitfähigkeiten verfügen.

[49] Vgl. https://inside.com/podcasting/posts/spotify-acquired-content-moderation-company-kinzen-318967, abgerufen am 09.10.23

3.3.2 Manuelle Verfahren

Als manuelle Verfahren werden alle Entscheidungsprozesse bezeichnet, bei denen Moderationsentscheidungen durch ein oder mehrere Menschen getroffen werden. Diese Entscheidungsprozesse können diensteabhängig auf vielfältige Weise organisiert sein und auf verschiedenen Hierarchieebenen erfolgen. Im Folgenden werden die wesentlichen Ebenen der manuellen Moderation beschrieben, die Implementierung der hier beschriebenen wesentlichen Ebenen kann sich bei sehr großen Online-Plattformen über verschiedene Konzerngesellschaften erstrecken.

3.3.2.1 Gemeinschaftsmoderation durch die Nutzer eines Diensts

Die Gemeinschaftsmoderation durch die Nutzer eines Dienstes („community-led moderation“) ist eines der ältesten Ansätze der Inhaltemoderation. Bereits in frühen nicht-kommerziellen und weitgehend idealistisch getriebenen Usenet-Foren und Bulletin Boards war es üblich, dass Nutzer hierarchisiert waren und über verschiedene Berechtigungen verfügten. „Foren-Admins“ kam üblicherweise die Aufgabe zu, Inhalte zu moderieren, zu löschen, zu verschieben etc. Die entsprechenden Berechtigungen konnte man sich auf Basis von Kriterien wie regelmäßiger ehrenamtlichen Mitarbeit, Sachkenntnis und Moderationsgeschick erarbeiten.

Diese Art der ehrenamtlichen Moderation findet auch heute noch bei zahlreichen Diensten Anwendung und ist dabei mitunter enger Bestandteil der Philosophie eines Dienstes insbesondere bei nicht-kommerziellen und interessengetriebenen Diensten.

Je spezifischer die Ansprache einer bestimmten Community bei einem Dienst oder in einem Teilbereich eines Dienstes ist, umso mehr ergibt es auch aus kommerziellen Erwägungen heraus Sinn, die Nutzer (bzw. besonders aktive Teile der Nutzerschaft) langfristig partizipativ zu binden und sie für die Fortentwicklung ihrer Community zu gewinnen. So können bspw. Nutzer die über einen längeren Zeitraum zuverlässig Verstöße gemeldet haben, zu Usermoderatoren zu ernennen, ggf. mit Zugang zum Moderationssystem inkl. Berechtigung zum Verbergen von Inhalten.

Die Gemeinschaftsmoderation durch die Nutzer eines Diensts beschränkt sich dabei nicht auf nicht-kommerzielle oder Nischenangebote. Auch große Online-Plattformen, wie etwa Twitch messen der Gemeinschaftsmoderation durch die eigenen Nutzer einen erheblichen Stellenwert bei und inkorporieren die Möglichkeit zur Gemeinschaftsmoderation innerhalb bestimmter Grenzen.

Bei Online-Plattformen mit stark live-getriebenen Inhalten (z. B. Live-Video-Streaming) und einer aktiven Nutzer-Gemeinschaft ist die Gemeinschaftsmoderation häufig das effektivste Instrument, um schnell auf unerwünschte Inhalte reagieren zu können. Automatisierte Systeme, insbes. KI-gestützte, arbeiten bislang nicht hinreichend latenzarm, um Live-Inhalte in Echtzeit überwachen zu können.

Eine reine Gemeinschaftsmoderation durch die eigenen Nutzer ist für kommerzielle Plattformen jedoch nicht praktikabel. Und die nachträgliche Entfernung/Sperrung von Inhalten erfolgt hier durch die im Auftrag des Dienstes tätigen Moderatoren.

3.3.2.2 Eigenmoderation durch Inhalte-Ersteller

Die Eigenmoderation durch Inhalte-Ersteller existiert typischerweise bei Plattformen, bei denen das Verhältnis zwischen Inhalte-Erstellern und Inhalte-Konsumenten stark asymmetrisch ist, etwa bei Live-Streaming-Plattformen wie YouTube oder Twitch. Die jeweiligen Inhalte-Ersteller (YouTuber, Streamer bzw. Creators) können auf ihren Kanälen eigenen Moderationsrichtlinien festlegen, die strenger als die Gemeinschaftsrichtlinien der jeweiligen Plattform sein können. Um diese durchzusetzen, stellen die Plattformen den Inhalte-Erstellern spezielle Moderationswerkzeuge bereit. Neben der Möglichkeit einzelne Kommentare zum eigenen Beitrag händisch zu löschen, werden zusätzliche Wortfilter-Systeme angeboten. Teilweise können auch Moderationswerkzeuge von freien Entwicklern über entsprechende Schnittstellen der Online-Plattformen durch die Inhalte-Ersteller eingebunden werden. Darüber hinaus können auch Mitglieder der eigenen Community mit Moderationsprivilegien ausstatten, damit diese z. B. Chat- und Kommentarfunktion in Echtzeit moderieren können, während der Inhalte-Ersteller selbst einen Live-Stream veranstaltet. Die Eigenmoderation durch Inhalte-Ersteller ersetzt hierbei nicht die im Auftrag des Dienstes tätigen Moderatoren, sondern ergänzt diese in bestimmten Umfeldern.

Die Umsetzung dieser Moderationstools ist hierbei vom Charakter und der Ausrichtung der Plattform abhängig: Auf Twitch sorgen Moderatoren etwa dafür, dass der Chat den vom Streamer festgelegten Benimm- und Inhaltsstandards folgt, indem sie anstößige Inhalte und Spam entfernen. Auf YouTube helfen Moderatoren bei der Überprüfung und Verwaltung von Kommentaren, die Leute zu einem Video hinterlassen, oder auch von Nachrichten, die Teilnehmer während des Live-Chats eines Streams senden. YouTube unterscheidet hierbei zwischen Standardmoderatoren und leitende Moderatoren, welche zusätzliche Optionen der Inhaltemoderation haben. Der Kanalbetreiber legt die Moderationsprivilegien individuell fest. Eine Auswahl der Moderationswerkzeuge von Twitch und YouTube zeigt die folgende Tabelle.

Tab. 10 Auswahl von Moderations-Werkzeugen auf Twitch und YouTube

HD*	Instrument	Eigenbeschreibung des Instruments auf der Plattform
Twitch	AutoMod	Automatisierte Methode zum Identifizieren potenziell riskanter Chatnachrichten
	Chatregeln	Kanalbetreiber können ihren eigenen Regelkatalog für ihren Kanal erstellen, um neue Zuschauer darüber zu informieren, welches Verhalten in dem Chat angemessen ist. Bei Twitch können auch Chatbots eingesetzt werden, die Nutzer darüber informieren, dass ihr Beitrag geblockt oder sie gesperrt werden.
	Moderatoren-Tools im Chat	Vom Streamer festgelegte Moderatoren können den Chat- und den Sperrverlauf von Chattern einsehen und Kommentare zu Benutzern hinterlassen und anzeigen.
	Links blockieren	Einstellung, die verhindert, dass Links auf dem Kanal gepostet werden
	Verzögerung im Nicht-Mod-Chat	Einstellung, die Nachrichten leicht verzögert im Chat erscheinen lässt
	E-Mail- & SMS-Verifizierung	Einstellung, die verhindert, dass Benutzer ohne verifizierte E-Mail-Adresse und Handynummer in den Chat schreiben können
	Nur-Follower- & Nur-Abonnenten-Modus	Optionen mit denen man festlegen kann, ob Benutzer einem folgen oder abonniert sein müssen, damit sie in den Chat schreiben dürfen

HD*	Instrument	Eigenbeschreibung des Instruments auf der Plattform
	Gesperrte Chatter	Liste mit Benutzern, die dauerhaft für das Chatten in dem Kanal gesperrt wurden.
YouTube	Den Kanal aufrufen	Wenn dir eine Nachricht im Livechat auffällt, kannst du direkt den Kanal eines Livechat-Teilnehmers aufrufen und so zuerst mehr über ihn erfahren
	Inhalte entfernen	Du kannst alle unangemessenen oder potenziell missbräuchlichen oder anstößigen Inhalte entfernen. Wenn du eine Nachricht löschst, wird sie zusammen mit allen Antworten endgültig aus dem Livechat entfernt.
	Nutzer vorübergehend blockieren	Du kannst für einen Zeitraum von fünf Minuten verhindern, dass jemand Nachrichten im Livechat sendet.
	Nutzer auf diesem Kanal ausblenden	Die Chatnachrichten und Kommentare dieser Person sind dann für andere Zuschauer nicht mehr sichtbar. Der ausgeblendete Nutzer wird darüber nicht benachrichtigt.
	Potenziell unangemessene Nachrichten prüfen	Du kannst Kommentare oder Nachrichten, die aufgrund deiner Community-Einstellungen zur Überprüfung zurückgehalten wurden, ein- oder ausblenden.
	Community-Standardeinstellungen auswählen	Du kannst Funktionen aktivieren, um mithilfe von Technologien automatisch Spam, Eigenwerbung, unsinnige und andere unangemessene Inhalte in Kommentaren zu erkennen.
	Livechat aktivieren/deaktivieren	Du kannst den Livechat auch nach Beginn der Veranstaltung jederzeit aktivieren oder deaktivieren.
	Teilnahmemodus ändern	Du kannst den Teilnahmemodus im Livechat anpassen und so nur Beiträge durch Abonnenten, Beiträge durch Mitglieder oder Livekommentare zulassen.
	Nachrichtenverzögerung einschalten	Du kannst einschränken, wie oft ein Nutzer eine Chatnachricht senden kann, indem du ein Limit für den zeitlichen Abstand zwischen Kommentaren festlegst.
	Gesperrte Wörter erkennen	Du kannst Nachrichten im Livechat sperren, die bestimmte Begriffe oder ähnliche Wörter enthalten.
	Hinweis	Als leitender Moderator hast du keinen Zugriff auf den Live Control Room oder auf YouTube Studio. Leitende Moderatoren können keine anderen Moderatoren ernennen.

**HD: Hostingdienste*
Quelle: Twitch „Einrichten der Moderationseinstellungen für deinen Twitch-Kanal" (Stand September 2023), online unter: https://help.twitch.tv/s/article/setting-up-moderation-for-your-twitch-channel?language=de, und https://dev.twitch.tv/docs/irc/, abgerufen am 07.09.23
YouTube „Moderationstools verwenden" (Stand September 2023), online unter: https://support.google.com/youtube/answer/10888907?hl=de&co=GENIE.Platform%3DAndroid, abgerufen am 07.09.23

Facebook bietet seinen Creators zum Beispiel die Möglichkeit, festzulegen, welche Art von Text-, Foto- oder Videobeiträgen Besucher auf der eigenen Seite posten können. Beiträge von anderen Personen lassen sich grundsätzlich erlauben oder deaktivieren. Weiter gibt es die Option, Foto- und Videobeiträge erst nach einer manuellen Überprüfung zu veröffentlichen. Kommentare zu den Beiträgen auf einer Seite lassen sich nicht deaktivieren, einzelne Kommentare können jedoch ausgeblendet oder gelöscht werden. Zusätzlich zu den automatisch greifenden Wortfiltern können Creator auch einen Wortfilter für „vulgäre Sprache" aktiv dazuschalten.[50] Zudem lassen sich einzelne Nutzer von

[50] Vgl. https://www.facebook.com/help/1017549069082358, abgerufen am 22.09.23

der Seite ausschließen oder es lassen sich Bewertungen für die Seite allgemein deaktivieren.[51]

TikTok bietet gleichfalls spezielle Moderationstools für Inhalteersteller und Live-Kanäle, für die bis zu hundert Nutzermoderatoren benannt werden können. Zudem gibt es die Möglichkeit für Nutzer, Schlüsselwörter aus der eigenen Nutzererfahrung auszublenden oder über eine Schaltfläche persönlich unerwünschte Inhalte oder Nutzer zu blockieren.

3.3.2.3 Manuelle Moderation im Auftrag eines Dienstes

Die manuelle Moderation im Auftrag eines Dienstes ist, nach wie vor, der bedeutendste Baustein innerhalb des Moderationsprozesses. Die Mehrheit der Meldungen automatisierter Verfahren[52] sowie sämtliche manuell über die Meldewege der Plattform eingegangenen Meldungen werden durch menschliche Moderatoren manuell überprüft und bearbeitet. Hierdurch kann bei großen Online-Plattformen ein beträchtlicher Personalaufwand entstehen (vgl. Kap. 4.1), der üblicherweise im Rahmen des Business Process Outsourcings auf Dienstleister für manuelle Moderation ausgelagert wird.[53]

Bei kleineren Anbietern ist es hingegen auch in Deutschland üblich, dass die manuelle Moderation vollständig innerhalb des eigenen Unternehmens durch eigene Angestellte realisiert wird.

Spezialisierte Outsourcing-Dienstleister stellen das Personal für die manuelle Moderation bereit und betreiben die Center, in denen die Moderatoren tätig werden, statten diese aus und bieten psychologische Unterstützung für die angestellten Moderatoren an. Aufgrund des sensiblen Arbeitsumfeldes und der mitunter entstehenden psychologischen Belastungen findet Content-Moderation in überwiegendem Maße innerhalb der Räumlichkeiten eines Moderations-Dienstleisters statt. Remote-Arbeit ist aus Gründen des Schutzes der Moderatoren und deren Angehörigen unüblich.

Als spezialisierte Dienstleister in einem stark arbeitsteiligen Prozess erarbeiten Dienstleister nicht die der Moderation zugrundeliegenden Richtlinien, sondern wenden diese nur an. Trotz ihrer Moderationserfahrung sind Outsourcing-Partner lediglich Dienstleister der Online-Plattformen. Strategische Beratung zu Fragen der Inhaltemoderation ist kein primäres Geschäftsfeld der Dienstleister. Jedoch sind Bestrebungen der Dienstleister für manuelle Moderation im Markt erkennbar, ihr Angebot auszuweiten, um zu Full-Service-Dienstleistern für Inhaltemoderation zu werden. Da sie Erfahrung im Umgang mit vielen Gemeinschaftsrichtlinien und Moderationsprozessen haben und denselben Rechtsrahmen für mehrere Kunden anwenden, fungieren sie auch als Transferzentrum für Best-Practices.

Der Markt für Outsourcing-Dienstleister zeichnet sich dadurch aus, dass die führenden Unternehmen global operieren und Moderationscenter in vielen Ländern auf verschiedenen Kontinenten betreiben. Anbieter, mit denen für diese Studie gesprochen wurde, unterhalten bis zu 20 Standorte in allen Regionen der Welt und beschäftigen bis zu

---

[51] Vgl. https://de-de.facebook.com/business/help/1323914937703529, abgerufen am 22.09.23

[52] Hierunter fallen nicht einfache Filtersysteme, die aufgrund von Binärlogiken entscheiden, sondern höher entwickelte, prädiktive Filtersysteme, die üblicherweise ab einer Erkennungswahrscheinlichkeit von 80 % bis 85 % einen manuellen Moderationsbedarf generieren.

[53] Auch große bzw. sehr große Online-Plattformen moderieren oft eine Stichprobe ihrer Inhalte inhäusig, allerdings ist dies vor allem im Kontext der plattforminternen Qualitätskontrolle und der Weiterentwicklung der Gemeinschaftsrichtlinien im Aufgabenspektrum des Legal-and-Policy Teams zu betrachten.

80.000 Mitarbeitende. Solche global operierenden Dienstleister werden bevorzugt von sehr großen Online-Plattformen beauftragt, da diese Dienstleister eine Inhaltemoderation nahe am Sprach- bzw. Kulturraum vieler regionaler Märkte aus einer Hand anbieten können. Zur Risikoverteilung arbeiten große Online-Plattformen jedoch i. d. R. (auch innerhalb desselben regionalen Marktes oder Sprachraums) mit mehreren Moderations-Dienstleistern zusammen.

Inhalte in deutscher Sprache werden von einigen Outsourcing-Dienstleistern auch in Deutschland moderiert, allerdings kommen aufgrund der Kostenvorteile auch Moderationscenter in anderen EU-Ländern zu Einsatz, solange auf den dortigen Arbeitsmärkten hinreichend deutschsprechende Arbeitskräfte rekrutiert werden können. Die Moderation von deutschsprachigen Inhalten in den großen Moderationsstandorten Asiens, Afrikas oder Indiens, ist hingegen aufgrund der fehlenden deutschen Sprachkenntnis nicht üblich.

Dienstleister für manuelle Moderation kommen bei allen großen und sehr großen Online-Plattformen zum Einsatz, unabhängig davon, wie sich die Plattform grundsätzlich im Spektrum zwischen rigider und liberaler Inhaltemoderation positioniert.[54] Hierunter zählen so auch Hostingdienste, die nicht dem DSA oder der TCO-VO unterfallen, wie etwa Messaging-Plattformen.

Für kleine Plattformen besteht kaum ein hinreichender Moderationsbedarf, der ein Outsourcing rechtfertigen würde. Die bestehenden Moderationsteams kleiner Plattformen sind zu klein, als dass sie mit einer signifikanten Aufwandreduktion outgesourct werden könnten (vgl. Kap.4.3). Zudem müssen die Moderationsprozesse hinreichend formalisiert sein, um den manuellen Moderationsprozess durch einen Dienstleister durchführen zu lassen.

Ein minimal besetztes Moderationsteam eines Dienstleisters, das durchgehend besetzt Inhalte moderiert, erfordert nach Aussagen der Anbieter eine Mindestteamgröße von ca. 10-20 Personen, die dediziert in einem festen Team für einen bestimmten Kunden eingesetzt werden.[55] Ein durchgehender Moderationsaufwand wird von kleineren Diensten jedoch nicht betrieben. Erst bei einem personellen Moderationsaufwand im deutlich zweistelligen Personalbereich überwiegend die Vorteile eines Outsourcings deren Nachteile.

Zukünftiger Stellenwert manueller Moderation

Die zunehmende Nutzung des Internets für Video-Anwendungen wird sich auch im Bereich der Online-Plattformen in den kommenden Jahren weiter fortsetzen. Social-Video-Plattformen wie etwa TikTok (mit kurzen vertikalen Videoclips/Reels) oder Twitch (mit langen Live-Streams) sind hierbei die Treiber der Entwicklung. Aber auch etabliertere soziale Netzwerke wie Facebook, Instagram und auch YouTube setzen zunehmend auf die Verbreitung kurzer Videoinhalte für die Smartphone-Nutzung (vertikale Videoclips/Reels).

---

[54] Ein Dienstleister für Moderationsdienstleistungen drückte dies mit „Kein Anbieter hat null Inhaltemoderation" aus und schloss damit Anbieter ein, die explizit zur freien Meinungsäußerung auf ihren Plattformen einladen.

[55] Aus Gründen der Vertraulichkeit und erforderlichen Kenntnisse der jeweiligen Gemeinschaftsrichtlinien werden einzelne Moderatoren nicht für verschiedenen Kunden parallel eingesetzt.

Hieraus resultieren steigende Anforderungen an die Inhaltemoderation, da mit Video- und Live-Inhalten besondere Herausforderungen einhergehen. Proaktive technische Systeme sind im Bereich Video bereits im Einsatz. Jedoch können bislang neben der Hashwert-gestützten Detektion von Urheberechtsverletzungen (Musik, Bilder, Videoausschnitte) nur wenige Verstöße gegen Gemeinschaftsrichtlinien oder Gesetze (z. B. Nacktheit) sicher maschinell detektiert werden. Im Vergleich zur Moderation von Texten, wo nahezu in Echtzeit Wortfilter und auch kurze Textanalysen greifen, bleibt die Moderation von Video-Inhalten stärker auf manuelle Verfahren angewiesen.

Live-Video-Inhalte stellen gegenüber hochgeladenen Videoinhalten nochmal erheblich gesteigerte Moderationsanforderungen dar, da Upload und Verbreitung zeitlich zusammenfallen und somit bislang keine klassischen Filtersysteme zum Einsatz kommen können.

Um der Herausforderungen zu begegnen, knüpft YouTube die mobile Live-Streaming-Funktion an bestimmte Bedingungen. Hierzu zählen:

- Verifizierter Kanal
- Mindestens 50 Abonnenten (bzw. 1.000 Abonnenten bei Jugendlichen)
- Keine Kanal-Einschränkungen beim Livestreaming in den letzten 90 Tagen
- Bis zu 24 Stunden Wartefrist bei erstmaliger Aktivierung

Darüber hinaus setzen alle Video-Plattformen beim Thema Live-Content weiterhin vornehmlich auf die Einbindung ihrer Nutzer in den Moderationsprozess (Nutzermeldungen/Gemeinschaftsmoderation).

Darüber hinaus stellen die sich ständig verändernden und auch neu aufkommende problematische (politisch radikale) Netzwerke und damit verbundene Schlagwörter und Begrifflichkeiten eine Herausforderung für die automatisierte Inhaltemoderation dar. Die proaktive Beobachtung und Analyse der Tätigkeiten von radikalen Netzwerken zur Früherkennung neuartiger Herausforderungen liegt weiter vornehmlich im Bereich der manuellen Inhaltemoderation. Einzelne Dienstleister haben sich ähnlich wie Strafverfolgungsbehörden bereits auf die Beobachtung öffentlich schwer zugänglicher Bereiche des Internets (Darknet) spezialisiert, um die sich ändernden Kommunikationsmuster und neuartige Entwicklungen in Nischen-Communities im Bereich terroristischer Bedrohungen frühzeitig erfassen zu können.[56]

Trotz der Entwicklungen insbes. im Bereich der KI-gestützten kontextsensitiven Analysen (vgl. Kap. 3.3.1) und dem prädikativen Einsatz automatisierter Verfahren wurde in den geführten Expertengesprächen mehrfach darauf hingewiesen, dass es sich dabei um eine Unterstützung der manuellen Moderation handelt. Das bedeutet, manuelle Moderationsentscheidungen bleiben bei komplexeren Inhalten vorherrschend. Mit dem verstärkten Einsatz prädikativer Verfahren, die Ereignisse und Lagen im Vorfeld erkennen sollen, ist daher auch eher mit einem weiter steigenden manuellen Moderationsaufwand zu rechnen.

Letztlich kann KI zwar aufgrund der Verarbeitung größerer und unstrukturierterer Datenmengen insgesamt die Vorhersagequalität steigern. Jedoch handelt es sich um ein

[56] Hierfür beobachten Spezialisten einschlägigen Foren mit spezifischen Bedrohungen, z. B. dschihadistischen Gruppen im Nahen Osten.

statistisches Modell, das im Rahmen gegebener Parameter Vorhersagen macht und bei vielen Moderationsaufgaben auch künftig autonom keine Entscheidungen treffen kann. Schwierig bleibt zudem die automatisierte prädiktive Erkennung von Aktivitäten, die von vorneherein in Betrugsabsicht durch hochmotivierte Täter geschehen, etwa durch Scammer, Hater oder durch Pädophile. Zu deren Verhaltensweisen gehört etwa der Einsatz von Tarnprofilen und die Verwendung von altersspezifischem Vokabular und Diktion, sodass ein NLP-Modell kaum effektiv eingesetzt werden kann.

Automatisierte Systeme werden darüber hinaus aus dem Bereich der organisierten Kriminalität herausgefordert. Aus Dienstleisterkreisen wird anekdotisch davon berichtet, dass selbst Firmen gezielt daran arbeiten, KI-basierte Moderationssysteme zu überlisten und erhebliche Ressourcen einsetzen um „ausgeklügelte Betrugsmethoden" zu entwickeln. Auch bei gänzlich neuen Betrugsschemata offenbaren KI-basierte Erkennungssysteme ihre Schwächen: So wurden die im Zuge der Corona-Pandemie aufkommenden Betrugsmaschen im Handel mit medizinischen Masken von automatisierten Systemen nur unzureichend erkannt, da sie hierauf nicht hinreichend trainiert waren. Insofern wird auch von den KI-Lösungsanbietern selbst klargestellt, dass KI im Moderationsumfeld nicht alle Aufgaben lösen kann oder wird.

3.4 Dienstleister für die Moderation von Inhalten

3.4.1 Marktstruktur

Es besteht ein großer Markt an Anbietern für die Moderation von nutzergenerierten Inhalten, deren Dienstleistungen aus unterschiedlichen Kombinationen der nachfolgenden Dienstleistungen bestehen:

a) **Moderations(management)systeme** (SaaS-Plattformen mit hinterlegten Eskalations-Workflows, Aufbereitung der Inhalte für die Moderatoren, statistische Erfassung etc.)

b) **Technische Systeme für die automatisierte Inhalteerkennung** (Filtersysteme, KI-basierte Moderationsunterstützung etc.)

c) **Manuelle Moderationsdienstleistungen** (Moderatoren-Teams, Moderationscenterbetrieb, Personalbetreuung, Schulungen etc.)

Je nach Dienstleistungsbereich unterscheiden sich die Geschäftsmodelle der Dienstleister:

- Moderationssysteme – die Software, welche die Moderatoren benutzen – wird überwiegend zu einem Festpreis berechnet, wobei der Preis variable Preisbestandteile, wie etwa die Zahl der angeschlossenen Moderatoren, beinhalten kann.
- Technische Systeme für die Inhalteerkennung werden überwiegend aufwands-/volumenabhängig berechnet, etwa nach der Anzahl der API-Calls oder Anzahl der überprüften Elemente, wobei der Preis eine fixe Basiskomponente beinhalten kann.[57]
- Moderationsdienstleistungen werden in der Regel auf Basis des eingesetzten Personals, bzw. auf Basis der vereinbarten Arbeitsstunden abgerechnet. Leistungs-

[57] Hierbei kann sich die Definition der Elemente stark vom Einsatzbereich abhängen, z. B. ein Wort, ein Bild, ein Video, ein Textbeitrag, ein Post inkl. Begleitinhalte usw.

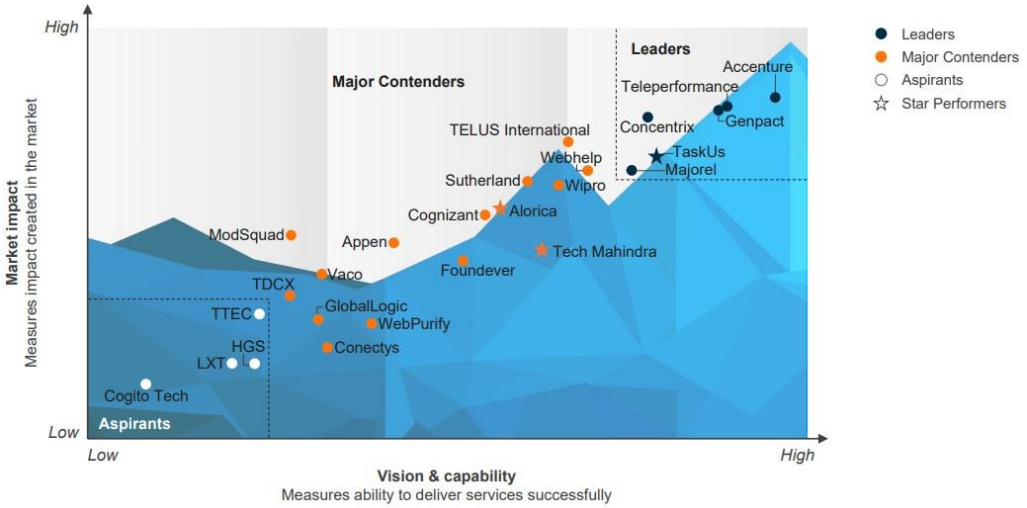
bzw. erfolgsabhängige Vergütungsmodelle, wie sie etwa bei Telekommunikationsdienstleistungen in Callcentern weit verbreitet sind, sind im Bereich der Inhaltemoderation unüblich.

Der Markt für die Inhaltemoderation lässt sich dabei im Wesentlichen unterteilen in:

1. **Spezialisierte Lösungsanbieter** im Bereich der **Moderationsmanagementsysteme**
2. **Spezialisierte Lösungsanbieter** im Bereich technischer Systeme für die **Inhalteerkennung**
3. **Sehr große IT-Konzerne**, die technische Systeme zur Inhaltemoderation als Teil ihrer Cloud-Lösungen anbieten und
4. **Dienstleister für Moderationsdienstleistungen** bzw. Full-Service-Anbieter, die vielfach aus dem Callcenter-basierten Outsourcing von Telekommunikationsdienstleistungen (Produktberatung, Kundensupport etc.) stammen, seit langem manuelle Inhaltemoderation anbieten und inzwischen vermehrt auch technische Systeme der Inhalteerkennung (in Kooperation oder als Zukauf) mit anbieten.

Die folgende Abbildung bietet eine exemplarische Übersicht der Anbieterlandschaft im US-amerikanischen Content-Moderations-Markt:

Abb. 6 Trust and Safety Services Assessment 2023 für den US-amerikanischen Markt



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

Anbieter im Bereich der Moderationssysteme bieten vielfach bereits einfache Filterlisten-Systeme als Lizenzbestandteil mit an. Komplexere, insbes. KI-gestützte Systeme der Inhalteerkennung sowohl im Bereich Text aber insbes. auch im Bereich Bild und Video sind bislang eine Domäne spezialisierter Anbieter.

Dies führt dazu, dass zahlreiche, zum Teil sehr spezialisierte Systeme am Markt angeboten werden. Gleichzeitig ist der Einsatz und die Anbindung vieler verschiedener Teilsysteme an das eigene Moderationssystem sehr herausfordernd für die Online-Plattformbetreiber.

Größere Anbieter im Bereich Moderationsdienstleistungen, die vornehmlich manuelle Moderation anbieten und eigene Moderationscenter betreiben, sind bestrebt sich durch Kauf von oder durch strategische Partnerschaften mit spezialisierten technischen Lösungsanbieter von KI-gestützten Moderationsinstrumenten künftig als vollwertige Full-Service-Dienstleister zu positionieren. Ziel ist es zum einen, eine höhere Wertschöpfung auch bei Großkunden zu erzielen, die derzeit gemeldete bzw. selbst detektierte Verstöße zur manuellen Moderation an die Dienstleister durchreichen. Zudem will man im Wettbewerb mit anderen Moderationsanbietern die Kosten für die Kunden weiter senken, indem man die heute manuell behandelten Fälle verstärkt automatisiert und damit eine Differenzierung erreicht. Darüber hinaus will man mit vornehmlich automatisierten Dienstleistungen auch den Markt der kleineren Online-Plattformen erreichen.

3.4.2 Dienstleister

Im Folgenden werden einige Dienstleister für die Moderation von Inhalten dargestellt, die auf die auf dem deutschen Markt ihre Dienstleistungen anbieten. Hierbei handelt es sich um eine exemplarische Darstellung einiger Dienste, sowie ihrer relevanten Dienstleistungen im Bereich Inhaltemoderation.

Tab. 11 Anbieter für ausschließlich automatisierte Moderationsverfahren und -lösungen (Stand: Juli 2023)

Produkt	Einsatz von KI	Erläuterung
ActiveFence	Ja	Bieten eine generische API mit Filtersystemen, aber entwickeln auch personalisierte Lösungen für ihre Kunden
Amazon Comprehend/ Recognition	Ja	Comprehend: Nutzt natürliche Sprachverarbeitung, um Zusammenhänge zu entdecken Recognition: Automatisiert und rationalisiert Bild- und Videomoderations-workflows
Azure AI Content Safety	Ja	Klassifiziert schädliche Inhalte gibt ihnen ein risikoabhängiges Rating
Azure AI Computer Vision	Ja	Nutzt visuelle Datenverarbeitung zur Kennzeichnung von Inhalten
Bodyguard.ai	Ja	Reine Softwarelösung zur Moderation
Community Sift	Ja	Chatfilter- und Inhaltsmoderationssystem für soziale Netzwerke
CleanSpeak	Ja	Bietet Software-Lösungen, die Kunden vor unangemessenen Inhalten schützt
Coral	Ja	Bietet Moderation von Kommentaren insbes. im Bereich der Online-Presse
Disqus	Ja	End-to-End-Plattform für die Einbindung von Nutzern des jeweiligen Dienstes. Bietet Moderation von Kommentaren bzw. Diskussionen
Ferret Go Conversario	Ja	Bietet Moderations-KI-Modelle für Dialog-Manager und Community Builder
Ferret Go Engagently	Ja	Soziale Plattform mit Lösungen für Kommentare und Community Management

Produkt	Einsatz von KI	Erläuterung
Hive Moderation	Ja	Bietet diverse automatisierte Lösungen zur Inhaltsmoderation für Text-, Bild- und Videoinhalte; Moderations-Dashboard; Erkennung von KI erzeugten Inhalten
Moderate Content	Ja	Programmierschnittstelle für Moderation von Bildinhalten in Echtzeit
Jigsaw Perspective	Ja	Ist in der Lage die Verarbeitung natürlicher Sprache zu nutzen, um das menschliche Verständnis für Wörter nachzuahmen
Respondology	Ja	Verwaltung und Anpassung der Moderation auf sozialen Medienplattformen der Kunden
Thorn Safer	Ja	Bietet Lösungen für Plattformen zur Ermittlung, Entfernung und Meldung von Material über sexuellen Kindesmissbrauch
Sightengine	Ja	Programmierschnittstelle (API) für Bild- und Videoinhalte; Anonymisierung von Bild- und Videoinhalten
WebPurify	ja	Bietet diverse automatisierte Lösungen zur Inhaltsmoderation für Text-, Bild- und Videoinhalte sowie für Metaverse-Anwendungen

Quelle: Goldmedia Analyse 2023

Technische Dienstleister haben in Hintergrundgesprächen betont, dass verschiedene Ausgangssprachen keine besondere Hürde für die Moderation von Inhalten darstellen. Die meisten Anbieter arbeiten mit Sprachmodellen, die über 100 Sprachen interpretieren können. Insbesondere Terrorismus-verdächtige Textinhalte seien hierbei technisch gut zu detektieren. Insofern ist in Zukunft mit dem Marktzutritt weiterer technischer Dienstleister zu rechnen, die bislang in Deutschland noch nicht tätig sind.

Neben kleineren, spezialisierten Anbietern technischer Systeme bieten inzwischen auch die **sehr großen IT-Konzerne Microsoft, Google und Amazon** automatisierte Moderationsinstrumente als externe Dienstleistung an.[58] Diese werden im folgenden Abschnitt detaillierter dargestellt. Der Vorteil dieser Angebote liegt darin, dass sie „off-the-shelf" für eine Vielzahl von Sprachen und Inhalte-Typen (Text, Bild, Video) genutzt werden können und zudem transparente, skalierbare Preismodelle (Abrechnung nach Datenmenge bzw. Items) ohne sprungfixe Kosten anbieten. Nachteil dieser Angebote ist, dass sie nur begrenzt individualisiert werden können (kein spezifisches Training der KI-Modelle auf Basis der Corpora eigener Hosting-Inhalte) oder nur spezifischen Moderationsaufgaben lösen, ohne eine Full-Service-Lösung zu sein. In der Regel ist auch die Nutzung weiterer Cloud- und Hostingdienste beim jeweiligen sehr großen IT-Konzern Voraussetzung für die Nutzung der automatisierten Moderationsverfahren. Zusätzlicher Aufwand entsteht für Dienste weiterhin im Rahmen der Schnittstellenanbindung und im Bereich der manuellen Moderation.

[58] Das Produkt von Microsoft ist sogar derzeit noch nicht offiziell veröffentlicht, sondern befindet sich noch in der „Preview"-Phase (Stand: September 2023).

Dies führt dazu, dass diese Instrumente bislang eher von größeren Plattformen wie Reddit oder der New York Times[59] genutzt werden bzw. von größeren Kunden, die bereits andere Clouddienstleistungen des Anbieters (z. B. AWS, Azure etc.) einsetzen.

Personaldienstleister, die in deutscher Sprache moderieren, nutzen hierfür Moderationsstandorte im Inland oder in anderen Ländern der EU wie Irland oder Malta, in denen Moderatoren arbeiten, die der deutschen Sprache mächtig sind. Maschinelle Übersetzungen werden nicht als Grundlage von menschlichen Moderationsentscheidungen genutzt. Hierfür ist lt. Aussage der Anbieter der lokale (kulturelle) Kontext zu entscheidend. Dies gilt insbes. für die Erkennung von Hassrede und der Verhinderung terroristischer Inhalte.[60] Zudem kommt es innerhalb der Dienstleister zu einer Spezialisierung einzelner Moderatoren nach Themenfeld bzw. Rechtsbereich.

Tab. 12 Anbieter für Moderationsverfahren und Lösungen, die auch Moderationscenter betreiben (Stand: Juli 2023)

Anbieter	Moderations-verfahren	Hinweise
Besedo	KI und manuelle Moderation	Bieten eine All-in-One-Lösung (Training eines Algorithmus und menschliche Moderation) oder auch Einzelkomponentenlösungen an.
Genpact	Manuelle Moderation	Moderieren Inhalte aus über 30 Ländern.
Majorel	Manuelle Moderation	Arbeiten international aus 20 Standorten mit 80.000 Mitarbeitern and manueller Inhaltemoderation.
Pexly	Manuelle Moderation	Moderieren Inhalte aus über 45 internationalen Standorten.
Telus International	KI und manuelle Moderation	Bieten eine All-in-One-Lösung (Training eines Algorithmus und menschliche Moderation) oder auch Einzelkomponentenlösungen an.
Webhelp	KI und manuelle Moderation	Bieten eine All-in-One-Lösung durch automatisierte KI- und manuelle Moderation.

Quelle: Goldmedia Analyse 2023

[59] Vgl. https://www.perspectiveapi.com/case-studies, abgerufen am 20.09.23

[60] Lokaler Kontext wird vor allem sprachlich oder über das Themenfeld definiert, da die genaue Bestimmung des Ursprungslandes einer Äußerung bei Online-Plattformen nicht möglich bzw. aufgrund des globalen Charakters der großen Plattformen auch nicht zielführend wäre. Im Fall der Terrorismusbekämpfung ist die Domänenexpertise, die in der Regel eine regionale Komponente beinhaltet, wichtiger als die exakte Bestimmung des Ursprungslands einer Äußerung.

3.4.3 Automatisierte Moderationsverfahren der sehr großen IT-Konzerne

Im Folgenden werden die automatisierten Moderationsinstrumente der sehr großen IT-Konzerne Microsoft (Azure), Alphabet (Jigsaw) und Amazon (AWS) näher betrachtet.

Microsoft: Azure AI Content Safety

Azure AI Content Safety[61] wurde 2023 vorgestellt und ist aktuell erst eingeschränkt in einer „Preview"-Phase verfügbar (Stand: September 2023). Azure AI bietet eine KI-basierte, multi-modale Lösung zur Inhaltemoderation, primär entwickelt für die Bereiche Gaming, Soziale Medien, E-Commerce, E-Learning, Medien und Werbung. Azure AI überprüft in sozialen Netzwerken u. a. Chats, Bilder, Nutzername und Avatare. Azure AI Content Safety kommt unter anderem bei ChatGPT zum Einsatz.

Für Gaming-Umgebungen ist Azure AI besonders geeignet, da auch Live-Streams von Multi-Player Games sowie zugehörige Chatverläufe durch das System überwacht werden können. Der spezialisierte Dienstleister Two Hat Security, der mit seiner Moderationslösung Community Sift bereits langjährige Erfahrungen in der Inhaltemoderation auf der Microsoft-Plattform Xbox Live verfügt, wurde 2021 von Microsoft vollständig übernommen.[62] Die Code-Basis von Community Sift ist auch Bestandteil des neuen Moderationsinstruments Azure AI Content Safety.

Azure AI Content Safety setzt Natural Language Processing ein, um die Bedeutung und den Kontext von Sprache semantisch zu verstehen und zu überprüfen. Dies funktioniert grundsätzlich sprachenunabhängig. Aktuell werden acht Sprachen durch Azure AI Content Safety unterstützt. Die Bilderkennung beruht auf dem Project Florence der Microsoft AI Cognitive Services Initiative.[63]

Die KI erkennt schädliche Inhalte wie Hass, sexuelle Inhalte, Selbstverletzung und Gewalt. Jeder Kategorie wird ein Schweregrad von 0 bis 6 zugeordnet. Basierend auf diesem Schweregrad kann ein Unternehmen die markierten Inhalte überprüfen, nach Prioritäten ordnen und entsprechende Maßnahmen ergreifen. Dies ist in gewissen Freiheitsgraden durch den Kunden konfigurierbar. Kunden können die gewählten APIs in ihr Moderationsmanagementsystem integrieren und zusätzlich das Content Safety Studio zum Erkunden und Testen der Funktionen des Dienstes nutzen.

Azure AI Content Safety wird variabel, in Abhängigkeit der zu moderierenden Text- und Bildmenge berechnet. In Deutschland wird das Produkt noch nicht angeboten, in der Schweiz belaufen sich die Kosten für eine kleine Online-Plattform für die Analyse von Nutzerkommentaren mit einer durchschnittlichen Länge von 1.000 Zeichen und bei einem Volumen von 100.000 Kommentaren mit dem Produkt „Language Understanding (LUIS) Standard" pro Monat auf 138,60 Euro. [64]

Die Erkennung von Bildern zu Moderationszwecken (mit dem Produkt „Content Moderator S0") kosten bei Azure AI Content Safety für 100.000 Bilder etwa 92,40 Euro (Stand: September 2023).[65]

---

[61] Vgl. https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety, abgerufen am 22.09.23
[62] Vgl. https://www.crunchbase.com/organization/community-sift, abgerufen am 21.09.23
[63] Vgl. https://www.microsoft.com/en-us/research/project/projectflorence/, online abgerufen am 18.09.23
[64] Vgl. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/, online abgerufen am 20.09.23
[65] Vgl. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/, online abgerufen am 20.09.23

Alphabet: Perspective by Jigsaw

Perspective ist ein Projekt von Jigsaw, einer altruistischen Tochtergesellschaft von Alphabet, die sich auf die Entwicklung von Technologien zur Bekämpfung von Online-Missbrauch und Online-Desinformation konzentriert.

Perspective ist ein Moderationsinstrument, das maschinelles Lernen und künstliche Intelligenz (KI) verwendet, um den Ton und die Qualität von Kommentaren in Online-Diskussionen zu bewerten. Perspective analysiert Textkommentare und gibt ihnen eine Bewertung hinsichtlich ihres "Toxicity Score" (Toxizitätsbewertung), die anzeigt, wie wahrscheinlich es ist, dass ein Kommentar beleidigend oder unangemessen ist.[66] Es entstand 2017 aus den Erfordernissen des Google "Counter Abuse Technology Teams"[67] und wurde zunächst bei der New York Times trainiert und eingesetzt. Das Instrument kommt auch bei weiteren Publishern zum Einsatz, ist aber auch Teil der Community-Management-Plattform Coral und Disqus[68] (vgl. Tab. 11), sodass Perspective auch in vielen weiteren Online-Communities zum Einsatz kommt. So nutzt in Deutschland etwa das Magazin Der Spiegel die Community-Management-Plattform Coral.[69]

Die Idee hinter Perspective ist es, die Online-Kommunikation sicherer und respektvoller zu gestalten, indem es Plattformen und Nutzern Werkzeuge zur Verfügung stellt, um den Ton der Diskussionen zu verbessern und den Missbrauch und die Verbreitung von Hassrede einzudämmen. Perspective ermöglicht so etwa das Echtzeit-Feedback an postende User. Einige Plattformen setzen daher Perspective dafür ein, Nutzern eine unmittelbare Rückmeldung in Bezug auf mögliche Toxizität zu ihren Beiträgen zu geben. Ein zusätzliches Merkmal von Perspective API ist die individuelle Kontrolle über die erfahrene Toxizität aus Nutzersicht. Die API bietet daher auch ein Werkzeug, mit den Grad der Toxizität von Kommentaren, denen man online begegnen möchte, selbst zu bestimmen.

Das Instrument läuft auf den Servern von Jigsaw und steht der Allgemeinheit kostenfrei zur Nutzung zur Verfügung. Dienste können die API von Perspective einbinden, um automatisierte Moderation zu implementieren oder um menschlichen Moderatoren dabei zu helfen, Kommentare effizienter zu überprüfen, ohne hierfür zahlen zu müssen. Eine Kommerzialisierung des Dienstes ist derzeit nicht geplant.

Amazon: Amazon Recognition und Amazon Comprehend

Amazon Comprehend ist ebenfalls ein Moderationsinstrument, der maschinelles Lernen (NLP) nutzt. Er bietet eine umfangreiche Palette an Funktionen, die verschiedene Analysefunktionen umfassen, darunter benutzerdefinierte Klassifizierung, Schlüsselextraktion, Stimmungsanalyse, Ereigniserkennung und Entitätserkennung. Der Dienst wird für eine Vielzahl an Ausgangssprachen angeboten. Die Nutzung von AWS-Servern ist Voraussetzung für die Nutzung des Dienstes.

Mit der Erkennungsfunktion von Comprehend wird es möglich, Ereignisstrukturen aus unstrukturierten Daten zu extrahieren. Dies ermöglicht es, relevante Informationen aus großen Texten zu filtern und für die weitere Nutzung in KI-Anwendungen aufzubereiten.

---

66 Vgl. https://www.perspectiveapi.com/#/home, abgerufen am 22.09.23

67 Vgl. https://jigsaw.google.com/the-current/toxicity/countermeasures/, abgerufen am 20.09.23

68 Vgl. https://www.perspectiveapi.com/case-studies/, abgerufen am 20.09.23

69 Vgl. ebd.

Der Dienst erkennt und benennt Entitäten wie Personen, Orte, Unternehmen und vieles mehr, die im gegebenen Text auftreten. Diese automatisierte Identifizierung der Entitäten ermöglicht eine effiziente Klassifizierung von Texten.

Die Targeted Sentiment-Analyse von Comprehend bietet darüber hinaus detaillierte Erkenntnisse über die Stimmung in Texten. Hierbei werden nicht nur grobe Stimmungsindikatoren wie positiv, neutral oder negativ erkannt, sondern es erfolgt eine präzise Identifizierung der Stimmung in Bezug auf bestimmte Entitäten im Text.

Amazon Recognition identifiziert Bilder und Videos mit unangemessenem oder sensiblem Inhalt, wie beispielsweise anstößige Darstellungen. Für jedes Bild oder Video wird die Wahrscheinlichkeit errechnet, ob es sich um einen problematischen Inhalt handelt.

Eine weitere Eigenschaft von Recognition ist die Möglichkeit, eigene Moderationsmodelle zu trainieren und einzusetzen. Unternehmen können auf dieser Basis spezifische Modelle entwickeln und implementieren, die auf ihre individuellen Anforderungen und Gemeinschaftsrichtlinien zugeschnitten sind.

Der Preis für die Nutzung des Dienstes ist variabel und wird von der Anzahl an Zeichen (bei Text) bzw. der Anzahl der Bilder berechnet, die überprüft werden.

Die Analyse von Nutzerkommentaren mit einer durchschnittlichen Länge von 1.000 Zeichen und bei einem Volumen von 100.000 Kommentaren pro Monat kostet mit Amazon Comprehend rd. 100 US-Dollar pro Monat. Die Erkennung von Bildern (Bild-Label-Erkennung und Bildeigenschaften) kosten mit Amazon Recognition für 100.000 Bilder etwa 175 US-Dollar (Stand: September 2023).[70]

Die Inhaltemoderation für Video-Inhalte wird minutenbasiert abgerechnet. Für die automatisierte Inhaltemoderation mit Amazon Recognition werden für die Moderation von 100.000 Videos mit einer durchschnittlichen Länge von 1 Minute beim aktuellen Minutenpreis von 0,12 USD/min. [71] demnach 12.000 US-Dollar berechnet (Stand: September 2023). [72]

3.4.4 Integration von technischen Lösungen und Dienstleistern

Neben eigenentwickelten Moderationssystemen werden auch eingekaufte Moderationssysteme sehr individuell an die jeweilige Online-Plattform angepasst. Systemwechsel sind daher eher unüblich. In der Regel arbeiten Online-Plattform und Anbieter der Moderationslösung langjährig zusammen und entwickeln die eingesetzten Systeme gemeinsam weiter.

Allerdings können einzelne Komponenten, wie z. B. Filtersysteme, relativ leicht in bestehende Konfigurationen integriert werden. Der technische Anschluss eines Dritt-Systems an eine bestehende Moderationsmanagementplattform stellt keine besondere Herausforderung dar, da diese Plattformen auf modulare Erweiterbarkeit hin entwickelt sind.

---

70 Vgl. https://aws.amazon.com/de/rekognition/pricing/, online abgerufen am 21.09.23

71 Vgl. https://aws.amazon.com/de/comprehend/pricing/, online abgerufen am 21.09.23

72 Für die Online-Plattform YouTube, auf die nach Unternehmensangaben ca. 500 Stunden Videos pro Minute hochgeladen werden, bedeutete dies einen rechnerischen Aufwand für Amazon Recognition von 62,2 Mio. US-Dollar pro Monat.

Die Einbindung eines weiteren Filtersystems über Schnittstellen in eine bestehende Moderationslösung bedeutet – je nach System – ein Entwicklungs- bzw. Anpassungsaufwand von einigen Tagen bis wenigen Wochen. Soll ein komplexes KI-basiertes System eingebunden werden, kann das Training der KI mehrere Monate in Anspruch nehmen, bis optimale Ergebnisse erzielt werden. Die Grundverfügbarkeit eines Systems auf Basis eines generischen Trainingssets sollte jedoch bereits in wenigen Wochen zur Verfügung stehen.

Bei Personaldienstleistungen ist der technische Integrationsaufwand ebenso schnell zu bewerkstelligen. Allerdings kann der Anstellungs- und Schulungsaufwand für Moderatoren deutlich zeitaufwändiger sein. Größere Dienstleister können aufgrund ihres Personalbestandes ggf. auf Neuanstellungen verzichten und das benötigte Personal für ein neues Team mithilfe bestehender Personalressourcen bestreiten. Jedoch kann für den Schulungsaufwand von einem mindestens 4-6-wöchigen Prozess ausgegangen werden.

4 Praxis der Inhaltemoderation in Deutschland

Im folgenden Kapitel wird die Praxis der Inhaltemoderation in Deutschland näher beschrieben. Dabei wird zunächst auf große soziale Netzwerke eingegangen, die bislang unter das NetzDG fielen. Hierbei wird zunächst eine Analyse der Angaben der NetzDG-Transparenzberichte vorgenommen. Im Anschluss daran werden einige ausgewählte große Online-Plattformen eingehender porträtiert, indem hierbei zusätzlich zu den Angaben aus den öffentlichen Transparenzberichten auch Angaben aus Gesprächen aufbereitet werden, die mit den Anbietern geführt wurden.

Im darauffolgenden Kapitel werden analog dazu kleine Online-Plattformen und – Anbieter porträtiert. Die Darstellung in diesem Kapitel stützt sich in einem größeren Umfang auf die Angaben aus den Gesprächen, da diese Anbieter bislang keine dem NetzDG vergleichbaren Berichtspflichten unterliegen.

Im darauffolgenden Kapitel werden die Erkenntnisse zum Aufwand für Inhaltemoderation kurz zusammengefasst, bevor im letzten Unterkapitel die Bedeutung terroristische Inhalte auf Online-Plattformen anhand der vorliegenden TCO-VO-Transparenzberichte ausgewählter Anbieter ausgewertet wird.

4.1 Große Soziale Netzwerke

Seit 2018 gilt das Netzwerkdurchsetzungsgesetz (NetzDG) für „soziale Netzwerke", die im Inland mehr als zwei Millionen registrierte Nutzer haben. Das Gesetz wird im Februar 2024 durch den DSA abgelöst.

Dienste-Anbieter, die dem NetzDG unterfallen, sind verpflichtet, halbjährlich über ihre Maßnahmen zur Inhaltemoderation Bericht zu erstatten. Diese Transparenzberichte gewähren Einblick in das Moderationsaufkommen und die Moderationsprozesse großer Online-Plattformen („Sozialer Medien"), die in Deutschland operieren. Allerdings umfas-

sen die Transparenzberichte nur den Teilbereich der strafrechtlich justiziablen Inhaltemoderation, der explizit über das NetzDG geregelt ist. Andere große Teilbereiche der Inhaltemoderation von Online-Plattformen, etwa im Bereich Betrugsbekämpfung und -vorkehrung, werden nicht von den Transparenzpflichten des NetzDG mitumfasst.

Um den gesetzlichen Anforderungen des NetzDG zu entsprechen, ein wirksames und transparentes Verfahren für den Umgang mit Beschwerden über rechtswidrige Inhalte vorzuhalten[73], wurden von den großen Online-Plattformen separate Moderationsprozesse (Cues) eingerichtet, die speziell für die Bearbeitung von NetzDG-Anfragen geschult sind. Nutzer eines Dienstes können einen strafrechtlich relevanten Beitrag in der Regel direkt an ein NetzDG-Team melden, oft wird hierbei der reguläre Moderationsprozess (nach den Gemeinschaftsrichtlinien des Anbieters) umgangen. In der Regel entscheiden die Nutzer bei der Meldung, nach welchem Prozess die Moderation erfolgen soll, in der Regel nach Gemeinschaftsrichtlinien oder nach NetzDG.

Für den Anbieter ist diese Unterscheidung in der Regel immateriell, da die Gemeinschaftsstandards zum weit überwiegenden Teil strenger sind als die gesetzlichen Standards des Strafrechts.[74] Hieraus resultieren einige methodische Herausforderungen bei der Interpretation der berichteten Daten der Online-Plattformen. Bei den Angaben in den Transparenzberichten kann verallgemeinernd nicht davon ausgegangen werden, dass sämtliche Inhaltemoderationen erfasst sind, die NetzDG-relevant sind, sondern nur, ob diese von den NetzDG-Teams moderiert wurden.[75] Die Angaben innerhalb der Transparenzberichte sollten also mit Vorsicht interpretiert werden, da sie im Allgemeinen keine Schlüsse auf das gesamte Moderationsgeschehen der jeweiligen Plattform zulassen.

Trotz dieser Einschränkungen für die quantitativen Angaben innerhalb der Berichte lassen sich den Transparenzberichten wesentliche qualitative Informationen zu den Moderationsprozessen der Plattformen entnehmen. Im Folgenden werden ausgewählte Angaben aus den NetzDG-Transparenzberichten führender Online-Plattformen tabellarisch zusammengefasst.

[73] Vgl. § 3 Abs. 1 NetzDG

[74] Nur in einigen Bereichen sind die Anforderungen des deutschen Strafrechts strenger als die Gemeinschaftsrichtlinien großer international operierender Sozialer Netzwerke, etwa bei der Darstellung von Kennzeichen verfassungswidriger Organisationen.

[75] Die Prozesse unterschieden sich dabei von Anbieter zu Anbieter, so gibt es durchaus Anbieter, die deutsche Inhalte grundsätzlich von zwei verschiedenen Teams prüfen lassen (Gemeinschaftsrichtlinien und NetzDG), allerdings ist dies nicht die Regel.

Tab. 13 Zentrale Aussagen der NetzDG-Berichte ausgewählter Online-Plattformen, Juli-Dezember 2022

Plattform	Nutzer tägl. in Dtl. *	Anz. Meldungen nach NetzDG	Automatisierte Erkennung	Organisation	Personelle Ausstattung
Facebook	14,11 Mio.	- 34.806 entfernte oder gesperrte Inhalte - davon 33.700, die gegen die Gemeinschaftsstandards verstoßen - davon 1.106 Verstöße gegen das NetzDG	- Rate limits (um Bots zu verhindern) - Abgleiche (Content Hashing) - Machine Learning: Entfernung automatisch, falls KI hinreichend sicher	Überprüfung erfolgt in zwei Stufen: geschulte Teams und Juristen	178 Personen in drei Teams zur Bearbeitung von NetzDG-Beschwerden
YouTube	25,4 Mio.**	- 233.440 Meldungen nach NetzDG - davon 5.166 Meldungen rechtswidrig - davon 109 Terror- Entfernungen	Automatisierter maschineller Abgleich: Verwendung von Hashes (digitale Fingerabdrücke) Automatische maschinelle Meldungen, um Inhalte einer menschlichen manuellen Prüfung zuzuführen	NetzDG Team in zwei Schichten, rund um die Uhr. Stufen bestehen aus Sachbearbeiter, Senior Content Reviewer, Rechtsabteilung von YT und Google sowie externe Kanzleien mit Strafrechtexpertise.	Spezielles Team für NetzDG-Beschwerden, bei externem Dienstleister in Deutschland werden 77 Personen beschäftigt
Twitter ***	2,82 Mio.	- Anzahl der eingegangen NetzDG-Beschwerden: 947.994 Anzahl der NetzDG Beschwerden mit Maßnahmen: 153.416	Verwendung von Heuristiken (Schlüsselwortmuster) sowie maschinelle Lernmethoden um auf neue Formen an Richtlinienverstöße zu reagieren	Spezielles Team für NetzDG-Meldungen nachgeschaltet: Erst Überprüfung auf Richtlinienverstöße, erst dann auf NetzDG-Verstöße	Netz-DG-Team: 150 Personen, 7 % direkt bei Twitter angestellt, alle weiteren bei Vertragspartnern

Plattform	Nutzer tägl. in Dtl. *	Anz. Meldungen nach NetzDG	Automatisierte Erkennung	Organisation	Personelle Ausstattung
Instagram	14,81 Mio.	- 4.273 entfernte Inhalte - davon 4.155 Verstößen gegen die Gemeinschafts-richtlinien, - 118 Verstöße gegen das NetzDG	- Rate limits (um Bots zu verhindern) - Abgleiche (Content Hashing) - Machine Learning: Entfernung automatisch, falls KI hinreichend sicher	Überprüfung erfolgt in zwei Stufen: geschulte Teams und Juristen	178 Personen in drei Teams zur Bearbeitung von NetzDG-Beschwerden
reddit	0,71 Mio.	- 1.066 NetzDG-Beschwerden - Hiervon führten 674 zu einer Entfernung oder Sperrung	- Einsatz von automatisierten Tools auf der Plattform für Verstöße gegen Inhaltsrichtlinien - Keine automatisierten Tools für NetzDG-Meldungen	Safety-Team für allgemeine Inhaltsverstöße Community-Team für Moderationsverstöße Plattform & Legal Policy Team für Entfernung rechtswidriger Inhalte	Plattform & Legal Policy Team besteht aus 12 Spezialisten Hiervon sind 4 speziell für die Bearbeitung von NetzDG-Beschwerden
- **Twitch**	1,41 Mio.	- Insgesamt 37.607 Meldungen - davon 929 Meldungen nach NetzDG	- Einsatz von Machine-Learning-Modellen um gegen anstößige Benutzernamen, Spam, Betrug, anstößige Emotes und Bot-Konten vorzugehen	- Zuerst Überprüfung auf Richtlinienverstöße, erst dann auf NetzDG-Verstöße, rechtswidrige Inhalte werden innerhalb 24 Stunden gesperrt	- Mind. 15 Moderatoren zu jeder Zeit für NetzDG-Meldungen. Falls genauere Prüfung erforderlich: Eskalation an internes Spezialistenteam. Bei komplexen Fällen auch an Juristen
Soundcloud	2,12 Mio. pro Woche	- Insgesamt 114 NetzDG-Meldungen - 47 weitere, nicht explizit NetzDG-bezogene Meldungen	Nutzen keine KI	Bearbeitung von NetzDG-Beschwerden über internes und externes Trust and Safety Team	Internes Trust & Safety Team: 12 Mitarbeiter Externes Trust & Safety Team: 6 Personen, dazu Rechts-abteilung zur Unterstützung

Plattform	Nutzer tägl. in Dtl. *	Anz. Meldungen nach NetzDG	Automatisierte Erkennung	Organisation	Personelle Ausstattung
TikTok	5,64 Mio.	- 226.479 NetzDG-Beschwerden - 24.534 Entfernungen oder Sperrungen nach NetzDG - 20.051 Entfernungen oder Sperrungen nach den Gemeinschaftsrichtlinien	Hochgeladene Videos durchlaufen automatisierte Überprüfung zur Erkennung und Klassifizierung (z. B. terroristische Symbole), ggf. automatisierte Entfernung oder Kennzeichnung der Videos für manuelle Moderation	Entwicklung von Moderationsrichtlinien durch Trust-and-Safety-Team (SIC!); Überprüfung der Inhalte durch Moderationsteams	Spezielles NetzDG-Team mit 28 Mitgliedern, hiervon sind 11 bei einem externen Dienstleister angestellt
Pinterest	4,94 Mio.	- 55 NetzDG Beschwerden - 5.406 im Zusammenhang mit einer Verletzung der Gemeinschaftsrichtlinien	- Automatisierte Tools kennzeichnen Inhalte die gegen die Gemeinschaftsrichtlinien verstoßen - Machine-Learning Modelle, die Bilder bewerten und bereits Durchsetzungsentscheidungen treffen können	Bearbeitung der Beschwerden von NetzDG-Team; Bearbeitung von komplexen Fällen durch Trust & Safety Leads und Rechtsabteilung sowie externe deutsche Rechtsberater	6 Personen bewerten die Beschwerden; 4 Mitarbeiter sind für die Moderation zuständig; Trust & Safety umfasst 2 Mitarbeiter

** Goldmedia-Berechnung auf Basis: ARD-ZDF-Onlinestudie 2022.*
*** Goldmedia-Berechnung auf Basis: die medienanstalten 2022. Intermediäre und Meinungsbildung 2022-I, *** Angaben von Twitter vor der Übernahme durch Elon Musk*
Quelle: Goldmedia Analyse 2023 auf Basis Bundesanzeiger

Die tabellarische Zusammenstellung offenbart teilweise recht deutliche Unterschiede bei der personellen Ausstattung der Online-Plattformen in Bezug auf die NetzDG-Moderation, wenn man diese in Beziehung zur Größe der Plattform bzw. zum Meldungsaufkommen setzt. Allerdings konnten uns Branchenexperten, Dienste-Anbieter und Moderationsdienstleister diese Beobachtung in Hintergrundgesprächen nicht in dem Umfang bestätigen, wie es die Angaben in den Transparenzberichten nahelegen. Zwar unterscheidet sich der Moderationsaufwand aufgrund plattformspezifischer Faktoren wie Ausrichtung, Zielgruppe und Sensibilität der Inhalte[76], jedoch sei der menschliche Moderationsaufwand sei im Wesentlichen vergleichbar, zumindest unter Berücksichtigung

[76] Gewisse Themenfelder, wie z. B. Politik oder einzelne Computerspiel-Titel, gelten grundsätzlich als moderationsintensiv, da es in diesen gehäuft zu Meldungen, u. a. im Bereich der hasserfüllten Rede, kommt.

plattformspezifischer Unterschiede. Die größeren Abweichungen, die die Transparenzberichte nach NetzDG nahelegen, scheinen vor allem auf unterschiedliche Erfassungsmetriken und definitorische Unschärfen zurückzugehen (s. u.), da die Gemeinschaftsrichtlinien in der Regel sehr große Schnittmengen zum NetzDG aufweisen. Wie hoch der NetzDG-relevante Moderationsaufwand innerhalb der Moderationsteams ist, die nur nach den Gemeinschaftsrichtlinien moderieren, ist vielen Anbietern dabei selbst nur schwer bis kaum abgrenzbar. In der nachfolgenden Tabelle wird das NetzDG-Moderationsaufkommen, insbesondere mit Bezug zu terroristischen Inhalten ausgewählter Plattformen, näher dargestellt.

Tab. 14 Zentrale Kennziffern der NetzDG-Berichte ausgewählter Online-Plattformen, Juli-Dezember 2022

Zeitraum	07.-12.22	07.-12.22	07.-12.22	07.-12.22	01.-06.22
Plattform	**Facebook**	**Instagram**	**YouTube**	**Twitter/ X**	**Twitch**
NetzDG-Beschwerden	**125.195**	**57.541**	**233.440**	**947.994**	**44.299**
davon "Terrorismus"* in %	21,1%	23,8%	k. A.	9,8%	k. A.
davon "Terrorismus und verfassungswidrig"** in %*	31,9%	35,4%	6,8%	15,2%	8,6%
Entfernte Inhalte	**34.806**	**4.273**	**32.150**	**153.416**	**2.894**
davon "Terrorismus" in %	6,7%	6,0%	k. A.	1,9%	k. A.
davon "Terrorismus und verfassungswidrig" in %	28,0%	34,4%	6,3%	19,8%	7,7%
Entfernung <24h	**92,9%**	**93,2%**	**85,6%**	**97,5%*** **	**98,6%**
davon Terrorismus in %	k. A.		87%	k. A.	k. A.
Entfernung >7 Tage	**0,3%**	**0,6%**	**0,4%**	**0,01%**	**0,1%**
davon Terrorismus* in %	k. A.		0,2%	k. A	k. A
Berechtigte Widersprüche	3.173	883	k. A.	k. A.	68
Anteil berecht. Widersprüche an entfernten Inhalten in %	**9,1%**	**20,7%**	**k. A.**	**k. A.**	**2,3%**
Personal NetzDG-Moderation lt. Berichten	178		77[77]	150	45
Stellen (in Vollzeit, Schätzung Goldmedia)** **	**118,7**		**51,3**	**100,0**	**30,0**
Beschwerden pro Vollzeitstelle in 6 Monaten	1.540		4.548	9.480	1.477

** „Terrorismus" umfasst die Paragrafen § 86, § 86a, § 89a, § 91, § 100a, § 129a und § 129b StGB*
***„Terrorismus und verfassungswidrige Inhalte" umfasst zusätzlich § 129, § 140, § 269 StGB*
**** Angabe beruht abweichend auf der Bearbeitung der NetzDG-Beschwerden und nicht auf der Bearbeitung der nach NetzDG entfernten Inhalte.*
***** Es wird pauschal angenommen, dass sich die Angaben zuModerationsteams aus 50 Prozent Vollzeitkräften und 50 Prozent Teilzeitkräften mit einer halben Vollzeitstelle zusammensetzen.*
Quelle: Goldmedia Analyse 2023

Eine grundsätzliche Herausforderung liegt bei der Interpretation der Angaben darin begründet, dass Online-Plattformen keine einheitliche Handhabe für Inhalte haben, die

[77] EU-weit kommen bei YouTube laut Presseberichten 16.974 Moderatoren zum Einsatz, vgl. https://ch.marketscreener.com/kurs/aktie/ALPHABET-INC-24203373/news/Musks-X-hat-nur-einen-Bruchteil-der-Moderatoren-der-Konkurrenz-sagt-die-EU-45299109, online abgerufen am 13.11.2023

sowohl nach den eigenen Gemeinschaftsrichtlinien als auch nach NetzDG nicht zulässig sind. Einige Anbieter prüfen Inhalte grundsätzlich nach beiden Maßstäben, während andere Anbieter im Falle einer Nicht-Zulässigkeit auf die Prüfung nach dem anderen Maßstab verzichten.

Auch sind die Angaben der Transparenzberichte verschiedener Plattformen mit Bezug zu terroristischen Inhalten nur bedingt miteinander vergleichbar. Plattformen wie Facebook oder Instagram weisen zwar die jeweils betroffenen Straftatbestände des StGB pro Beschwerde auf, allerdings kommt es hierbei in einem nicht bekannten Umfang zu Doppelausweisungen. Die prozentualen Angaben auf Beschwerde-Ebene dürften daher stark nach oben verzerren. Plattformen wie YouTube und Twitch gruppieren die Beschwerden in größere Gegenstandsbereiche – allerdings ist der hierbei gewählte Maßstab von „Terrorismus und verfassungswidrigen Inhalten deutlich umfangreicher als die Terrorismusdefinition nach TCO-VO, sodass auch diese Angaben stark nach oben verzerren.

Aussagekräftiger sind daher die Angaben zu den entfernten Inhalten. Bei den betrachteten Online-Plattformen liegt die der Anteil der terroristischen bzw. terroristischen oder verfassungswidrigen Inhalten zwischen 2 Prozent und 7 Prozent der auf Grundlage der nach NetzDG entfernten Inhalte. Dabei liegt die Zeit von der Meldung bis zur Entfernung bei den betrachteten Plattformen zwischen 85,6 Prozent und 98,6 Prozent. Der Anteil der Gegenvorstellungen, denen nach erneuter Prüfung entsprochen wurde, liegt bei den betrachteten Online-Plattformen zwischen 2,3 Prozent und 20,7 Prozent.

Die Angaben zum Moderationspersonal sind in den Transparenzberichten sehr unscharf definiert, und werden mitunter als „zur Verfügung stehend“[78] bezeichnet, was einen erheblichen Interpretationsspielraum zulässt und sich nicht auf das Personal beschränkt, dass unmittelbar zur Inhaltemoderation eingesetzt wird („tätig“). Zudem macht etwa der Meta-Konzern inhaltsgleiche Aussagen zum eingesetzten Personal sowohl für Facebook als auch Instagram, was stark darauf hindeutet, dass hierbei dasselbe Personal für beide Plattformen tätig wird.[79] Die Aussagen zum durchschnittlichen Moderationsaufkommen pro Vollzeitstelle sind daher nur bedingt geeignet, um einen Personalschlüssel für Inhaltemoderation abzuleiten.

4.1.1 Sehr große Online-Plattform im Bereich soziale Netzwerke

Proaktive, automatisierte Moderationsverfahren sind grundsätzlich ein Industriestandard, wie die sehr große Online-Plattform TikTok betont. Alle automatisierten Moderationsverfahren und -instrumente werden bei TikTok intern entwickelt.

Der Einsatz von KI in der Moderation wird grundsätzlich als sehr interessante Entwicklung gesehen, die sich innerhalb der Branche schnell entwickelt, und bei der das Ende der Entwicklung nicht absehbar ist. Anwendungsfelder liegen sowohl bei der Durchsetzung der Gemeinschaftsrichtlinien (Erkennung von Inhalten), aber auch in nutzerweisenden Merkmalen zur Unterstützung der Nutzererfahrung.

---

[78] Vgl. Facebook: NetzDG Transparenzbericht Januar 2023, Abschnitt 5B

[79] Hierauf wird jedoch in den Transparenzberichten der beiden Plattformen nicht hingewiesen. In Tab. 14 wird dies jedoch unterstellt und bei der Bildung des Quotienten in der letzten Zeile entsprechend berücksichtigt.

Bei der automatisierten audiovisuellen Moderation ist der Einsatz von KI im Kommen, die Implementierung muss allerdings „extrem vorsichtig" geprüft werden. TikTok setzt auch auf ein fortgeschrittenes KI-gestütztes Verfahren zur Erkennung unerwünschter Inhalte, das konstant anhand der Moderationsentscheidungen der menschlichen Moderatoren weiter trainiert wird. TikTok ist auch offen für andere, neue Lösungen, aktuell werden bei TikTok weitere, auf NLP basierende audiovisuelle Erkennungsmechanismen implementiert. Deren Resultate werden aber von ihren Moderatoren manuell überwacht.

Die Leitfrage, die sich TikTok für die Zukunft stellt, ist: "Wie viel KI können wir sicher einsetzen"? Grundsätzlich gelten für KI-gestützte Verfahren die gleichen Standards wie für manuelle Moderationsprozesse. Generell müssten automatisierte KI-gestützte Verfahren so weit entwickelt sein, dass sie dasselbe oder ein noch höheres Sicherheitsniveau gewährleisten können.

Die manuelle Moderation bei TikTok findet branchenüblich an verschiedenen Standorten statt, neben den USA etwa in Dublin, London und in Deutschland. Auch, weil lokale Nuancen eine Schlüsselrolle bei der Erkennung von Hassrede und Terrorismus spielen. In den spezifischen Cues sind Moderatoren besonders geschult, um solche Verstöße bzw. Bedrohungen zu erkennen.

Das strategische Ziel beim Einsatz des KI-gestützten Verfahrens ist es, Fälle zu identifizieren, um die manuelle Moderation zu unterstützen. Maschinell identifizierte Fälle werden nach Unternehmensrichtlinien immer manuell moderiert. Insbesondere bei sensiblen Moderationsaufgaben wie bei Terrorismus wird weiterhin menschliche Moderation erforderlich sein, um die Sicherheit der Nutzer zu gewährleisten. Gerade bei solchen sensiblen Moderationsbereichen wie Terrorismus und Kindesmissbrauch ist die Zusammenarbeit mit Meldedatenbanken der Branche (vgl. Kap. 8.2), etwa von Tech Against Terrorism essentiell, um von den Erfahrungen anderer Online-Plattformen profitieren zu können.

Abschließend betont TikTok die Bedeutung von Transparenz gegenüber den Nutzern, welche Inhalte unerwünscht sind und wie Inhalte moderiert werden. Informationen darüber (Gemeinschaftsrichtlinien, Datenschutzeinstellungen, Möglichkeiten zur Eigenmoderation für Inhalteersteller, Regeln zur Kooperation mit Strafverfolgungsbehörden) sollten deshalb zentral in einem „Sicherheitscenter" vorgehalten werden und leicht für die Nutzer auffindbar sein. Dort wird auch ein Transparenzreport zur TCO-VO veröffentlicht („EU Terrorist Content Online Regulation (EU) 2021/784 Transparency Report").[80] Wesentliche Angaben aus diesem zeigt die folgende Tabelle:

[80] Vgl. https://www.tiktok.com/transparency/en/tco-report/, abgerufen am 22.09.23

Tab. 15 Spezifische Kennziffern zur Verhinderung terroristischer Inhalte bei TikTok

Kenngröße	Angabe	Zeitpunkt
Entfernungen (gesamt)	53.385	07.06.22- 31.12.22
Einsprüche von Nutzern (% an gesamt)	22,1 %	07.06.22- 31.12.22
Erfolgreiche Einsprüche, bei denen die Inhalte wiederhergestellt wurden (% an gesamt)	11,5 %	07.06.22- 31.12.22

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

Im Berichtszeitraum hat TikTok u. a. ein Netzwerk aus 6 Nutzerkonten und 368.644 Followern deaktiviert, dass aus Deutschland heraus operierte und mit unauthentischen Nutzerkonten versuchte, den Diskurs um die ägyptische Regierung zu beeinflussen und für ein rein ägyptisches Publikum bestimmt war. TikTok hat von den europäischen Behörden im Zeitraum vom 07.06.22-31.12.22 keine Entfernungsanordnungen gemäß der TCO-Verordnung erhalten.[81]

Allgemeine Kennziffern zur Inhaltemoderation bei TikTok zeigt folgende Tabelle:

Tab. 16 Allgemeine Kennziffern zur Inhaltemoderation bei TikTok

Kenngröße	Angabe	Zeitpunkt	Anmerkungen
Unique User pro Monat in Deutschland	20,6 Mio. monatl. 5,64 Mio. tägl.	2023 2022	Schätzung[82] vgl. Tab. 13
Durchschnittliche Nutzer beiträge pro Monat (Gesamt)	5,056 Mrd.	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
Durchschnittlich entfernte Nutzerbeiträge (an Gesamt)	0,6 Prozentpunkte	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
Durchschnittlich entfernte Nutzerbeiträge in Deutschland (an Gesamt)	0,0066 Prozentpunkte	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
Durchschnittlich entfernte Nutzerbeiträge im Bereich gewalttätiger Extremismus (an Gesamt)	0,0084 Prozentpunkte	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
davon proaktiv entfernt	94,9 %	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
davon entfernt vor Verbreitung	77,4 %	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
davon entfernt in unter 24 Stunden	85,9 %	1. Quartal 2023	TikTok Community Guidelines Enforcement Report
Erfolgreich durch Nutzer beanstandete Moderationsentscheidungen	6,8 %	1. Quartal 2023	TikTok Community Guidelines Enforcement Report

[81] Vgl. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-1/, abgerufen am 14.09.23

[82] Vgl. https://www.smart-home-fox.de/tiktok-nutzer-statistiken, abgerufen am 14.09.23

Mitarbeiter gesamt	450 Personen (nur Deutschland)	30.08. 2022	Verdi[83]
Angestellte Moderatoren des Dienstes	28 Personen (nur NetzDG)	2. Halb-jahr 2022	TikTok NetzDG-Transparenzbericht
Moderatoren für Deutsch-land (an Moderatoren ge-samt)	2,3 %	1. Quartal 2023	TikTok Community Guide-lines Enforcement Report

Quelle: Goldmedia Analyse 2023

4.1.2 Große Online-Plattform im Bereich Social Video

Mit zwischen 30 und 32 Millionen aktiven Nutzern weltweit zählt die Social-Video-Plattform Twitch nicht zu den sehr großen Online-Plattformen. Von aktiven Nutzern erstellen etwa 7 bis 8 Mio. aktiv Inhalte und verbreiten sie über den Dienst.

Auf der Social-Video-Plattform kommen sowohl externe aus inhäusig entwickelte automatisierte Verfahren zum Einsatz. Bei der manuellen Moderation arbeitet der Dienst branchenüblich mit einem externen Dienstleister zusammen.

Die spezifische Herausforderung für die Social-Video-Plattform ist, dass sie hauptsächlich Live-Inhalte bietet. Das bedeutet, dass alle Inhalte, die auf die Plattform hochgeladen werden, im gleichen Moment produziert werden, in dem sie verbreitet werden. Der Großteil der Inhalte verbleibt auch nicht auf der Plattform, sondern ist nur eine gewisse Zeit (zum Beispiel für 48 Stunden) zugänglich. Nur ein geringer Anteil der Inhalte auf der Plattform bleibt dauerhaft verfügbar.

Der aktuelle Trend zu verstärkter Automatisierung wird auch von diesem Dienst stark wahrgenommen. Es bestehen große Ambitionen bei den Dienstleistern und es werden viele Systeme entwickelt, auch innerhalb des Dienstes. So kommt intern u. a. die Machine-Learning-Anwendung „Ally" von Spirit AI für eigene Moderationslösungen zum Einsatz.

Die Ergebnisse der bisher getesteten externen Dienstleister noch einigermaßen ernüchternd. Die Produktversprechen übersteigen aktuell noch die tatsächliche Leistungsfähigkeit. Die Erkennung ist oft noch zu wenig kontextbasiert, um die spezifischen Inhalte des Dienstes moderieren zu können. Damit ist Zuverlässigkeit ist noch nicht in dem erforderlichen Maße gegeben. Eine Wunderwaffe seien die KI-gestützten Verfahren nicht.

Die Implementierung automatisierter Verfahren bei Live-Content ist deutlich schwieriger, weshalb sich die Sicherheitsarchitektur stark von anderen Anbietern unterscheidet, bei denen die Veröffentlichung erst nach dem Upload erfolgt. Der strategische Schwerpunkt liegt auf community-led moderation (vgl. Kap. 3.3.2.1). Vor allem hierfür werden (mitunter KI-gestützte) Tools entwickelt und zur Verfügung gestellt.

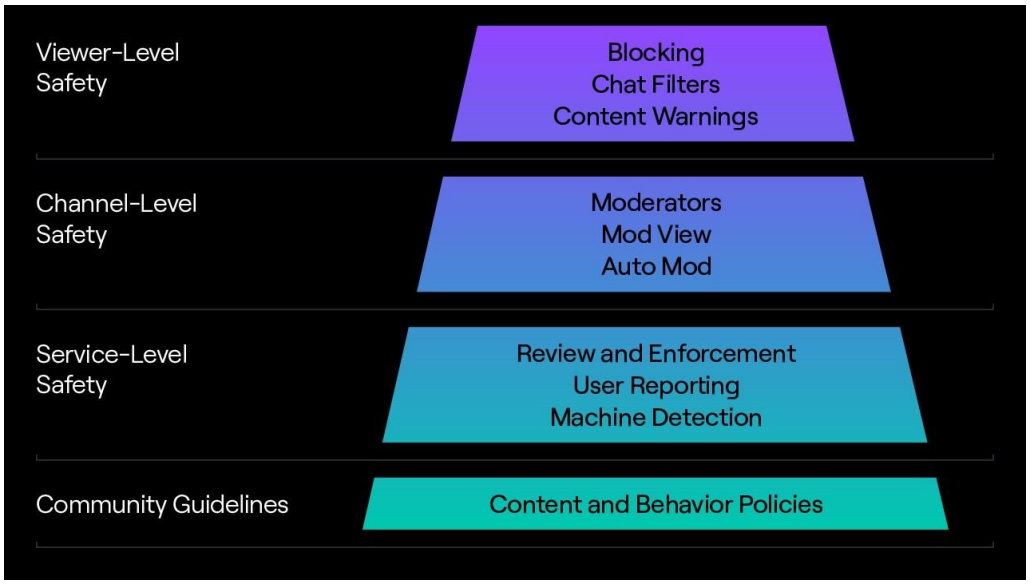
Mit besonderen Anwendungsszenarien, wie der Chat-Moderation besonders schriller Inhalte (z. B. Beleidigungen und andere „statische" Begriffe) können die NLP-basierten

[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.

Systeme gut umgehen. Auch die etablierten Hash-Matching-Tools liefern gute Resultate.

Abschließend betont der Anbieter, dass Sicherheit ein mehrschichtiges Konzept ist und die Suche nach einer alleinseligmachenden Lösung für die vielfältigen Herausforderungen in der Inhaltemoderation nicht sinnvoll ist. Der Mensch innerhalb des Moderationsprozesses wird auch künftig der Schlüssel zu einer erfolgreichen Inhaltemoderation bleiben.

Abb. 7 Mehrschichtiges Moderationskonzept der Social-Video-Plattform



Quelle: Twitch, online unter: https://safety.twitch.tv/s/article/Safety-at-Twitch?language=en_US, *abgerufen am 31.07.23*

Der Dienst veröffentlicht einen Transparenzreport zur TCO-VO („EU Terrorist Content Online Regulation 2022 Transparency Report").[84] Im Berichtszeitraum hat der Dienst von europäischen Behörden keine Entfernungsanordnungen gemäß der TCO-Verordnung erhalten.

Allgemeine Kennziffern zur Inhaltemoderation des Dienstes zeigt folgende Tabelle (vgl. Tab. 17). Abgebildet sind nur Moderationsentscheidungen (Enforcement Actions) der Moderatoren des Dienstes, die Moderationsentscheidungen im Rahmen der community-led moderation sind hierin nicht berücksichtigt. Entfernungen von Inhalten spielen auf der Plattform nur eine untergeordnete Rolle, da die Inhalte in der Regel beim Fällen einer Moderationsentscheidung bereits nicht mehr verbreitet werden[85], daher berichtet der Dienst nicht die Zahl der Entfernungen, sondern der sanktionsbewehrten Moderationsentscheidungen (enforcement actions).

[84] Vgl. https://safety.twitch.tv/s/article/2022-EU-Terrorist-Content-Transparency-Report?language=en_US, abgerufen am 22.09.23

[85] Sollte dies noch der Fall sein, würden beanstandete Inhalte durch die Moderatoren entfernt.

Tab. 17 Allgem. Kennziffern zur Inhaltemoderation des Social-Video-Dienstes

Kenngröße	Angabe	Zeitpunkt	Anmerkungen
Unique User pro Monat in Deutschland	3,58 Mio. monatl. 1,41 Mio. tägl.	März 2020 2022	agof daily digital facts vgl. Tab. 13
Durchschnittlich gesehene Stunden pro Monat	1,847 Mrd.	2. Hj 2022	Transparency Report H2 2022
Durchschnittliche Enforcement Actions pro Monat	194.000	2. Hj 2022	Transparency Report H2 2022
Durchschnittliche Enforcement Actions pro Monat im Bereich Terrorismus	17	2. Hj 2022	Transparency Report H2 2022
Erfolgreich durch Nutzer beanstandete Enforcement Actions	852	2. Hj 2022	Transparency Report H2 2022

Quelle: Goldmedia Analyse 2023

4.1.3 Große Online-Plattform im Bereich Social Gaming

Mit rd. 66 Millionen täglichen aktiven Nutzern weltweit (Stand: 1. Quartal 2023) und durchschnittlich 27,4 Mio. monatlich aktiven Nutzern in der EU (Stand: August 2023)[86] zählt die Social-Gaming-Plattform Roblox nicht zu den sehr großen Online-Plattformen.

Auf der Online-Plattform können Nutzer in einer intuitiven Entwicklungsumgebung ihre eigenen Spiele kreieren, Spiele entdecken, die von anderen Benutzern erstellt wurden und sich gemeinsam über ihre Spielerfahrungen auszutauschen. Die Online-Plattform ist besonders bei jüngeren Spielern beliebt, da die Spiele eine ausgeprägte soziale Komponente bieten, bei der Nutzer mit anderen chatten und zusammen spielen können. Fast 50 Prozent der Nutzer sind unter 13 Jahre alt.

Der Dienst betont die Bedeutung einer klaren Kommunikation der Gemeinschaftsstandards. Dazu gehört auch die Kommunikation der potenziellen Konsequenzen, die drohen, wenn nicht erwünschte Inhalte gepostet werden. Hierzu zählen neben Verwarnungen und der Entfernung solcher Inhalte auch die dauerhafte Sperrung des eigenen Nutzerkontos oder eine polizeiliche Meldung, falls eine unmittelbare Bedrohungslage besteht.

Der Dienst wendet sowohl technologische als auch manuelle Verfahren (rund um die Uhr) an, um Inhalte zu moderieren. Technisch werden Inhalte zunächst anhand von Branchendatenbanken mit bereits bekannten illegalen Inhalten (z. B. terroristische Inhalte und Material über sexuellen Missbrauch von Kindern) abgeglichen. Erwähnt werden in diesem Zusammenhang vor allem das European Internet Forum (EUIF) und die Initiative Tech Against Terrorism.

Zusätzlich zu diesen Branchendatenbanken werden Inhalte anhand einer eigenen Datenbank überprüft, in welcher die zuvor von der Plattform entfernten Inhalte als Hash-

[86] Vgl. https://en.help.roblox.com/hc/de/articles/13061336948244-Digital-Services-Act, abgerufen am 18.09.23

Werte gespeichert sind. Aufgrund der Plattformspezifik kommen Chat-Filtern eine wichtige Bedeutung zu, so wird u. a. die Lösung Community Sift von Two Hat Security eingesetzt[87]. Aber auch inhäusig entwickelten NLP-basierten Verfahren kommen vermehrt zum Einsatz.

Der Dienst gibt an, dass bis zu 1.000 Personen auf der Plattform Inhalte auf Basis der Gemeinschaftsrichtlinien manuell moderieren. Diese befinden sich in der Regel bei externen Dienstleistern in der Region des Ursprungslandes, da für die manuelle Moderation der lokale Kontext entscheidend ist. So verändern Nutzer bewusst die Schreibweisen so weit, bis automatisierte Filtersysteme nicht mehr anschlagen. Für die richtige Einordnung solcher Äußerungen („Leetspeek") ist die Kenntnis des lokalen Sprachgebrauchs daher wesentlich.

Unter den Inhalte-Moderatoren gibt es auch ein spezialisiertes (internes) Team, dass sich ausschließlich um die Prävention terroristischer Inhalte („terrorism and violent extremism", TVE) kümmert. Die Mitarbeiter im Bereich Terrorismusprävention hätten entsprechende berufliche Vorerfahrungen, etwa durch eine Tätigkeit bei Nachrichtendiensten oder beim FBI und betreuen terroristische Themen, Namen, Memes, Ikonographien und mehr. Aufgrund neuer Bedrohungslagen, die vor allem in kleinen Splittergruppen entstehen, ergibt es für die Online-Plattform Sinn, eine auf Terrorismus spezialisierte Einheit zu beschäftigen, die solche sich dynamisch entwickelnden Lagen und Gruppen besonders beobachtet. Diese Mitarbeiter entwickeln auch die grundlegende Schulung in Terrorismusprävention für sämtliche Moderatoren des Dienstes.

Der Dienst unterscheidet Sicherheitssysteme und Meldewege. Zu den Sicherheitssystemen zählen:

- Automatisierte und manuelle Bildüberprüfung
- Automatisierte Chat-Filter und Regeln
- Spezielle Chat-Einschränkungen für Nutzer unter 13 Jahre
- Sicherheitskontrollcenter für Nutzer und deren Eltern
- Nutzermeldungen aus der Community.

Zu den Meldewegen zählen:

1. Monitoring (automatisierte Verfahren)
2. Nutzermeldungen
3. Trusted flaggers

Gemeldete Inhalte werden, unabhängig vom Meldeweg, von den Moderatoren manuell überprüft. Der Großteil des Moderationsaufkommens rührt aus den Meldungen der Nutzer her.

Der Dienst veröffentlicht einen Transparenzreport zur TCO-VO („TCO Annual Report") auf seiner Webseite.[88] Im Berichtszeitraum hat der Dienst von europäischen Behörden keine Entfernungsanordnungen gemäß der TCO-Verordnung erhalten. Besondere Anpassungen an der eigenen Infrastruktur wurden durch das Inkrafttreten der TCO-VO

[87] Vgl. https://roblox.fandom.com/wiki/Moderation_system, abgerufen am 20.09.23

[88] Vgl. https://corp.roblox.com/safety-civility-resources/?section=Tools&article=tco-annual-report

nicht nötig, zu jenem Zeitpunkt existierten bereits die aktuellen internen Prozesse zur Terrorismusprävention.

Abschließen mahnt der Dienst im Gespräch, das für diese Studie geführt wurde, dass Sicherheit eine Reise ist, aber kein Ziel, das erreicht bleiben kann. Insbesondere im Bereich Terrorismus und gewalttätiger Extremismus wird die Lage von neuen Ereignissen und Gruppen, bzw. Umfirmierungen und neuen Erscheinungsbildern geprägt, sodass Sicherheit in diesem Bereich ein bewegliches Ziel bleibt.

4.2 Kleine Online-Plattformen und -Anbieter

Nachfolgend wird dargestellt, wie

a) kleine Online-Plattformen mit weniger als zwei Millionen registrierten Nutzer im Inland, die nicht dem NetzDG unterliegen sowie

b) kleine Hostingdienste, die nicht als Online-Plattformen betrachtet werden, da die Verbreitung von Nutzerinhalten nur eine unbedeutende und untrennbar mit einem anderen Dienst verbundene Nebenfunktion darstellt, wie exemplarisch die Kommentarbereiche von Online-Zeitungen[89]

ihr Content-Management betreiben.

Die Darstellungen in diesem Abschnitt beruhen im Wesentlichen auf den Aussagen von Anbietern kleinerer Online-Plattformen, mit denen für diese Studie Hintergrundgespräche auf Geschäftsführungs- bzw. auch auf Moderatoren-Ebene geführt wurden.

4.2.1 Anbieter aus dem Bereich Games-Nachrichten

Der Online-Dienst im Bereich Games-Nachrichten wird von einem Verlag im Bereich Entertainment betrieben, der unter verschiedenen Markenauftritten redaktionelle Informationen und markenübergreifende Community-Angebote (textbasierte Angebote wie Foren, Kommentarfunktionen etc.) betreibt und seine Inhalte auch auf eigenen Webseiten und über Social-Media ausspielt. Seine Angebote werden durch Abonnements und durch Werbung finanziert. Ein Teil der Inhalte ist frei verfügbar. Als vorwiegend redaktionelles Angebote unterfällt der Dienst nicht dem DSA. Das Games-Nachrichtenportal wird um eine große Community-Plattform mit vielen verschiedenen Foren zum Thema Games ergänzt.

Eigene KI-Lösung auf Basis von ChatGPT

Zur Unterstützung der diensteeigenen Moderatoren wird seit 2023 eine durch das interne Entwicklungsteam des Anbieters entwickelte KI-Lösung getestet, die auf dem Large Language Model ChatGPT von OpenAI in seiner aktuellen Version GPT-4 aufsetzt[90]. Hierbei lautet der Prompt im Kern, dass sich das Modell wie ein Foren-Moderator

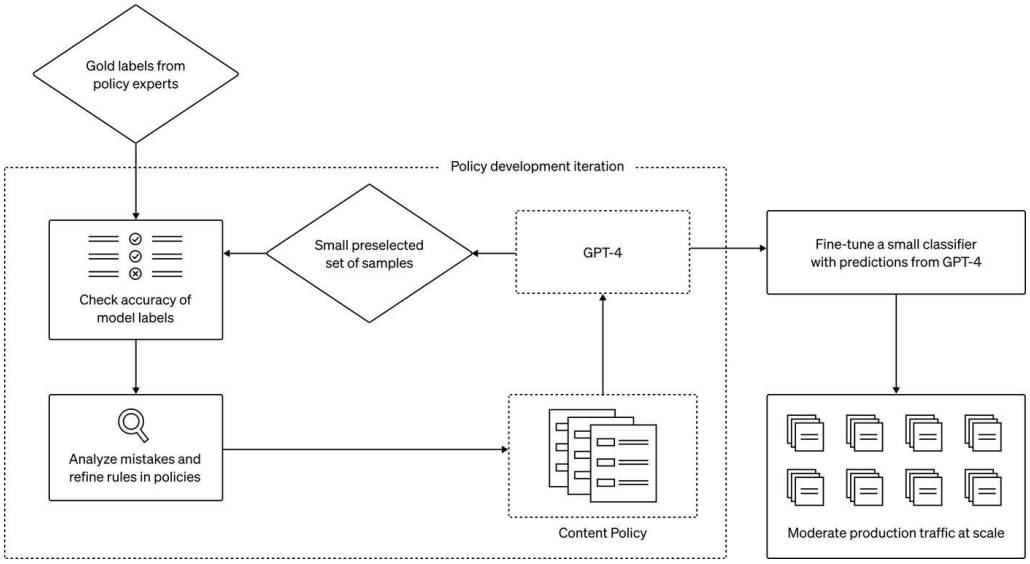
[89] Vgl. Erwägungsgrund 13 DSA

[90] Dies hängt damit zusammen, dass auch bei einem allgemeinen Large Language Model wie ChatGPT (manuelle) Moderationsentscheidungen eine herausgehobene Bedeutung beim Training des KI-Modells haben. So wurden zum Training der KI von ChatGPT mitunter dieselben Dienstleister eingesetzt, die auch für sehr große Online-Plattformen Inhalte moderieren.
Vgl. https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai, aufgerufen am 31.08.23

auf Basis gängiger Community-Richtlinien verhalten und die vorgelegten Inhalte entsprechend flaggen bzw. im automatischen Modus auch moderieren soll. Eine spezifische Anpassung an die eigenen Angebote wurde nicht vorgenommen.

Auch wenn ChatGPT nicht auf den Daten des Anbieters trainiert werden kann, liefert es bei allgemeinen Aufgaben der Textmoderation bereits sehr zufriedenstellende Ergebnisse. OpenAI selbst nutzt ChatGPT auch für die Moderation der Inhalte, die durch ChatGPT generiert und verbreitet werden, sowie zur Fortentwicklung ihrer eigenen Inhalterichtlinien[91].

Abb. 8 Prozess der Nutzung von GPT-4 für die Moderation von Inhalten und zur Entwicklung von Moderationsrichtlinien



Quelle: OpenAI, online unter: https://openai.com/blog/using-gpt-4-for-content-moderation, abgerufen am 07.09.23

Tab. 18 Kennziffern zur Inhaltemoderation der Games-Community

Kenngröße	Angabe	Zeitpunkt	Anmerkungen
Unique User pro Monat	6,9 Mio.	August-Oktober 2022	AGOF, Unternehmensangabe
Visits pro Monat	32,2 Mio.*	April 2023	IVW
Durchschnittliche Nutzerbeiträge pro Monat	1,03 Mio.	Juni 2023	Expertengespräch
Angestellte Moderatoren des Dienstes	4,5 (inkl. freie Mitarbeiter)	Juni 2023	Expertengespräch

** nur Hauptangebot, weitere Angebote des Anbieters im Bereich Gaming nicht IVW-gelistet*
Quelle: Goldmedia Analyse 2023

Die Moderatoren arbeiten täglich während üblicher Bürozeiten, nachts moderiert die KI-Lösung maschinell.

[91] Vgl. https://openai.com/blog/using-gpt-4-for-content-moderation, abgerufen am 31.08.23

Zudem wird sehr stark das Instrument der Gemeinschaftsmoderation genutzt, um die angestellten Moderatoren zu entlasten. Hierbei handelt es sich um etwa 100 aktive Nutzer des Dienstes, die vom Dienst zu Nutzer-Moderatoren ernannt wurden. Auswahlkriterien des Dienstes für User-Moderatoren ist eine Quote von mindestens 30 freiwillig gemeldeten Inhalten pro Monat mit einer Zuverlässigkeit der korrekten Meldung von min. 85 Prozent. Die Nutzer-Moderatoren gliedern sich in drei Hierarchiestufen, von denen die höchste auch über Entfernungsrechte verfügt. Die Qualität der Nutzermoderation werden vom festangestellten Moderationsteam überwacht.

4.2.2 Anbieter aus dem Bereich Q&A-Plattform

Der Dienst ist eine überwiegend textbasierte Frage-Antwort-Plattform, die sich auf keine besonderen Themenbereiche beschränkt, sondern Wissen, Erfahrung und Meinungen zu einem großen Themenbereich bietet und mit einer breiten Meinungsvielfalt abbildet. Der Dienst wird durch Werbung und durch „Business Partner“[92] finanziert. Als Online-Plattform unterfällt der Dienst dem DSA.

Eigene KI-Lösung

Der Dienst verwendet einen eigenen Algorithmus zur proaktiven Moderation, der vom inhäusigen Data-Team selbst entwickelt wurde. Die Entscheidung zur Entwicklung einer eigenen Lösung fiel, da die letzten Tests (ca. 2019) marktverfügbarer Lösungen auf Basis generischer Trainingssets unzufriedenstellend verliefen. Der Algorithmus ist auf den historischen Moderationsentscheidungen der Online-Plattform geschult worden. Besonderes Augenmerk lag hierbei auf dem Ausfiltern von hassschürenden Elementen (Hatespeech, Gewaltaufrufe), der auch verlässlich terroristische Inhalte erkennen sollte. Der Algorithmus vergibt einen Score für die Löschwahrscheinlichkeit zwischen 0 und 1. Ab einem Score von 0,8 wird ein Inhalt manuell von Moderatoren überprüft.

Rolle der Usermoderatoren

Zur Unterstützung der diensteeigenen Moderatoren wird insbesondere auf das Instrument der Gemeinschaftsmoderation durch die Nutzer des Dienstes eingesetzt. Hierfür werden durch den Dienst bestimmte Benutzer durch das Community-Management zu Usermoderatoren ernannt, wenn sich bestimmte Kriterien erfüllen:

- Eigeninitiative Meldung von mindestens 30 Inhalten pro Monat
- Trefferquote für problematische Inhalte bei diesen eigeninitiativen Meldungen: mindestens 85 Prozent

Diese Usermoderatoren haben wiederum eine eigene Hierarchie, die sich durch unterschiedliche administrative Rechte auszeichnen:

- „Junior-Moderator“
- „Light-Moderator“
- „User-Moderator“

In der höchsten Hierarchiestufe verfügen die Usermoderatoren auch über Löschrechte und können Inhalte von der Plattform entfernen. Die Tätigkeiten der Usermoderatoren

[92] Mit Business Partnern kann über die Plattform direkt kommuniziert werden, es können ihnen direkt Fragen gestellt werden, vergleichbar zu anderen Firmenpräsenzen in sozialen Medien.

und die Einhaltung der Gemeinschaftsrichtlinie wird vom Community-Management des Dienstes überwacht. Das Community-Management ist die den Moderatoren grundsätzlich übergeordnete Ebene, die auch die Moderatorenschulungen durchführt. Das Community-Management bearbeitet auch mutmaßliche Falschmeldungen. Derzeit liegt das Aufkommen etwa bei 100 von solchen mutmaßlichen Falschmeldungen pro Tag.

Tab. 19 Kennziffern zur Inhaltemoderation der Q&A-Plattform

Kenngröße	Angabe	Zeitpunkt	Anmerkungen
Unique User pro Monat	22,3 Mio.	August 2023	Unternehmensangabe
Aktive Nutzer pro Monat	1,85 Mio.	August 2023	Unternehmensangabe
Durchschnittliche Nutzerbeiträge pro Monat	1,98 Mio.	Juli 2023	Expertengespräch
Durchschnittlich entfernte Nutzerbeiträge	2,8 %	Juli 2023	Expertengespräch
Durch Nutzer beanstandete Moderationsentscheidungen	5,6 %	Juli 2023	Expertengespräch
Mitarbeiter gesamt	> 50	Juli 2023	Expertengespräch
Angestellte Moderatoren des Dienstes	13 (inkl. Teilzeitmitarbeiter)	Juli 2023	Expertengespräch
Zeitgleich aktive Moderatoren	2-4, aufkommensabhängig	Juli 2023	Expertengespräch
Ehrenamtliche Nutzer-Moderatoren	72	April 2017	Pressebericht
Angestellte im Community-Management des Dienstes	9	Juli 2023	Expertengespräch

Quelle: Goldmedia Analyse 2023

Die Moderatoren arbeiten täglich von 8 Uhr bis 24 Uhr. In der Nacht moderierten die eigene KI-Lösung maschinell und die Usermoderatoren manuell. Nach Aussagen des Anbieters funktioniert das Zusammenspiel zwischen technischer Moderation durch die KI-Lösung und Usermoderatoren zuverlässig: Auch in der Nacht bleiben gemeldete Inhalte durchschnittlich maximal nur 3 bis 6 Minuten online, bevor sie durch einen Moderator gelöscht werden.

4.2.3 Anbieter aus dem Bereich Online-Nachrichten

Der Anbieter ist ein größeres Tageszeitungs-Medienhaus, dass im Kerngeschäft Nachrichten über Print-Ausgaben und Online-Portalen/Apps verbreitet und zahlreichen zusätzliche Social-Media-Kanäle nutzt. Seine Angebote werden durch Abonnements, Print-Einzelverkauf und durch Werbung finanziert. Ein Teil der Inhalte sind online frei verfügbar. Als Anbieter, der vorwiegend Nachrichten publiziert, handelt es sich nicht um eine Online-Plattform im Sinne des DSA.

Das hier dargestellte Online-Diskussionsforum ist konkret an ein Online-Tageszeitungsangebot gebunden und lässt ausschließlich Textinhalte zu. Das Diskussionsforum kann ausschließlich von Abonnenten der Zeitungsmarke genutzt werden. Die weiteren Aus-

führungen beziehen sich ausschließlich auf diesen Markenauftritt und nicht auf die anderen Nachrichtenangebote bzw. Unternehmungen des Medienhauses. Für den Anbieter ist das Diskussionsangebot ein wichtiges Instrument zur Abonnentengewinnung und zur Kundenbindung. Die Nutzerforschung des Verlages zeigt, dass Diskussionsteilnehmer eine engere Bindung zum Produkt aufbauen und das Angebot länger abonnieren als der Durchschnitt. Dementsprechend wird das Angebot auch weiterhin aktiv weiterentwickelt und ausgebaut.

Externer Dienstleister für maschinelle Moderation

Als Moderationstool kommt die Lösung Engagently des deutschen Dienstleisters Ferret zum Einsatz. Das Moderationstool wird nach den Eigenangaben des Dienstleisters von vielen öffentlich-rechtlichen und kommerziellen deutschsprachigen Medienanbietern verwendet, um Diskussionen in sozialen Netzwerken zu moderieren. Das Tool arbeitet auf Basis einer vordefinierten Liste mit problematischen Schlüsselbegriffen. Die Kosten für den Einsatz des Tools liegen für den Anbieter bei rund 50.000 Euro pro Jahr.

Manuelle Moderation

Es gibt ein internes Moderationsteam des Anbieters, was primär die manuelle Moderation der Inhalte übernimmt. Darüber hinaus existiert ein übergeordneter Kundenservice, der neben anderen Aufgaben auch als Eskalationskontakt bei Beschwerden über Moderationsentscheidungen dient. Zudem existiert ein Social-Media-Team, welche Inhalte für soziale Netzwerke aufbereitet und dort die Diskussionsverläufe mit den Tools der jeweiligen Plattformen moderiert. Die Tätigkeiten des Social-Media-Teams werden jedoch im Folgenden nicht weiter betrachtet.

Die manuelle Moderation wird inhäusig von 7 Moderatorenteam an jedem Tag des Jahres vorgenommen. Im Durchschnitt arbeiten 2,8 Moderatoren parallel, wobei die Zahl stark von der aktuellen Nachrichtenlage abhängig ist: Aus redaktionellen Gründen werden die Kommentare unter manchen Meldungen ausschließlich manuell moderiert, was den manuellen Moderationsaufwand stark erhöht und maximal für 2-3 Meldungen pro Tag leistbar ist.

Tab. 20 Kennziffern zur Inhaltemoderation der Community einer Online-Zeitung

Kenngröße	Angabe	Zeitpunkt	Anmerkungen
Aktive Nutzer pro Monat	29.000	August 2023	Expertengespräch
Durchschnittliche Nutzer-beiträge pro Monat	617.000	August 2023	Expertengespräch
Durchschnittlich entfernte Nutzerbeiträge	9,9 %	August 2023	Expertengespräch
Angestellte Moderatoren des Angebotes	7	August 2023	Expertengespräch
Zeitgleich aktive Moderatoren	2,8	August 2023	Expertengespräch
Angestellte im Kunden-dienst des Angebotes	7	August 2023	Expertengespräch

Quelle: Goldmedia Analyse 2023

Die Moderatoren arbeiten täglich an jedem Tag des Jahres. Nachts moderiert die maschinelle Moderationslösung. Auch nachts gehen daher Nutzerkommentare direkt online, solange sie von der maschinellen Vorprüfung nicht beanstandet wurden.

Das Kommentaraufkommen ist stetig zunehmend, auch wenn lediglich Abonnenten des Angebotes Zugriff auf diese Funktion haben. Aufgrund des moderationsintensiven Umfeldes geht der Anbieter nicht davon aus, in Zukunft durch den vermehrten Einsatz maschineller Lösungen signifikant den Gesamtaufwand für die Inhaltemoderation reduzieren zu können. Hintergrund hier ist, dass viele Beiträge mit Andeutungen und Bezügen arbeiten, welche die Transferleistung der Leser mit einbeziehen. Diese Bedeutung wird von KI-Tools nicht erkannt.

4.2.4 KI-Lösung Zöe von Zeit Online

2016 wurde von Zeit Online hausintern eine experimentelle künstliche Intelligenz entwickelt, um zu evaluieren, inwiefern das Moderatorenteam von Zeit Online unterstützt werden kann. In einer Entwicklungszeit von sieben Tagen wurde ein experimenteller Prototyp für einen Kommentar-Bot mit Spamfilter-Funktion durch den „Bordmathematiker" von Zeit Online erstellt, der lediglich 250 Zeilen Code umfasste und bereits eine Übereinstimmungsquote von 75 Prozent mit den menschlichen Moderatoren aufwies.[93]

Die Entwicklung erfolgte in der Programmiersprache Python auf Basis frei zugänglicher Softwarebibliotheken. Es wurde die Python-Bibliothek Keras verwendet, die für neuronale Netzwerke besonders geeignet ist. Die KI basiert auf der Softwarebibliothek Tensorflow des Google-Brain-Teams, welche unter Open-Source-Lizenz veröffentlicht ist.

Die Herausforderung bei der Entwicklung war nach Angaben des Entwicklers nicht die Programmierung des neuronalen Netzwerks, sondern dessen Design: „Was ist sinnvoll für die Textklassifizierung? Wie viele Ebenen staple ich übereinander? Wie verbinde ich die Neuronen? Welche Typen von Layern verwende ich am besten?"[94]

[93] Dieses Vorhaben wurde in einem Zeit-Online-Artikel ausführlich dokumentiert, Vgl. Zeit Online „Mein Bot und ich", https://www.zeit.de/digital/2016-09/kuenstliche-intelligenz-kommentar-bot-zeit/komplettansicht, abgerufen am 20.09.23

[94] Vgl. ebd.

Im Jahr 2018, als das Kommentaraufkommen von Zeit Online rd. 350.000 Beiträgen pro Monat betrug, wurde das 2016 zunächst experimentell entwickelte Tool unter dem Namen „Zöe" in den Wirkbetrieb übernommen und unterstützt seitdem das Moderatorenteam.[95]

4.2.5 Fazit der Analyse Inhaltemoderation bei kleineren Plattformen

In der nachfolgenden Tabelle wird ein Personalindikator für die manuelle Moderation auf Basis der drei dargestellten Fallbeispiele errechnet.

Tab. 21 Personalaufwand zur manuellen Moderation kleinerer Anbieter

	Reichweite pro Monat (Unique User)	Beiträge pro Monat	Moderatoren[96] (in Vollzeit-äquivalenten)	Beiträge pro Moderator pro Monat
Anbieter Gaming	6,9 Mio.*	1,03 Mio.	3,38	305.000
Anbieter Q&A	1,85 Mio. **	1,98 Mio.	9,75	203.000
Anbieter Nachrichten/Politik	29.000 **	617.000	5,25	118.000

** Gesamtangebot, ** Aktive Nutzer des Forums*
Quelle: Goldmedia 2023

Im Ergebnis zeigen sich die Aufwandunterschiede zwischen den Plattformen, die sich sowohl aus der Moderationsintensität des Themenumfeldes als auch aus der gewählten Moderationsstrategie ergeben.

Der Anbieter im Bereich Gaming operiert in einem eher moderationsintensiven Themenumfeld, allerdings mit einem Nischenangebot mit einer starken Community-Komponente, indem im Vergleich zu großen Online-Plattformen vergleichsweise starke „selbstregulative" Kräfte vorhanden sind. Auf einen Vollzeit-Moderator entfallen rechnerisch rund 305.000 veröffentlichte Beiträge pro Monat.

Der Anbieter im Bereich Q&A operiert in einem nicht übermäßig moderationsintensiven Umfeld, hat aber nach außen wenige wahrnehmbare Community-Elemente, da die Plattform eher auf darauf ausgelegt ist, ein niedrigschwelliges Angebot für neue aktive Nutzer ohne spezifische Interessen zu machen, als eine „eingeschworene" Nutzergemeinschaft zu bedienen. Die Plattform stellt den Communityaspekt nicht als besonderen Anreiz zur Nutzung der Plattform heraus und kommuniziert auch nicht aktiv ihre Usermoderatoren. Auf einen Vollzeit-Moderatoren entfallen rechnerisch rund 203.000 veröffentlichte Beiträge pro Monat.

Der Anbieter im Bereich Online-Nachrichten operiert in einem moderationsintensiven Umfeld. Zwar operiert die Plattform auch mit einem Gemeinschaftsgedanken, da nur die Gemeinschaft der Abonnenten Zugang zum Angebot haben. Allerdings wird die Community ausschließlich hierarchisch von oben gesteuert. Der Anbieter bestimmt etwa, bei

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

[96] Die Angaben aus den Gesprächen spiegeln den Headcount wider, der sich sowohl aus festen Mitarbeitern in Vollzeit und zusätzlichen Mitarbeitern zusammensetzt. Im Folgenden wird pauschal angenommen, dass sich die Moderationsteams aus 50 Prozent Vollzeitkräften und 50 Prozent Teilzeitkräften mit einer halben Vollzeitstelle zusammensetzen.

welchen Nachrichten und Meldungen die Kommentarfunktion abgeschaltet oder eingeschränkt wird. So werden Kommentare und einigen Meldungen ausschließlich von Hand moderiert, was den manuellen Moderationsaufwand erheblich steigert. Auf einen Vollzeit-Moderator entfallen rechnerisch rund 118.000 veröffentlichte Beiträge pro Monat.

Der Aufwand für technische Moderationssysteme ist im Vergleich zum Personalaufwand nachrangig. Im Falle des kleinen Anbieters, der auf eine technische Lösung eines Dienstleisters setzt, beträgt der Personalaufwand etwa 80 Prozent des Gesamtaufwands für die Inhaltemoderation.

4.3 Terroristische Inhalte auf Online-Plattformen

Welchen Stellenwert terroristische Inhalte im Bereich der Inhaltemoderation haben, kann seit 2022 nicht nur der allgemeinen Transparenzberichterstattung der sehr großen Plattformen entnommen werden, sondern auch den spezifischen Transparenzberichten zu terroristischen Inhalten, die in Folge der Transparenzpflichten nach Art. 7 TCO-VO seit 2022 von den Hostingdiensten veröffentlicht werden. In der folgenden Tabelle werden die Selbstauskünfte zu terroristischen Inhalten ausgewählter sehr großer Online-Plattformen wiedergegeben

Tab. 22 Auskünfte zu terroristischen Inhalten in der EU auf Basis der TCO-Transparenzberichte, 2. Halbjahr 2022

Artikel in TCO-VO		Face-book	Instagram	YouTube	Twitter/ X	TikTok
	Zeitraum	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
	Entfernungen durch eigene Inhaltemod.	4,1 Mio.[97]	1,5 Mio.	1,4 Mio.[98]	k. A.	53.385
Art. 14(5)	Proaktive Meldung an Strafverfolgung	k. A.	k. A.	k. A.	3	k. A.
Art. 7(3)(d)	Beschwerden	672.000	74.800	15.295	7.270	11.816
Art. 7(3)(g)	Berechtigte Beschwerden	90.800	11.900k	1.783	401	6.153
Art. 7(3)(g)	Anteil berechtigter Widersprüche an Entfernungen	2,2 %	0,8 %	0,1 %	k. A.	11,5 %
	Nicht-berechtigte Entfernungsanord.	25	125	k. A.	k. A.	k. A.
Art. 7(3)(c)	Entfernungsanordnungen durch qualifizierte Behörden	0	0	0	0	0
Art. 7(3)(e)	Behördliche oder gerichtliche Überprüfungsverfahren	0	0	0	0	0

Quelle: Goldmedia Analyse nach TCO-Transparenzberichten der Online-Plattformen 2023

Die Zahl der Entfernungen ist bei den betrachteten sehr großen Online-Plattformen bei Facebook am größten (4,1 Mio.), bei TikTok am niedrigsten (53 Tsd.). Der Grund für die

[97] Definition Facebook: Zahl entspricht der Entfernungen, die gegen die Facebook-Gemeinschaftsrichtlinien über gefährliche Organisationen und Einzelpersonen, Gewalt und Aufwiegelung sowie Koordinierung von Schaden und Förderung von Straftaten verstoßen haben

[98] Definition YouTube: Anzahl der Objekte mit terroristischem Inhalt, die aufgrund von Verstößen gegen die Gemeinschaftsrichtlinien oder aufgrund rechtlicher Anordnungen entfernt wurden.

relativ großen Unterschiede zwischen den Plattformen dürften vor allem aus einer unterschiedlichen Definition der ausgewiesenen Inhalte herrühren, so ist die Terrorismusabgrenzung, die Facebook innerhalb des TCO-VO-Transparenzberichtes zugrunde legt, relativ weit.

Die Anzahl der berechtigten Beschwerden bei Sperrungen aufgrund von terroristischen Inhalten schwankt in der betrachteten Stichprobe zwischen 0,1 Prozent (YouTube) und 11,5 Prozent (TikTok), wobei eine Quote berechtigter Beschwerden im zweistelligen Prozentbereich ungewöhnlich hoch ist.

Keine der betrachteten sehr großen Online-Plattformen hat im zweiten Halbjahr 2022 eine Entfernungsanordnung einer qualifizierten Behörde erhalten. In den Gesprächen, die im Rahmen der Studienerstellung mit sehr großen Online-Plattformen geführt wurden, wurde die Moderation von terroristischen Inhalten im Allgemeinen und die Umsetzung der TCO-VO im Speziellen auch nicht als besondere Herausforderungen für die Dienste eingeschätzt. Die grundlegende Infrastruktur, die für ein TCO-konformes Verhalten benötigt wird, ist auch ohne regulatorisches Erfordernis grundlegender Bestandteil der allgemeinen Infrastruktur der Inhaltemoderation.

Auch kleine Dienste teilten diese Einschätzung durchgehend. Die Prozesse, die für eine TCO-VO-konforme Moderation benötigt werden, sind bereits Bestandteil der bestehenden Moderationsprozesse. Hinzu kommt der Fakt, dass terroristische Inhalte bei kleinen Diensten in der Regel in Nischenphänomen darstellen, da kleinere Dienste, auch aufgrund ihrer begrenzten Reichweite, nicht zu den bevorzugten Verbreitungswegen von terroristischen Inhalten zählen. Entfernungen auf Grund von terroristischen Inhalten wurden als Einzelfälle, oder zumindest als sehr selten (weniger als 1 Prozent der Inhalte) vorkommend geschildert.

4.4 Gesamtfazit

Die wesentlichen Erkenntnisse zum Status-quo der Inhaltemoderation insbes. in Deutschland lassen sich wie folgt zusammenfassen:

Gemeinschaftsrichtlinien und Transparenz der Moderationsentscheidungen

- Zentraler Hebel für eine professionelle Inhaltemoderation ist die Qualität der Gemeinschaftsrichtlinien. Sie müssen konkret, operationabel und widerspruchsfrei sein. Die Effektivität der anderen Ebenen wird hierdurch bedingt.
- Die Kommunikation der Richtlinien sollte über eine leicht (auch für Nicht-Nutzer) zugängliche Meldeplattform erfolgen.

Automatisierte Verfahren der Inhaltemoderation

- Automatisierte Moderationsverfahren werden von sämtlichen Anbietern zur Unterstützung und Entscheidungsvorbereitung eingesetzt. Hierzu zählen
  - Wortfilter (Diese sind bei vielen externen Moderationssystemen inkludiert. Die Listen müssen jedoch plattformspezifisch erweitert und gepflegt werden.)
  - Hash-Abgleiche für Bild/Ton (Hashwert-Plattformen stehen i. d. R. kostenfrei zur Verfügung. Zur Nutzung müssen jedoch Schnittstellen programmiert werden.)

  - KI-Sentimentanalysen (Diese sind seit ca. 2017 für Textinhalte bei vielen externen Dienstleistern zu geringen Kosten verfügbar.)
  - Automatisierte Full-Service-Lösungen sind aktuell bereits ab 4.000 €/Monat verfügbar.
- Automatisierte Lösungen arbeiten insbesondere im Bereich der Textanalyse sehr überzeugend.
- Lösungen zur Analyse von anderen Mediengattungen wie Video oder Audioinhalten sind in ihrer Marktverfügbarkeit im Vergleich deutlich eingeschränkter und die Ergebnisse weniger reliabel. Insbesondere Live-Content (Streaming) stellt automatisierte Moderationsverfahren noch vor hohe moderative Hürden.
- Bei moderationsintensiven Inhalten (bestimmte Politikfelder, Gamestitel) unterstützen automatisierte Systeme kaum.
- Prädiktive, kontext-sensitive KI-gestützte Verfahren befinden sich aktuell in der Entwicklung, werden jedoch noch nicht in großem Stil eingesetzt.

Manuelle Moderation

- Die manuelle Moderation durch menschliche Entscheidungsträger bleiben das Fundament jeglicher Inhaltemoderation.
- Die manuelle Moderation durch hinreichend sprachkundiges Personal ist für das Verständnis des kulturellen Kontexts von Äußerungen von großer Bedeutung.
- Der manuellen Moderation, durch angestellte Moderatoren des Dienstes, aber auch durch Inhalteersteller (Creators) und Nutzer, kommt bei Video- oder Live-Inhalten daher eine noch stärkere Bedeutung zu als bei der Moderation von textbasierten Inhalten.
- Das Ergebnis der manuellen Moderation ist abhängig von der Qualität der Gemeinschaftsrichtlinien (s. o.), der Moderationsprozesse und der personellen Ausstattung.
- Richtwert: Je nach Komplexität muss für 100 Tsd. bis 300 Tsd. Nutzerbeiträge pro Monat mit einer Vollzeitstelle für die manuelle Moderation kalkuliert werden.
- Die Personalkosten für manuelle Moderation sind in Relation zum Kommentaraufkommen selbst bei den kleinsten Anbietern personell abbildbar.
- Der Moderationsaufwand steigt mittelfristig aufgrund der weiter steigendenden Nutzung der Dienste. Online-Plattformbetreiber vergrößern aktuell ihre Moderationsteams oder haben diese kürzlich erweitert. Kein externer Dienstleister plant, seine Teams zu verringern oder rechnet damit, dass sich aufgrund verbesserter automatisierter Verfahren der Personalbedarf verringern könnte. Hier ist eher das Gegenteil der Fall: Kontext-sensitive KI-basierte Systeme werden das manuelle Moderationsaufkommen weiter vergrößern.

Terroristische Inhalte und TCO-VO

- Notice-and-Takedownverfahren insb. durch Trusted Flagger sind branchenweit etabliert und als Selbstregulierungsinstrument sehr erfolgreich. Die Verfahren funktionieren grenzübergreifend/international schnell und effektiv.
- TCO-Entfernungsanordnungen bilden derzeit noch die Ausnahme im Rahmen der Hinweise, die Hostingdienste bzw. Online-Plattformen von Behörden erhalten.

- TCO-Vorgaben machen in den allermeisten Fällen keinen zusätzlichen manuellen/automatisierten Monitoring- und Prüfaufwand erforderlich.
- Die zeitliche Frist zur Reaktion auf eine Entfernungsanordnung innerhalb einer Stunde gem. Art. 3 Abs. 3 TCO-VO kann durch automatisierte Löschung und spätere Nachprüfung eingehalten werden.
- Insbesondere Anbieter, deren Moderationsprozess in zentralen Punkten Schwächen aufweist, werden auch Probleme bei der Einhaltung der TCO-VO-Vorgaben haben.

5 Mindeststandards der Inhaltemoderation

5.1 Herleitung abstrakter Mindeststandards

Hostingdienste, die terroristischen Inhalten gem. Art. 5 Abs. 4 TCO-VO „ausgesetzt" sind, müssen gem. Art. 5 Abs. 2 TCO-VO „spezifische Maßnahmen" zur Eindämmung dieser Inhalte ergreifen. Aufgabe der BNetzA ist es, die Wirksamkeit und Angemessenheit dieser spezifischen Maßnahmen gem. Art. 5 Abs. 3 TCO-VO zu beurteilen.

Eine Beurteilung der Wirksamkeit implementierter spezifischer Maßnahmen kommt dabei einer Beurteilung der Qualität der gesamten Inhaltemoderation eines Hostingdienstes gleich. Insbesondere Anbieter, deren Moderationsprozesse in zentralen Punkten Schwächen aufweisen, werden auch Probleme bei der Einhaltung der TCO-VO-Vorgaben haben.

Nachfolgend werden auf Basis der bisherigen Analyse der Praxis der Inhaltemoderation in Deutschland mögliche und sachgerechte Maßnahmen zur Inhaltemoderation dargestellt, die ein Hostingdienst ergreifen kann, um der Verbreitung illegaler und insbesondere terroristischer Inhalte über seine Plattform entgegenzuwirken. Hierbei wird in Hinblick auf die Vorgabe der (wirtschaftlichen) Angemessenheit der Maßnahmen für unterschiedlich große Hostingdienste gem. Art. 5 Abs. 3 lit. b TCO-VO auf Basis der vorangegangenen Analyse auf Kosten und Aufwand der Maßnahmen eingegangen.

Alle beschriebenen spezifischen Maßnahmen sind einschlägige Marktpraktiken, die von Hostingdiensteanbietern zur Moderation von Inhalten genutzt werden. Die Maßnahmen werden in die folgenden Bereiche unterteil:

a) Gemeinschaftsrichtlinien
b) Moderationsprozess
c) Manuelle Moderationsverfahren
d) Automatisierte Moderationsverfahren
e) Kooperation mit Dritten

Die Maßnahmen werden tabellarisch in einer Matrix aufbereitet, die differenziert zwischen „Alle Dienste" und „Große Dienste/VLOPs" sowie zwischen „Good Practice der Branche" und „Erweiterte Maßnahmen".

Maßnahmen, die selbst für kleine Hostingdiensteanbieter mit Fokus auf Deutschland allgemein anerkannt sind und praktiziert werden, finden sich jeweils in der Zeile „Alle

Dienste". Hiervon getrennt dargestellt werden größere Dienste, die auf verschiedenen Märkten und in verschiedenen Sprachen operieren. Hierzu zählen etwa VLOPs/VLOSEs sowie weitere Sozialen Netzwerke, die dem NetzDG unterfielen.

Zugleich werden in der Matrix allgemein anerkannte Maßnahmen, die in der Regel von allen relevanten Hostingdiensten praktiziert werden, in die Kategorie „Good Practices der Branche" einsortiert. Darüber hinausgehende Maßnahmen werden als „erweiterte Maßnahmen" bezeichnet, die von einigen Diensten aufgrund struktureller Gründe eingesetzt werden. Zu diesen strukturellen Gründen zählen eine potenziell größere Gefährdung aufgrund der adressierten Zielgruppe (z. B. Minderjährige) oder der thematischen Ausrichtung (z. B. Diskussion politischer Themen) sowie technische Notwendigkeiten aufgrund der vorherrschenden Signal-Art (z. B. Live-Inhalte, Video-Inhalte).

Die allgemein anerkannten „Good Practices der Branche" entsprechen dabei den spezifischen Mindestanforderungen des Art. 5 Abs. 3 TCO-VO und sind für alle Hostingdienste grundsätzlich umsetzbar und sachgerecht und nur mittelbar von dessen Leistungsfähigkeit (Größe, Finanzkraft, Personalstärke etc.) abhängig. Im Falle einer konkreten Gefährdung durch oder vermehrten Betroffenheit von rechtswidrigen Inhalten können jedoch erweiterte Maßnahmen notwendig werden.

Insbesondere für Hostingdienste, die offiziell als „terroristischen Inhalten ausgesetzt" gelten, kann es im Einzelfall notwendig werden, erweiterte Maßnahmen zu ergreifen, wenn die „Good Practices der Branche" bereits implementiert sind und dennoch die Verbreitung rechtswidriger Inhalte, insbesondere terroristische Inhalte, nicht zufriedenstellend eingedämmt werden kann.

Tab. 23 Tabellenmuster „Mögliche spezifische Maßnahmen, nach Größe des Dienstes"

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	▪ Konsens der Branche ▪ allgemein anerkannt und praktiziert ▪ Maßnahmen für alle Dienste realisierbar	▪ über den Branchenstandard hinausgehende Maßnahmen ▪ für einige Dienste ggf. nicht sinnvoll bzw. vom Aufwand her nicht darstellbar
Große Dienste/ VLOPs	▪ Konsens unter VLOPs ▪ für kleinere Dienste oftmals vom Aufwand nicht darstellbar	▪ über den VLOP-Standard hinausgehende Maßnahmen ▪ für einige Dienste oftmals nicht sinnvoll

Quelle: Goldmedia Analyse 2023

Erweiterte Maßnahmen können in Art und Umfang stärker von der individuellen Leistungsfähigkeit des Hostingdiensteanbieters abhängen. Jedoch bezieht sich ein Großteil der erweiterten Maßnahmen auf organisatorische Aspekte, die nicht unbedingt zu substantiell erhöhten finanziellen Aufwänden für die Inhaltemoderation führen müssen.

5.2 Spezifische Maßnahmen zur Erreichung der Mindeststandards der Inhaltemoderation

5.2.1 Gemeinschaftsrichtlinien

Die Gestaltung der Gemeinschaftsrichtlinien stellen einen wesentlichen Grundpfeiler des gesamten Prozesses der Inhaltemoderation dar. Jenseits gesetzlicher Bestimmungen gestaltet der Dienst hierüber, welche Inhalte auf seiner Plattform zulässig sind und wie mit Verstößen gegen die Gemeinschaftsrichtlinien umgegangen wird.

Der Anbieter definiert hiermit seinen eigenen Maßstab, an dem seine Anstrengungen zur Inhaltemoderation gemessen werden können. Eine konkrete, operationalisierbare Definition von etwa unerwünschten Inhalten sowie von Sanktionsmechanismen ist wesentlich für die Anwendung dieser Richtlinien innerhalb der Moderationsprozesse und -verfahren.

Tab. 24 Mögliche spezifische Maßnahmen, nach Größe des Dienstes

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	▪ adäquate, operationalisierbare Gemeinschaftsrichtlinien ▪ transparente Maßstäbe der Inhaltemoderation ▪ Kommunikation der Gemeinschaftsrichtlinien	▪ Präzisierung der Gemeinschaftsrichtlinien
Große Dienste/ VLOP	-	-

Quelle: Goldmedia Analyse 2023

Insofern besteht hierin ein objektiver und einfach zu prüfender Maßstab, mit dem zunächst überprüft werden kann, ob ein Anbieter sich hinreichend mit Aspekten der Inhaltemoderation, insbesondere mit Bezug zu rechtswidrigen Inhalten, auseinandergesetzt hat.

Ebenfalls ist aus der Strukturierung und Nutzerführung des Dienstes zu erkennen, welchen Stellenwert ein Anbieter darauf legt, seine eigenen Maßstäbe an seine Nutzer zu vermitteln. Hierbei steht neben der reinen Auffindbarkeit der Gemeinschaftsrichtlinien auch die aktive Kommunikation der Maßstäbe und die damit verbundene Sensibilisierung der Nutzer im Umfeld des Inhalte-Uploads im Fokus

Eine Überprüfung der Gemeinschaftsrichtlinien eines Dienstes ist auch im Vergleich zu anderen Diensten (Good Practice der Branche) einfach zu realisieren, da die Gemeinschaftsrichtlinien aller Anbieter transparent sind. Schwachstellen in der konkreten Ausgestaltung von Gemeinschaftsrichtlinien (mangelnde Begriffsbestimmungen, mangelnde Operationalisierung/Konkretisierung, nicht adressierte strafrechtliche Bereiche) lassen sich bereits ohne Kenntnis von technischen oder organisatorischen Details von Moderationsprozessen identifizieren.

Sollten Unschärfen oder Mängel bei der Durchsicht von Gemeinschaftsrichtlinien auffallen, kann der Dienst um Änderung im Sinne einer Präzisierung aufgefordert werden. Im Vergleich mit anderen Maßnahmen können Änderungen, die sich lediglich auf der Policy-Ebene eines Dienstes bewegen, sehr zeitnah umgesetzt werden. Hierfür ist i. d. R.

keine Einbindung externe Dienstleister oder der Aufbau zusätzlicher interner Ressourcen notwendig. Alle Hostingdienste sollten in der Lage sein, kurzfristig auf eine solche Änderungsempfehlung reagieren zu können.

5.2.2 Transparenz der Inhaltemoderation

Für die Beurteilung der Effektivität der Moderationspraktiken eines Anbieters kommt internen Statistiken eines Dienstes zur Inhaltemoderation eine entscheidende Bedeutung zu. Aufgrund gesetzlicher Vorgaben sind hiervon wesentliche Angaben veröffentlichungspflichtig. So bieten die Transparenzberichtspflichten nach Artikel 15 DSA und die zuvor bestehenden Transparenzpflichten nach § 2 NetzDG bereits einen guten Überblick über die Praxis der Inhaltemoderation eines Dienstes. Künftig wird zudem die gem. Art. 20 Abs. 3 DSA bestehende Meldepflicht an die DSA-Online-Transparency-Datenbank eine weitere Informationsquelle darstellen, um sich einen aktuellen Überblick über den Umfang der moderierten Beiträge, das Aufkommen rechtswidriger Inhalte und der Detektionsmethoden zu verschaffen.

Allerdings ist ein quantitatives Benchmarking eines Dienstes mit anderen Diensten auf Basis der Transparenzberichte bzw. der Online-Transparency-Datenbank nur eingeschränkt möglich, da sich die Moderations- und die damit verbundenen Dokumentationsprozesse stark unterscheiden können. Hilfreicher ist daher der Längsschnittvergleich der Transparenzberichte/Online-Transparency-Datenbank eines Dienstes über verschiedene Berichtszeiträume. Dies wird jedoch erst mittelfristig möglich sein, da die Transparenzberichtspflichten ab 2024 umfassend (mit Ausnahme von Kleinst- und Kleinunternehmen) für Hostingdienste gelten werden.

Tab. 25 Mögliche spezifische Maßnahmen, nach Größe des Dienstes

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	▪ Transparenzberichte nach Art. 15 DSA (oder vergleichbar)	▪ Um detaillierte Kennziffern erweiterte Transparenzberichte ▪ Engmaschigere (z. B. monatliche) nicht-öffentliche Berichtspflichten
Große Dienste/ VLOP	▪ Transparenzberichte zusätzlich auf den Metriken des NetzDG	-

Quelle: Goldmedia Analyse 2023

Die Effektivität der Moderation von rechtswidrigen Inhalten bemisst sich primär an der Bearbeitungszeit ab Eingang der Meldung und sowie an der Zuverlässigkeit der Erkennung rechtswidriger Inhalte („false negatives"/"false positives").

Die vorgesehene Berichtspflicht nach Art. 15 DSA, nach dem nur die, bis zur Entscheidung benötigte Mediendauer von Vermittlungsdiensteanbietern zu berichten ist, ist zur Beurteilung leider nicht hinreichend aussagekräftig. Der Berichterstattungsmaßstab des § 2 Abs. 2 Nr. 9 NetzDG (innerhalb von 24 Stunden, innerhalb von 48 Stunden, innerhalb einer Woche etc.) wäre für eine eingehende Beurteilung besser geeignet. Sollten Dienste in ihren öffentlichen Transparenzplichten lediglich auf dem gesetzlichen Mindestniveau berichten, wird es daher künftig dennoch im Bedarfsfall notwendig werden, Anbieter um detailliertere Auswertung der Moderationszeiträume zu bitten.

Die Zuverlässigkeit der Erkennung rechtswidriger Inhalte sollte hingegen aus den Angaben von Artikel 15 Absatz 1 lit. d DSA bereits hinreichend erkennbar werden.

Weiterhin wünschenswert kann im Bedarfsfall eine separate Darstellung der Bearbeitungszeiten und der Zuverlässigkeit der Erkennung für die Teilmenge der terroristischen Online-Inhalte an der Gesamtmenge der rechtswidrigen Inhalte, sowie die Definition terroristischer Inhalte, die der Dienst in seiner Moderationspraxis zugrunde legt. Diese Angaben werden in Transparenzberichten zwar häufig auf freiwilliger Basis gemacht, im Bedarfsfall müsste bei einem potenziell terroristischen Inhalten ausgesetztem Anbieter jedoch sichergestellt werden, dass diese Angaben für eine behördliche Überprüfung vorliegen.

Zudem scheint die gesetzlich verankerte jährliche Berichterstattungsplicht für einen potenziell ausgesetzten Anbieter kaum hinreichend zu sein, um akute Problemstellungen, die durch besondere Lagen oder spezifische operative Einschränkungen entstehen, erkennbar werden zu lassen. Im Bedarfsfall könnte mit ausgesetzten Anbietern daher im Rahmen eines Bewährungszeitraums ein Reporting in kürzeren Abständen vereinbart werden, um den Dialog über konkrete spezifische Maßnahmen zielgerichtet führen zu können.

5.2.3 Prozess der Inhaltemoderation

Der Prozess der Inhaltemoderation regelt alle organisatorischen Aspekte der Inhaltemoderation, vom Meldungseingang, über die Zuordnung und Priorisierung von Meldungen, die Verzahnung der technischen Systeme mit manuellen Moderationsverfahren sowie die Erfassung und das Monitoring aller Moderationsvorgänge.

Alle Prozessbestandteile müssen vom Dienst für das Funktionieren seiner internen Betriebsabläufe hinreichend dokumentiert sein. Insbesondere bei größeren Anbietern, bei denen die Inhaltemoderation stark arbeitsteilig erfolgt, müssen Aufgaben- und Rollenprofile sowie technische und personelle Schnittstellen präzise und widerspruchsfrei definiert sein. Eine Prozessevaluation der Inhaltemoderation kann daher grundsätzlich auf Basis dieser betriebsinternen Prozessbeschreibungen dokumentenbasiert erfolgen.

Tab. 26 Mögliche spezifische Maßnahmen, nach Größe des Dienstes

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	▪ Moderationsleitfaden ▪ Prozess für Schulung/Weiterbildung ▪ Guidelines/Hilfestellungen für die Gemeinschaftsmoderation ▪ einfache Meldefunktionen ▪ Widerspruchsverfahren ▪ Ergebnisdokumentation	▪ Prominentere Meldefunktionen für (Nicht-)Nutzer ▪ Kontinuierliche Verbesserung (PDCA) und Richtlinienfortschreibung ▪ Einrichtung eines zentralen Safety-Centers für Nutzer ▪ Stärkere Sensibilisierung sowie Einbezug der Nutzer in die Moderation
Große Dienste/ VLOP	▪ externe Dienstleister (BPO) ▪ Kontinuierliche Verbesserung (PDCA) ▪ Zentrales Safety-Center für Nutzer	▪ DSA-Vorgaben für VLOPs ▪ Risikomanagement und Krisenreaktion (auditiert) ▪ Unabhängige Compliance-Abteilung

Quelle: Goldmedia Analyse 2023

Zentrale Dokumente stellen hierbei die Gesamtheit an Moderationsleitfäden und Hilfestellungen dar, in denen die Gemeinschaftsrichtlinien so operationalisiert sind, dass die

angestellten Moderatoren in die Lage versetzt werden, einheitliche Moderationsentscheidungen zu treffen. Hinzu kommen Schulungs- und Ausbildungsmaterialien, in denen die konkreten Maßstäbe für Moderationsentscheidungen in der Regel anhand vergangener Moderationsentscheidungen illustriert und didaktisch vermittelt werden.[99] Im Falle von Diensten mit Elementen von Gemeinschaftsmoderation kommen hierzu Unterlagen, die speziell für die Nutzer-Moderatoren erstellt sind.

Bei größeren Diensten kommen Vereinbarungen mit externen Dienstleistern hinzu, in denen zusätzlich zu den Moderations- und Schulungsunterlagen auch der Personaleinsatz, Arbeitszeiten, technische Schnittstellenspezifikationen, Response-Zeiten und Service-Levels definiert sein können. Auch Regelungen zur Qualitätsmessung und Kontrolle der getroffenen Moderationsentscheidungen sollten bei großen Plattformen definiert sein. Ohne strukturierte Mechanismen der Qualitätsprüfung wird ein Dienst nicht in der Lage sein, seine Compliance mit seinen eigenen Richtlinien zu evaluieren.

Sämtliche Dienste sollten in der Lage sein, darzustellen, wie der Fortentwicklungsprozess der Moderationsleitfäden strukturiert ist. In Regel gibt es hierfür formalisierte Feedback-Prozesse und Meetings auf Team-Ebene. Aufgrund der Dynamik im Bereich der Inhaltemoderation ist davon auszugehen, dass auf einer granularen Regelungsebene vergleichsweise häufig und kurzfristig Anpassungen erfolgen können. Zugleich sollte es jedoch auch mehrfach im Jahr Weiterbildungsveranstaltungen geben, in denen team- bis standortübergreifend fortgebildet wird. Sollten Moderationsleifäden und Schulungsunterlagen nicht aktuell sein, ist eine der thematischen Dynamik angemessene Inhaltemoderation in kritischen Moderationsumfeldern wir Terrorismus herausfordernd.

Neben der internen Prozessdokumentation ist auch die Schnittstelle zwischen Nutzern des Dienstes und den internen Moderationsprozessen ein zusätzlicher Aspekt, der im Kontext der Prozessanalyse betrachtet werden sollte. Folgende Fragen stehen hier im Fokus:

- Inwiefern ist es Nutzern (und Nicht-Nutzern) generell möglich, Inhalte zu melden?
- Sind die Meldefunktionen nutzerfreundlich gestaltet und intuitiv zu auffindbar?
- Werden die Nutzer hinreichend zu unerwünschten Inhalten informiert und über die Konsequenzen einer Meldung aufgeklärt?
- Gibt es eine zentrale Stelle, an dem sich ein Nutzer zu den Maßstäben der Inhaltemoderation informieren kann und ggf. eigene Moderationseinstellungen vornehmen kann (Sicherheitscenter)?

Sehr große Online-Dienste sollten darüber hinaus über Risikomanagement- und Krisenreaktionsmechanismen unter Einbindung höherer Ebenen verfügen.

Aus der Gesamtschau der Prozessunterlagen ergibt sich i. d. R. eine gute Übersicht über die Mechanismen, die ein Dienst vorsieht, um mit der Dynamik im Bereich Inhaltemoderation umzugehen. Auf dieser Basis können bereits spezifische Empfehlungen zur Verbesserung der Inhaltemoderation ausgesprochen werden, ohne auf die Ausgestaltung der Gemeinschaftsrichtlinien eines Dienstens Bezug nehmen zu müssen.

---

[99] Beispielhaft hierfür sind etwa die Schulungsunterlagen von Facebook zum Umgang mit terroristischen Inhalten, die The Guardian 2017 veröffentlicht hat: https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content, abgerufen am 27.10.23

5.2.4 Manuelle Moderation

Die manuelle Moderation durch Mitarbeiter des Dienstes (Safety- und Security-Team) stellt das Fundament jeglicher Inhaltemoderation dar. Ohne manuelle Moderation bzw. Mitarbeiter eines Dienstes, die sich mit der Moderation der Inhalte befassen, ist eine adäquate Inhaltemoderation grundsätzlich nicht zu gewährleisten. Insofern ist die Existenz einer Abteilung, die innerhalb eines Dienstes für die Inhaltemoderation zuständig ist, zwingend für Online-Plattformen, die Inhalte im Auftrag ihrer Nutzer veröffentlichen.

Hierbei spielt die Größe des Dienstes keine ausschlaggebende Rolle. Die Personalkosten für manuelle Moderation sind selbst bei kleinen Anbietern personell abbildbar: Bei den kleinen Diensten, mit denen im Rahmen der Studienerstellung gesprochen wurde, entfielen durchschnittlich zwischen 100 Tsd. Und 300 Tsd. Nutzerbeiträge pro Monat auf eine Vollzeitstelle in der Inhaltemoderation. Typische Personalstärken der Moderationsabteilung lagen bei 3-12 Vollzeitstellen, die üblicherweise werktäglich zu Bürozeiten und in den nutzungsstarken Abendstunden moderieren.

Als „Cost of doing business" ist der Anteil der Moderatoren am Gesamtpersonalbestand bei kleineren Diensten zwar höher als bei großen Diensten, Moderationsteams mit bis zu einem Dutzend Mitarbeitenden stellen jedoch auch für kleine Dienste keine unbotmäßige Belastung dar. Eine Rund-um-die-Uhr-Moderation ist für kleine Abteilungen mit weniger als 10 Mitarbeitenden in der Inhaltemoderation nicht darstellbar.

Ab einer deutlich zweistelligen Mitarbeiterzahl im Bereich Inhaltemoderation überwiegen die Vorteile des (teilweisen) Outsourcings des Safety- und Security-Teams auf hierfür spezialisierte Dienstleister. Hierbei ist insbesondere der initiale Aufwand relativ hoch, da eine externe Inhaltemoderation auf deutlich umfangreiche Moderationsleitfäden und Prozessbeschreibungen angewiesen ist als bei der Inhaltemoderation innerhalb einer kleinen Moderatorengruppe, die nur an einem einzelnen Standort moderiert.

Tab. 27 Mögliche spezifische Maßnahmen, nach Größe des Dienstes

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	ja, Manuelle Moderation zwingend erforderlich ▪ interne Abteilung	▪ Manuelle Moderation 24/7/365 ▪ Ggf. eigener Cue Terrorismus
Große Dienste/ VLOP	▪ 24/7/365-Moderation ▪ Spezialisierte Cues (eigene Teams für verschiedene Bereiche der Gemeinschaftsrichtlinien)	▪ Verstärkung des Cues für Terrorismus ▪ Aufbau/Verstärkung des internen Analyseteams für terroristische Bedrohungen ▪ Hinzunahme eines anerkannten, externen Moderationsdienstleisters

Quelle: Goldmedia Analyse 2023

Je nach Moderationsstrategie des Anbieters stellen die Gemeinschaftsmoderation durch die Nutzer und die Eigenmoderation durch Inhalte-Ersteller ein zentrales oder ergänzendes Elemente der manuellen Moderation dar. Die Gemeinschaftsmoderation ist ein typisches Element von Online-Foren. Die großen Social-Media-Plattformen ermöglichen hingegen nur eine Eigenmoderation der jeweiligen Accounts durch die Nutzer.

Während einige Dienste sehr stark auf automatisierte Verfahren zur Früherkennung setzen, steht bei anderen Diensten die Moderation von Beschwerden durch Nutzer im Vordergrund. Daher ist der reine Vergleich der personellen Ausstattung zwischen verschiedenen Diensten kaum aussagekräftig. Bei der Beurteilung des Personalschlüssel muss z. B. berücksichtigt werden, wie relevant für die jeweilige Plattform die Pflege der Beziehung zu den beitragerstellenden Nutzern ist (z. B. kostenfreie Nutzung ggü. abonnement-basierte Nutzung).

Sollte ein Dienst terroristischen Inhalten potenziell ausgesetzt sein, sollte zunächst untersucht werden, ob in der Organisation der manuellen Moderation strukturelle Probleme vorliegen:

- Sind die Moderations-Teams hinreichend in die internen Safety- und Security-Prozesses des Dienstes eingebunden?
- Sind die internen Schulungs- und Weiterbildungsprozesse ausreichend formalisiert?
- Lassen sich Moderationsprozesse durch eine stärkere Spezialisierung der Moderatoren auf bestimmte Moderationsgebiete beschleunigen?
- Gibt es fehlende Sachkenntnis in der Erkennung von terroristischen Inhalten?
- Wurden rechtswidrige Inhalte aufgrund fehlenden kulturellen Kontextes nicht erkannt?
- Häufen sich Probleme in Zeiten, in denen nicht manuell moderiert wird und muss daher der Zeitraum, indem manuell moderiert wird, ausgeweitet werden?

Abhängig vom Ergebnis kann es erforderlich werden, die Aufgaben innerhalb der manuellen Moderation neu zu strukturieren, gezielt Sachkenntnis in spezialisierten Teams aufzubauen oder auch die allgemeine Verfügbarkeit der manuellen Moderation zu verbessern. In Abhängigkeit von der Größe und Leistungsfähigkeit des Anbieters kann dies neben der Ausweitung der manuellen Moderation auch den Einbezug von externem Know-how von spezialisierten Dienstleistern bedeuten. Insbesondere externe Moderationsdienstleister verfügen in der Regel über Erfahrung in der Früherkennung und Moderation terroristischer Inhalte. Eine entsprechende Größe des Dienstes vorausgesetzt, kann die Hinzuziehung externer Expertise kurzfristig Abhilfe schaffen, sollte ein potenziell terroristischen Inhalten ausgesetzter Dienst bislang über keine dedizierten Teams zur Früherkennung bzw. Moderation von terroristischen Inhalten verfügen.

5.2.5 Automatisierten Moderationsverfahren

Der Einsatz automatisierter Moderationsverfahren ist Standard im Bereich der Inhaltemoderation. Sämtliche Anbieter nutzen automatisierte Verfahren, um die Inhaltemoderation zu unterstützen, vor allem bei Text-Inhalten kommt automatisierten Verfahren eine wesentliche Bedeutung bei der Früherkennung problematischer Inhalte zu.

All-in-One-Moderationslösungen externer Dienstleister beinhalten auch (vortrainierte) KI-Unterstützung für die Erkennung problematischer Inhalte. Zum Teil stehen KI-Anwendungen auch als Open Source zur Verfügung, sodass Dienste, ihre eigenen Anwendungen damit entwickeln können (vgl. Kap. 8.2).

KI-gestützte Moderationsinstrumente der sehr großen IT-Konzerne sind im Bereich der Textanalyse vergleichsweise günstig zu beziehen und einfach für Dienste in ihre bestehende Moderationsarchitektur zu implementieren: Marktgängige cloudbasierte Moderationsinstrumente verursachen Kosten zwischen rd. 90-180 Euro für die Moderation von 100.000 Textkommentaren bzw. Bildern.

Tab. 28 Mögliche spezifische Maßnahmen, nach Größe des Dienstes

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	ja, Automatisierte Moderation zwingend erforderlich (für Text)	▪ Abgleich mit Hash-Datenbanken (Text, Bilder, Videos) ▪ externer Anbieter mit TCO-Spezialisierung
Große Dienste/ VLOP	▪ Abgleich mit Hash-Datenbanken (Text, Bilder, Videos) ▪ KI-Training auf eigenen Datensätzen	-

Quelle: Goldmedia Analyse 2023

Der Einsatz automatisierter Systeme im Bereich der Inhaltemoderation zählt zum Mindeststandard im Bereich der Textmoderation und kann von allen Anbietern erwartet werden. Für diese Systeme wird weder ein hoher Entwicklungsaufwand benötigt noch entstehen signifikante Betriebskosten. Der Einsatz automatisierter Systeme für Bild- und Video-Inhalte ist bei kleinen Diensten bislang nicht verbreitet. In der Regel sind nutzergenerierte Inhalte bei kleinen Diensten jedoch vornehmlich textbasiert.[100]

Insofern ist ein potenziell terroristischen Inhalten ausgesetzter Dienst dringend dazu angehalten, automatisierte Verfahren der Inhaltemoderation einzusetzen. Hier gilt es, auf Basis laufender Marktanalysen zu prüfen, ob ein Dienst seine bestehende Moderationsarchitektur mit sinnvollen, ggf. neu verfügbaren Lösungen für die Textmoderation optimieren kann.

Die Landschaft externer Moderationsanbieter ist vielfältig und differenziert. Einige Anbieter stellen insbesondere auf ihre Erfahrung in der Früherkennung terroristischer Inhalte ab. Da grundsätzlich der Aufwand für automatisierte Lösungen im Vergleich zum Personalaufwand für manuelle Moderation gering ist, erscheint es zumutbar, dass sich terroristischen Inhalten ausgesetzte Anbieter bei solchen Anbietern gezielt ein Angebot einholen. Sofern keine zusätzlichen Lösungen implementiert werden, müsste dies begründet werden.

Neben Text-Inhalten moderieren bislang nur größere Dienste auch Bild- und Video-Inhalte mit KI-gestützten automatisierten Verfahren, die in der Regel selbst entwickelt und auf den eigenen Daten trainiert sind.[101] Diese dienen insbesondere der Früherkennung von bestimmten rechtswidrigen Inhalten (Missbrauchsdarstellungen, terroristische Propaganda etc.).

[100] Bild- und Video-Inhalte werden vor allem von spezialisierten großen bzw. sehr großen Online-Plattformen verbreitet, die nicht nur die Verbreitung der Inhalte übernehmen, sondern über ihre Plattform auch durch die Bereitstellung von „Creator-Tools" die Erstellung der Inhalte vereinfachen.

[101] Marktverfügbare externe Lösungen, die auf generischen Datensätzen trainiert sind, eignen sich bislang nur bedingt, da der Einsatz noch vergleichsweise teuer und vergleichsweise wenig reliabel ist.

5.2.6 Kooperation mit Dritten (Strafverfolgung und NGOs)

Ein spezifischer Punkt der Transparenz der Inhaltemoderation ist die Beschreibung der Zusammenarbeit eines Hostingdienstes mit externen Stellen und mit Einrichtungen, die die Moderationsleistung einer Plattform unterstützen können.

Dienste sollten zunächst zwingend ihre eigenen Meldewege zu Strafverfolgungsbehörden auf Prozessebene definiert haben, insbesondere für terroristische Inhalte und andere schwerwiegende Rechtsverstöße (z. B. Kindesmissbrauch) sowie aktuelle Gefahrenlagen.

Neben der gem. Art. 9 DSA verpflichtenden Zusammenarbeit mit Strafverfolgungsbehörden besteht die Möglichkeit, eigene Kommunikationskanäle zu vertrauenswürdigen Community-Mitgliedern oder Meldestellen von NGOs bereitzustellen, deren Hinweise priorisiert behandelt werden (eigene Trusted-Flagger-Programme).

Tab. 29 Mögliche spezifische Maßnahmen, nach Größe des Dienstes

Größe	Good Practices der Branche	Erweiterte Maßnahmen
Alle Dienste	▪ Zusammenarbeit mit Strafverfolgungsbehörden	▪ Hash-Datenbanken ▪ Ernennung von Trusted Flagger
Große Dienste/ VLOP	▪ Hash-Datenbanken ▪ Trusted Flagger Programm ▪ Teilnahme an (internationalen) Branchenforen	-

Quelle: Goldmedia Analyse 2023

Darüber hinaus besteht die Möglichkeit, die Inhalte der Nutzer in Hashwerte umzuwandeln und mit verschiedenen Hashwert-Datenbanken abzugleichen (z. B. GIFCT Hash-Sharing Database, TCAP Terrorist Content Analytics Platform, NCMEC „Take It Down" oder die Check-the-Web-Anwendung der EU IRU). Diesen branchenübergreifenden Hash-Datenbanken (vgl. Kap. 8.2) kommt eine besondere Bedeutung zu, da mit diesen der Klassifizierungsaufwand für alle Anbieter deutlich reduziert wird. Einmal als rechtswidriger Inhalt klassifiziert, lassen sich inhaltsgleiche Kopien eines Inhalts auf anderen Plattformen entfernen, ohne dass diese erneut manuell gesichtet werden müssen.

Bislang scheinen vor allen große Dienste an den branchenweiten Hash-Datenbanken teilzunehmen, auch weil die Mitgliedschaft im Global Internet Forum to Counter Terrorism (GIFCT) an einen umfangreichen Aufnahmeprozess geknüpft ist. Sollte jedoch ein Dienst wiederholt damit auffallen, Inhalte zu verbreiten, die bereits in den einschlägigen Hash-Plattformen des GIFCT indiziert sind, sollte eine Mitgliedschaft dringend empfohlen werden.

Zudem gibt es mittlerweile mehrere internationale Foren, die sich zum Thema Content-Moderation rechtswidriger Inhalte auseinandersetzen. Deren Veröffentlichungen und Datenbanken geben Hinweise für die Weiterentwicklung von Moderationsprozessen. Alle EU-weit relevanten Meldestellen und Datenbanken zum Thema Eindämmung rechtswidriger Inhalte auf Hostingdiensten werden im Anhang in den Kapiteln 7 und 8 dargestellt.

5.3 Zusammenfassung und Empfehlung

Ob die Anzahl der eingesetzten Moderatoren sowie die implementierten technischen Systeme zur automatisierten Unterstützung der Inhaltemoderation ausreichend sind, um insbes. terroristische Inhalte dauerhaft einzudämmen, lässt sich aufgrund der großen Unterschiede zwischen den verschiedenen Diensten in Bezug auf die Inhalte, Medienformen und Kommunikationsstile nicht anhand fester Kennziffern beurteilen.

Allerdings liefern Transparenzberichte Hinweise auf anbieterspezifische Spezifika, die für eine vertiefte Diskussion mit dem Dienst über Moderationsprozesse hilfreich sein können. Die Fähigkeit zu einer adäquaten Moderationsleistung eines Hostingdienstes lässt sich in Grundzügen aus den öffentlich verfügbaren Gemeinschaftsrichtlinien und Transparenzberichten erfassen. Auf Basis ergänzender, interner Dokumente des Diensts (Moderationsleitfäden, Prozessbeschreibungen etc.) und der spezifischen Umstände, die zur Klassifikation als „terroristischen Inhalten ausgesetzter Hostingdienst" geführt haben, lassen sich dann Empfehlungen zur weiteren Konkretisierung der selbst auferlegten Gemeinschaftsrichtlinien, der Community-Guidelines und der internen Organisationsabläufe formulieren.

Manche Auffälligkeiten, wie eine sehr geringe Anzahl an Moderatoren im Verhältnis zum monatlichen Nutzeraufkommen, die Nicht-Nutzung zugänglicher Quellen für die Identifikation von terroristischen Inhalten (z. B. GIFCT, vgl. Anhang Kap. 8.2) oder das Fehlen proaktiver, automatisierter bzw. auch KI-gestützter Filtersysteme, können auch direkt adressiert werden. Hier gilt die Erkenntnis dieser Studie, dass eine Implementierung oder Aufstockung manueller und insbes. auch automatisierter Verfahren der Inhaltemoderation grundsätzlich technisch realisierbar und finanziell zumutbar sind.

Inwieweit die vorgelegten Prozessdokumentationen in der moderativen Praxis Anwendung finden, kann hingegen nur durch eine nachfolgende Erfolgskontrolle betroffener Hostingdienste überprüft werden. Für eine nachhaltige Implementierung der zusätzlichen spezifischen Maßnahmen ist daher eine fortlaufende Überprüfung der Moderationsstatistiken des jeweiligen Dienstes erforderlich. Dies erfolgt im besten Fall in Verbindung mit einer freiwilligen Vereinbarung mit dem betroffenen Dienst, welche messbaren Ziele mit den zusätzlichen spezifischen Maßnahmen in welchem Zeitraum erreicht werden sollen.

Für ein Monitoring der Moderationsstatistiken wäre es zielführend, mit den Hostingdiensten, die terroristischen Inhalten ausgesetzt sind, innerhalb eines Bewährungszeitraums zusätzlich zu den jährlich erforderlichen Transparenzberichten Fortschrittsberichte mit kürzeren Zeiträumen (z. B. 3-6 Monate) zu vereinbaren, in denen auch die qualitative Weiterentwicklung der Moderationsmaßnahmen skizziert werden sollte.[102] Zugleich kann auch die Übermittlung neuer Versionen der Moderationsleitfäden und Schulungsunterlagen vereinbart werden, sobald diese in der Moderation eingesetzt werden.

Eine weitere, künftige Möglichkeit, die Fortschritte im Bereich der Inhaltemoderation von Online-Plattformen effektiver zu überwachen, besteht über statistische Auswertungen der DSA Transparency Database. Diese können für die einzelnen Plattformen auf

[102] Sofern ausgesetzte Hostingdienste in die Kategorie der „Kleinst- und Kleinunternehmen" fallen, ist die Erstellung solcher Berichte dringend anzuempfehlen, da sie gemäß DSA keiner gesetzlichen Transparenzauflage unterliegen.

Länderebene ausgewertet werden. Organisatorische Fortschritte können auf dieser Basis jedoch nur deduktiv erfasst werden.

Darüber hinaus kann die BNetzA die Inanspruchnahme externer Beratung zur Content-Moderation empfehlen, wenn Hostingdienste, die terroristischen Inhalten ausgesetzte sind, nach einem angemessenen Übergangszeitraum keine signifikanten Fortschritte bei der Weiterentwicklung ihrer Moderationsleistung nachweisen können.

Insbesondere Full-Service-Anbieter, die sowohl manuelle Moderation als auch eigene Moderationssysteme und Content-Filter anbieten, können Hostingdienste über den sinnvollen Einsatz der unterschiedlichen Moderationsverfahren beraten. Auch die Wirtschaftsprüfungsgesellschaften, die Audit-Aufgaben im Kontext von Art. 37 DSA für VLOPs übernehmen, positionieren sich hier als Berater. Es ist davon auszugehen, dass aufgrund der im Vergleich zum NetzDG deutlichen Vergrößerung des Kreises der Hostingdienste, die nun gem. DSA einen Transparenzbericht abgeben müssen, der Markt für Beratung insbes. auch für kleine und mittlere Unternehmen weiter wachsen wird.

Anhang

6 Am Prozess der Inhaltemoderation beteiligte nationale Behörden

6.1 Rolle des Bundeskriminalamtes

Das **Bundeskriminalamt** (BKA) ist die zentrale Stelle zur Bekämpfung von Kriminalität im Internet. In der Zentralen Meldestelle für strafbare Inhalte im Internet (ZMI BKA) werden etwa dezentrale Meldestrukturen, die in den Bundesländern zur Bekämpfung von Hass und Hetze im Internet bestehen, zusammengeführt. Die ZMI BKA dient einer effektiven Strafverfolgung von Straftaten im Internet wie Propagandadelikten, Volksverhetzungen oder Bedrohungen. Zu den kooperierenden Meldestellen gehören etwa: „HessenGegenHetze", „REspect!", „die medienanstalten" und „Justiz und Medien – konsequent gegen Hass" (vgl. Tabelle 30).[103]

Zudem ist das BKA federführend bei der Bekämpfung terroristischer Inhalte. Die Abteilung Polizeilicher Staatsschutz mit Unterstützung der Abteilung Islamistisch motivierter Terrorismus/Extremismus übernimmt beim BKA die zentrale Rolle bei der Umsetzung der TCO-VO in Deutschland. Es ist die allein bevollmächtigte Behörde für den Erlass und die Überprüfung von Entfernungsanordnungen der TCO-VO und die Bearbeitung von Gefahrensachverhalten der Hostingdienste.

Das BKA steht hierbei im engen Austausch mit anderen staatlichen Einrichtungen, die ebenfalls auf die Verfolgung von Internetkriminalität spezialisiert sind, etwa der Zentral- und Ansprechstelle Cybercrime NRW bei der Staatsanwaltschaft Köln (ZAC NRW) oder der Zentralstelle zur Bekämpfung der Internet- und Computerkriminalität in Hessen (ZIT Hessen).

Das BKA bietet jedoch kein allgemeines Meldeportal für illegale Internetinhalte i. S. der TCO-VO an. Die Kontaktstelle des BKA besteht für die Kommunikation mit einem von einer Entfernungsanordnung betroffenen Hostingdienstes. Bürger können strafrechtlich relevante Inhalte mit Internetbezug an ihre zuständigen Polizeibehörden oder andere Meldestellen melden.

[103] Vgl. https://www.bka.de/DE/KontaktAufnehmen/HinweisGeben/MeldestelleHetzeImInternet/meldestelle_node.html, abgerufen am 22.09.23

6.2 Weitere Behörden in Deutschland

Neben dem BKA sind weitere Stellen im TCO-VO Prozess eingebunden.

- Die **Bundesnetzagentur** ist betraut mit der Kontrolle und ggf. Sanktionierung von Hostingdienstanbietern, die Entfernungsanordnungen nicht nachkommen. Ihr obliegt die Zuständigkeit zur Beurteilung der spezifischen Maßnahmen nach Art. 5 Abs. 4-8 TCO-VO sowie weitere Bußgeldzuständigkeiten (vgl. hierzu Kap. 1).
- Ebenfalls übermitteln Online-Wachen der **Polizei bzw. der Landeskriminalämter** Sachverhalte an das Bundeskriminalamt, zum Erlass von Entfernungsanordnungen[104]

6.3 Rolle der Landesmedienanstalten

Die 14 **Landesmedienanstalten** in Deutschland, welche unter der Dachmarke „die medienanstalten" gemeinsam auftreten, sind gem. Medienstaatsvertrag zuständig für die Zulassung und Aufsicht der privaten Radio- und Fernsehveranstalter. Sie prüfen die Einhaltung von Werberegeln und setzen sich zudem für die Sicherung der Vielfalt im privaten Rundfunk und im Internet ein. Zugleich sind sie auf Basis des Jugendmedienschutz-Staatsvertrags zur Einhaltung des Jugendschutzes im Rundfunk und im Internet ein. Da ein größerer Teil rechtswidriger Inhalte auf Online-Plattformen auftaucht, die insbes. von Jugendlichen genutzt werden, engagieren sich die Landesmedienanstalten stark im Bereich der Eindämmung von Hassrede und Verletzung der Menschenwürde im Internet. Sie fördern Projekte zur Vermittlung von Medienkompetenz und betreiben eigene Programme, die sich gegen illegale Inhalte im Internet richten (vgl. Tab. 30). Sie zählen damit ähnlich wie Einrichtungen der EU und verschiedene NGOs (vgl. Kap. 7.1 und 8.1) zu den aktiven und professionellen Meldestellen und haben bei vielen Online-Plattformen den Status eines „Trusted Flaggers". Im Sinne der TCO-VO werden sie zur Prüfung vor Erlass von Entfernungsanordnungen beteiligt, falls dies geboten erscheint.[105]

Tab. 30 Landesmedienanstalten und deren Projekte gegen illegale Inhalte im Internet (Stand: August 2023)

Medienanstalt	Projekt gegen illegale Inhalte
Landesanstalt für Medien NRW	▪ Verfolgen statt nur löschen (Meldestelle)
Bayerische Landeszentrale für neue Medien (BLM)	▪ Justiz und Medien – konsequent gegen Hass: Meldestelle mit gesicherter Online-Cloud für die Beweissicherung und Schnittstelle zur Generalstaatsanwaltschaft
Landesanstalt für Kommunikation Baden-Württemberg (LFK)	▪ REspect! (Meldestelle) ▪ handysektor
Niedersächsische Landesmedienanstalt (NLM)	▪ Zentralstelle zur Bekämpfung von Hasskriminalität im Internet – Niedersachsen

[104] Vgl. https://www.bka.de/DE/UnsereAufgaben/Deliktsbereiche/PMK/TCO-VO/TCO-VO_node.html#:~:text=Das%20BKA%20bietet%20kein%20allgemeines,an%20ihre%20zust%C3%A4ndigen%20Polizeibeh%C3%B6rden%20melden, abgerufen am 22.09.23
Vgl. https://www.bka.de/DE/KontaktAufnehmen/HinweisGeben/MeldestelleHetzeImInternet/ZMIProzess/zmiprozess_node.html, abgerufen am 22.09.23

[105] Vgl. § 2 TerrOIBG

Medienanstalt	Projekt gegen illegale Inhalte
	▪ Kooperationsvereinbarung Hatespeech darf nicht folgenlos bleiben.
Medienanstalt Hessen	▪ HessenGegenHetze (Meldestelle) ▪ #KeineMachtdemHass ▪ MeldeHelden App
Medienanstalt Rheinland-Pfalz	▪ Verfolgen und Löschen
Sächsische Landesanstalt für privaten Rundfunk und neue Medien (SLM)	
Medienanstalt Berlin-Brandenburg (mabb)	▪ Verfolgen statt nur löschen
Medienanstalt Hamburg / Schleswig-Holstein (MA HSH)	▪ Hingucker
Medienanstalt Sachsen-Anhalt	-
Thüringer Landesmedienanstalt (TLM)	-
Medienanstalt Mecklenburg-Vorpommern (MMV)	-
Landesmedienanstalt Saarland (LMS)	▪ Courage im Netz – Gemeinsam gegen Hass und Hetz ▪ Medienkompetenzzentrum der Landesmedienanstalt Saarland
Bremische Landesmedienanstalt (brema)	▪ RIKO – Resignation ist keine Option

Quelle: Goldmedia Analyse 2023

KI-Instrument KIVI der Landesanstalt für Medien NRW

Zur besseren Erfüllung ihrer Aufsichtsfunktion für Telemedienangebote[106] nutzt die Landesanstalt für Medien NRW (LfM) das KI-Instrument KIVI. Der Name KIVI steht für die Verschmelzung der Begriffe KI und vigilare (lat. für wachsam sein).[107]

Das automatisierte Verfahren überprüft Social-Media-Plattformen und Webseiten auf potenzielle Rechtsverstöße, identifiziert sie und bereitet sie zur Prüfung vor.

Das Instrument wurde 2020 im Auftrag der LfM durch die Condat AG in Berlin entwickelt. Die einmaligen Entwicklungskosten lagen zwischen 150.000 und 200.000 Euro. Trainiert wurde die KI anhand von Bild- und Textbeispielen, die von der LfM in der Vergangenheit als Verstoß bewertet wurden. Zu den konkreten Verstoßkategorien zählen Gewaltdarstellungen, Volksverhetzung, die Verwendung verfassungsfeindlicher Kennzeichen und frei zugängliche Pornografie.

Das Instrument durchsucht sieben Online-Plattformen, etwa Twitter/X, YouTube, Telegram und VK (russisches soziales Netzwerk). Täglich wird alternierend eine Online-Plattform für etwa 6 Stunden gescannt. Dabei werden mehr als 10.000 Seiten automatisch durchsucht. Aktuell nicht durchsucht werden Plattformen des Meta-Konzerns, da dies

[106] Vgl. § 88 Abs. 4 Landesmediengesetz Nordrhein-Westfalen

[107] Vgl. https://www.medienanstalt-nrw.de/zum-nachlesen/recht-und-aufsicht/mit-kuenstlicher-intelligenz-zu-einer-modernen-medienaufsicht.html, abgerufen am 22.09.23

der Konzern nicht zulässt (Stand: September 2023). Eine Ausweitung des KIVI-Monitorings wäre möglich, hierfür wären jedoch weitere Ressourcen notwendig.

Identifizierte Verdachtsfälle, die das Tool ausweist, werden von ca. 5 Mitarbeitenden der LfM zu regulären Bürozeiten geprüft.[108] Wenn sich der Verdacht bestätigt, wird der Fall an hauseigene Juristinnen und Juristen weitergegeben. Bei justiziablen Fällen werden die Online-Plattformen informiert, dies geschieht etwa 25-mal pro Monat.

Die LfM zeigt sich außerordentlich zufrieden mit der Leistungsfähigkeit des Instruments. Nach eigenen Angaben konnten die Zahl der Strafanzeigen zu früheren Vergleichsmonaten verdoppelt werden.[109] Aktuell arbeiten fast alle Medienanstalten in Deutschland mit der durch die LfM entwickelten KI-Lösung. Auch Nutzungsanfragen aus anderen europäischen Ländern hat die LfM für Ihr KI-Instrument KIVI bereits erhalten.

7 Meldende und unterstützende Strafverfolgungsbehörden und Einrichtungen der EU

7.1 Strafverfolgungsbehörden der EU

Die **EU Internet Referral Unit** (EU IRU) spürt terroristische und gewalttätige extremistische Inhalte im Internet und in sozialen Medien auf und analysiert sie. Die EU IRU wurde 2015 gegründet und ist beim Europäischen Zentrum für Terrorismusbekämpfung von Europol angesiedelt. Ihre Arbeit umfasst mehrere Sprachgruppen und Rechtsordnungen. Sie liefert strategische Erkenntnisse über den dschihadistischen Terrorismus und Informationen, die bei strafrechtlichen Ermittlungen verwendet werden können.

Die EU IRU hat die folgenden Kernaufgaben:

- Unterstützung der zuständigen EU-Behörden durch die Bereitstellung strategischer und operativer Analysen;
- Kennzeichnung von terroristischen und gewalttätigen extremistischen Online-Inhalten und Weitergabe an die zuständigen Partner;
- Aufspüren von Internet-Inhalten, die von Schleusernetzwerken genutzt werden, um Migranten und Flüchtlinge anzulocken, und deren Entfernung beantragen;
- Zügige Durchführung und Unterstützung des Verweisungsverfahrens in enger Zusammenarbeit mit der Branche[110]

Das **European Union Internet Forum** (EUIF) ist eine öffentlich-private Partnerschaft der EU zur Bekämpfung terroristischer Inhalte sowie von sexuellem Kindesmissbrauch

[108] Vgl. https://www.bpb.de/lernen/digitale-bildung/werkstatt/513732/ki-in-der-medienaufsicht-was-leistet-das-tool-kivi, abgerufen am 22.09.23

[109] Vgl. https://www.medienanstalt-nrw.de/zum-nachlesen/recht-und-aufsicht/mit-kuenstlicher-intelligenz-zu-einer-modernen-medienaufsicht.html, abgerufen am 22.09.23

[110] Vgl. https://www.europol.europa.eu/about-europol/european-counter-terrorism-centre-ectc/eu-internet-referal-unit-eu-iru, abgerufen am 22.09.23

und gewalttätigem Rechtsextremismus im Internet. Es wurde von der EU-Kommission kurz nach der Gründung der EU IRU im Jahr 2015 ins Leben gerufen. Die EU IRU ist seither Mitglied des EUIF.

Die vorrangigen Bereiche, in die das Forum seine Bemühungen richtet, sind:

- die Umsetzung des EU-Krisenprotokolls (EUCP),
- Reaktionen auf rechtsextreme und gewalttätige Online-Inhalte und
- Reaktionen auf neue Herausforderungen.[111]

Beim **Europäischen Zentrum zur Terrorismusbekämpfung** (ECTC) werden seit 2016 die Kapazitäten von Europol, im Bereich der Strafverfolgung und zur Bekämpfung des Terrorismus, zusammengeführt. Dazu gehören beispielsweise Analyseprojekte im Bereich Terrorismusbekämpfung. Durch die Gründung des Zentrums konnte der Informationsaustausch zwischen den Behörden deutlich gesteigert werden.[112]

Das **Digital Europe Programme** (DIGITAL) der Europäischen Union unterstützt Safer Internet Centres in 27 europäischen Ländern mit dem Ziel, die Medienkompetenz von Kindern, Eltern und Lehrern zu fördern, für mögliche Risiken im Internet zu sensibilisieren und Kindern und Jugendlichen eine telefonische Beratung bei Online-Problemen anzubieten. Außerdem werden Meldestellen für illegale Inhalte finanziert. In Deutschland wird das Safer Internet Centre durch den **Verbund Safer Internet DE** umgesetzt. Diesem gehören neben dem Awareness Centre klicksafe die Internet-Hotlines internet-beschwerdestelle.de (betrieben von eco und FSM) und jugendschutz.net sowie die Helpline für Kinder- und Jugendliche Nummer gegen Kummer an.

7.2 Plattformen und Datenbanken der EU zur Strafverfolgung

Die EU IRU betreibt die von Europol entwickelte **EU-Plattform zur Bekämpfung von illegalen Online-Inhalten PERCI** (Plateforme Européenne de Retraits de Contenus Illégaux sur Internet). Über diese Plattform können die Mitgliedstaaten Meldungen und verbindliche Entfernungsanordnungen gemäß TCO-VO an Hostingdienste übermitteln, um alle Arten von terroristischen Inhalten zu entfernen und damit die Umsetzung der TCO-VO zu erleichtern. PERCI wurde im Juli 2023 in Betrieb genommen. In der Übergangsphase wurde die **Internet Referral Management Application (IRMA**) für Referrals genutzt. Die Entfernungsanordnungen erfolg(t)en über die **Secure Information Exchange Network Application (SIENA)**.[113] Die EU IRU betreibt zudem das Portal

[111] Vgl. https://home-affairs.ec.europa.eu/networks/european-union-internet-forum-euif_en, abgerufen am 07.09.23

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

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

Check the Web – eine Referenzbibliothek zu dschihadistischer Online-Terrorpropaganda einschl. rechtsterroristischer Inhalte. Hierbei handelt es sich um ein operatives Instrument zur Unterstützung der EU-Mitgliedstaaten bei der Erkennung von neuen Inhalten, Trends und Mustern im Kontext von Terrorpropaganda.

7.3 Weitere Projekte der EU zur Terrorismusbekämpfung

Das **EU-Zentrum für Informationsgewinnung und -analyse (INTCEN**) (Intelligence Analysis Centre) (vor März 2012 Joint Situation Centre, SitCen oder JSC) ist ein Organ des Europäischen Auswärtigen Dienstes und hat neben dem Satellitenzentrum der Europäischen Union (SatCen) und der Intelligence Division nachrichtendienstliche Aufgaben.[114]

SIRIUS ist ein von der EU finanziertes Projekt, das Strafverfolgungs- und Justizbehörden den Zugang zu grenzüberschreitenden elektronischen Beweismitteln im Rahmen von strafrechtlichen Ermittlungen und Verfahren erleichtert.

Horizont 2020 war das EU-Forschungs- und Innovationsförderprogramm für den Zeitraum 2014-2020 mit einem Budget von fast 80 Milliarden Euro. Das Programm wurde von Horizont Europa abgelöst. Für Horizont 2020 gab es eine Ausschreibung für Management von Informations- und Datenströmen zur Bekämpfung von (Cyber-)Kriminalität und Terrorismus. In der folgenden Tabelle werden ausgewählte Projekte aufgeführt, die vom Rahmenprogramm Horizont 2020 finanziert wurden.

Tab. 31 Auswahl EU geförderte Horizont2020 Projekte

Projekt	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 organized 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 organized crime.	https://roxanne-euproject.org/

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

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 analyzing terrorist-related online contents and financing activities	https://www.h2020-dante.eu/
PROTON	Modelling the processes leading to organized crime and terrorist networks	https://www.projectproton.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 organizations.	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/

Quelle: Goldmedia Analyse 2023

8 Nicht-staatliche Meldestellen und Datenbanken

Rechtswidrige oder anderweitig problematische Angebote, die durch Hostingdienste verbreitet werden, können zusätzlich zu den sie direkt verbreitenden Hostingdiensten auch anderen Stellen gemeldet werden. Hierzu zählen behördliche oder öffentliche Stellen, aber auch nicht-staatliche Stellen, die etwa von Internet-Branchenverbänden als Instrument der Selbstregulierung betrieben werden. Zudem existieren eine Reihe von Nicht-Regierungsorganisationen, die sich zivilgesellschaftlich mit Aspekten der Inhaltemoderation auseinandersetzen, insbesondere im Bereich antisemitischer oder anderweitig hasserfüllter Inhalte. Diese unterhalten in der Regel keine eigenen Meldestellen, können aber mit Online-Plattformen im Bereich der Prävention kooperieren.

8.1 Nicht-staatliche Meldestellen in Deutschland

Melde- bzw. Beschwerdestellen sind ein wirksames Instrument, um Hostingdienste auf potenzielle Verstöße gegen geltendes Recht zu informieren, die nicht durch die interne Inhaltemoderation der Anbieter erkannt wurden. Die relevantesten Meldestellen in Deutschland werden im Folgenden kurz vorgestellt.

Die Ausdrücke „Meldungen" und „Beschwerden" werden häufig austauschbar verwendet, um auf Mitteilungen hinzuweisen, die von Internetnutzern eingereicht werden, wenn sie auf problematische oder unangemessene Inhalte stoßen. Die Begriffe „Meldung" und „Beschwerde" werden auch im Folgenden synonym verwendet.

Internet-Beschwerdestelle.de

Bei einer Beschwerdestelle können Nutzer über Onlineformulare Internet-Inhalte melden, die Sie für rechtswidrig halten. Eine wichtige Online-Beschwerdestelle in Deutschland ist unter **Internet-Beschwerdestelle.de** erreichbar. Das Angebot unter Internet-Beschwerdestelle.de ist ein gemeinsames Projekt von eco - Verband der Internetwirtschaft (eco) und der Freiwilligen Selbstkontrolle Multimedia-Diensteanbieter (FSM) (s.u.).

Mit Internet-Beschwerdestelle.de bieten eco und FSM seit 2004 eine Anlaufstelle für Internetnutzer, um sich über den sichereren Umgang mit dem Internet zu informieren und Beschwerden einzureichen. Sie sind zudem Gründungsmitglieder der Vereinigung **International Association of Internet Hotlines (INHOPE)**, dem Dachverband von Internet-Hotlines, über den Hotlines auf internationaler Ebene zusammenarbeiten.

Dort werden eingegangene Beschwerden von der jeweiligen Institution juristisch geprüft und, wenn der gemeldete Inhalt gegen die einschlägigen Jugendmedienschutzgesetze bzw. einschlägigen Strafgesetze verstößt, können weitere Schritte eingeleitet werden:

Der Inhalte-Anbieter wird direkt aufgefordert, den Inhalt abzuändern bzw. der Host-Provider gebeten, die Entfernung des Inhaltes zu veranlassen. In gravierenden Fällen kann die Beschwerde in anonymisierter Form auch direkt an die zuständige staatliche Stelle weitergeleitet werden. Beschwerden über illegale Online-Inhalte, die nicht auf einem Server in Deutschland liegen, leiten eco und FSM an die zuständige INHOPE-Hotline weiter.

Beschwerden, die über Internet-Beschwerdestelle.de eingereicht werden, bearbeiten eco und FSM grundsätzlich entsprechend einer Aufgaben- und Zuständigkeitsverteilung. FSM: World Wide Web (bei Bezug zu einem eco Mitglied in Kooperation mit eco), Mobile Inhalte & Apps und Chat. eco: Newsgroups, Spam / E-Mail, Diskussionsforen sowie Peer-to-Peer.

eco – Verband der Internetwirtschaft

Der eco – Verband der Internetwirtschaft ist seit 1995 die Interessenvertretung der Internetwirtschaft in Deutschland und mit rund 900 Mitgliedern der größte Verband der Internetwirtschaft in Europa. An seine seit 1996 bestehende Beschwerdestelle kann sich jeder Internetnutzer – auf Wunsch auch anonym – wenden, der im Internet auf jugendmedienschutzrelevante Inhalte stößt oder sich über unerlaubte E-Mail-Werbung beschweren möchte. Zum Teil nutzen auch Dienste die Beschwerdestelle, um sich bei komplexeren Fällen bei der Inhaltemoderation eine externe juristische Meinung einzuholen.

Beim eco werden etwa 5 Personen werktäglich in der Beschwerdestelle beschäftigt. Alle haben eine erfolgreiche juristische Ausbildung (mind. Staatsexamen) absolviert. Bei der Sichtung und rechtlichen Bewertung der Beschwerden werden keine automatisierten Verfahren eingesetzt.

Im Jahr 2022 erreichte die Beschwerdestelle rd. 18.000 Beschwerden, von denen sich 49,2 Prozent als berechtigt herausstellten. Der Großteil der berechtigten Meldungen (91,6 Prozent) bezog sich auf den Bereich Kinderpornografie, ein sehr geringer Anteil bezog sich auf verfassungsfeindliche Inhalte (0,4 Prozent).

In 8.904 Fällen kam es durch Notice-and-Takedown zur Entfernung von Inhalten durch die Hostingdienste. Die Erfolgsquote lag hierbei insgesamt bei 97,7 Prozent, auch wenn nur 29,6 Prozent der inkriminierten Inhalte bei deutschen Diensten gehostet wurde. Die Entfernungsquote bei Kinderpornografie bei deutschen Hostingdiensteanbietern liegt bei 100 Prozent. Durchschnittlich werden Inhalte nach 2,5 Tagen durch die Hostingdienste entfernt.[115]

Aus Sicht des eco sind es nicht nur die sehr großen Online-Plattformen, bei denen extremistische Inhalte verbreitet werden. Aus ihrer Erfahrung sind kleine und mitunter private Diskussionsforen weiter sehr aktiv. Da es bei diesen oft an leicht auffindbaren Meldeverfahren mangelt, treten Nutzer oft mit der Beschwerdestelle des eco in Kontakt, um dort Inhalte zu melden.

Bei der Strafverfolgung kooperiert die Beschwerdestelle u. a. mit dem BKA, der Zentral- und Ansprechstelle Cybercrime NRW (ZAC) sowie normalen Polizeidienststellen. Bei der internationalen Strafverfolgung kommt dem INHOPE-Netzwerk eine herausgehobene Bedeutung zu.

FSM – Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V.

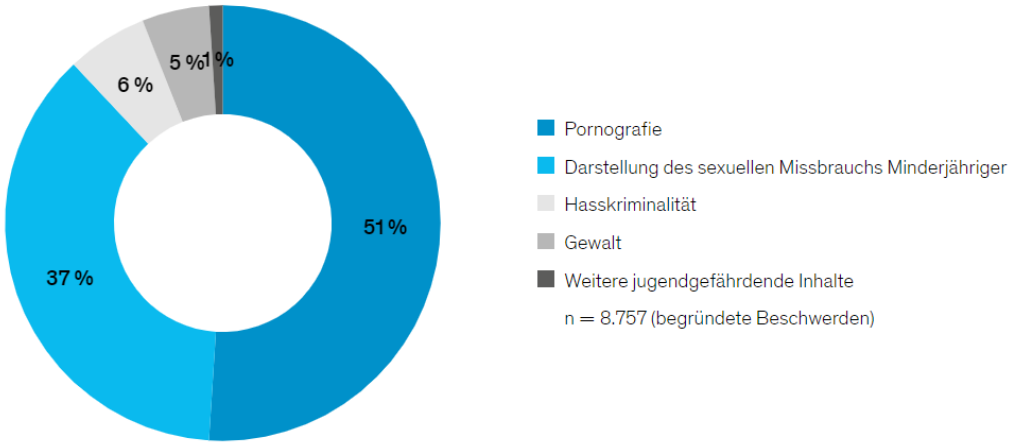
Die Freiwillige Selbstkontrolle Multimedia-Diensteanbieter e.V. (FSM) widmet sich in erster Linie dem Jugendschutz und der Bekämpfung illegaler, jugendgefährdender und entwicklungsbeeinträchtigender Inhalte in Online-Medien. Im Mittelpunkt der Arbeit der FSM steht vor allem die Arbeit als Selbstkontrolle der Diensteanbieter sowie die

[115] Vgl. https://www.eco.de/wp-content/uploads/2023/03/eco_beschwerdestelle_jahresbericht_2022.pdf, abgerufen am 22.09.23

Bearbeitung von Beschwerden über rechtswidrige, jugendgefährdende und entwicklungsbeeinträchtigende Inhalte von Mitgliedsunternehmen, aber auch von Nicht-Mitgliedern. Grundsätzlich wird jede Beschwerde bei der FSM manuell gesichtet und gegebenenfalls auch einem unabhängigen Entscheidungsgremium, dem Beschwerdeausschuss, vorgelegt.[116]

Im Jahr 2022 gingen bei der FSM-Beschwerdestelle insgesamt 12.956 Beschwerden über illegale oder jugendgefährdende Online-Inhalte ein.

Abb. 9: Aufteilung begründeter Beschwerden der FSM-Beschwerdestelle nach Beschwerdegrund 2022



Quelle: FSM Jahresbericht 2022 „Beschwerdestelle", online unter: https://jahresbericht.fsm.de/2022/beschwerdestelle/, abgerufen am 19.09.23

In 68 Prozent der Fälle (8.757 Meldungen) handelte es sich um begründete Beschwerden, d. h. um Inhalte, die gegen deutsche Jugendmedienschutzgesetze verstoßen. Von den begründeten Beschwerden bezogen sich 51 Prozent (4.455 Fälle) auf pornografische Inhalte und 37 Prozent (3.224 Fälle) auf Darstellungen des sexuellen Missbrauchs Minderjähriger.

Insgesamt wurden 55 Prozent der geprüften Missbrauchsdarstellungen Minderjähriger auf deutschen Servern gehostet. Diese leitet die FSM sofort an das BKA weiter und informiert, nach abgeschlossener Beweissicherung, im Notice-and-Takedown-Verfahren den Hostingdienst. Die Löschung solcher Inhalte dauerte im Schnitt 1,5 Tage nach Eingang der Beschwerde, wobei die insgesamte Entfernungsquote solcher Inhalte bei 100 Prozent lag.

Jugendschutz.net

Jugendschutz.net wurde 1997 gegründet und fungiert als gemeinsames Kompetenzzentrum von Bund und Ländern für den Schutz von Kindern und Jugendlichen im Internet. Jugendschutz.net sichtet Angebote im Netz auf Verstöße gegen den Jugendschutz und nehmen Beschwerden entgegen. Im Fokus stehen dabei Themen und Dienste, die bei Risiken für Kinder und Jugendliche besonders relevant sind. Das Team von jugendschutz.net sichtet die gemeldeten Angebote, bewertet die Inhalte unter rechtlichen Aspekten und prüft, wer für das Angebot verantwortlich ist.

[116] Vgl. https://www.fsm.de/unternehmen/angebot/#beschwerdeausschuss, abgerufen am 19.09.23

Jugendschutz.net geht mit Beschwerden wie folgt vor:

- Wenn ein Verantwortlicher eines Dienstes bekannt ist, nimmt jugendschutz.net Kontakt auf und fordert die Beseitigung von Verstößen gegen Jugendschutzbestimmungen. Falls der Verantwortliche in Deutschland ansässig ist und nicht reagiert oder ggf. seine freiwillige Selbstkontrolleinrichtung untätig bleibt, wird der Fall an die Kommission für Jugendmedienschutz (KJM) weitergeleitet, die medienrechtliche Verfahren einleitet.
- Wenn der Verantwortliche nicht identifiziert werden kann, wird der Anbieter des Hostingdienstes gebeten, die Verstöße zu beseitigen. Falls der Anbieter Mitglied einer freiwilligen Selbstkontrolleinrichtung (z. B. eco) ist, wird der Verstoß an diese weitergeleitet. Darüber hinaus leitet jugendschutz.net Angebote zur Indizierung durch die Prüfstelle für jugendgefährdende Medien (Bundeszentrale für Kinder- und Jugendmedienschutz), sodass diese Angebote aus Suchmaschinen entfernt werden können.
- In Fällen von ausländischen Hostingdiensten ohne Kontakt in Deutschland leitet jugendschutz.net die Fälle an kooperierende Beschwerdestellen durch internationale Netzwerken wie INHOPE (International Association of Internet Hotlines) und INACH (International Network Against Cyber Hate) weiter.

Hilfezentren in Deutschland

Das Awareness Center **Klicksafe** ist eine EU-Initiative für mehr Sicherheit im Netz. Klicksafe hat zum Ziel, die Online-Kompetenz der Menschen zu fördern und sie mit vielfältigen Angeboten beim kompetenten und kritischen Umgang mit dem Internet zu unterstützen. Es wird koordiniert von der Medienanstalt Rheinland-Pfalz und gemeinsam mit der Landesanstalt für Medien NRW umgesetzt. Seit 2008 koordiniert Klicksafe auch das Safer Internet Centre DE.[117]

HateAid ist eine gemeinnützige Organisation, die sich für Menschenrechte im digitalen Raum einsetzt und sich auf gesellschaftlicher wie politischer Ebene gegen digitale Gewalt und ihre Folgen engagiert.[118]

Als Vernetzungsstelle gegen hate Speech verbessert **Das NETTZ** die Rahmenbedingungen für Engagement gegen Hass im Netz. Sie verbinden Initiativen und Aktivisten aus der Community der digitalen Zivilcourage (aktuell rd. 138 Akteure). Dafür arbeiten sie eng mit zivilgesellschaftlichen Organisationen, politischen Instanzen und IT-Unternehmen zusammen.[119]

Stark im Amt ist ein Portal für Kommunalpolitik gegen Hass und Gewalt. Sie ist die erste zentrale Anlaufstelle, die Volksvertretern auf kommunaler Ebene mit Informationen und Orientierung der Strafverfolgung von Hassnachrichten versorgt.[120]

Die **Amadeu Antonio Stiftung** setzt sich dafür ein, die Zivilgesellschaft in Deutschland gegen Antisemitismus, Rassismus und Rechtsextremismus zu stärken. Dazu unterstützt sie über 1000 lokale Initiativen und Projekte in Jugendkultur, Schulen, Opferschutz,

---

[117] Vgl. https://www.klicksafe.de/die-initiative, abgerufen am 22.09.23

[118] Vgl. https://hateaid.org/, abgerufen am 22.09.23

[119] Vgl. https://www.das-nettz.de/, abgerufen am 22.09.23

[120] Vgl. https://www.stark-im-amt.de/rat-und-tat/online-hetze/anzeigen-von-hassnachrichten/, abgerufen am 22.09.23

Flüchtlingsinitiativen oder Demokratieprojekte finanziell, durch Aufklärung, Öffentlichkeitsarbeit und kommunale Netzwerke.

Nummer gegen Kummer e. V. ist die Dachorganisation des größten, kostenfreien, telefonischen Beratungsangebotes für Kinder, Jugendliche und Eltern in Deutschland.

Violence Prevention Network ist eine deutsche Nichtregierungsorganisation, die im Bereich Extremismusprävention sowie Deradikalisierung extremistisch motivierter Gewalttäter aktiv ist. Ihre Digitalsparte (**Violence Prevention Network Digital**) bringt die Erkenntnisse und Erfahrungen der verschiedenen Off- und Online Projekte des Trägers zusammen, entwickelt neue digitale Projekte, die die bestehenden Strukturen der Präventionspraxis zielführend ergänzen und erprobt neue Ansätze der internetbasierten Radikalisierungsprävention.

8.2 Internationale Organisationen, Gremien und Datenbanken

Global Internet Forum to Counter Terrorism

Das **Global Internet Forum to Counter Terrorism (GIFCT)** ist eine Nichtregierungsorganisation, die Terroristen und gewalttätige Extremisten daran hindern soll, digitale Plattformen zu nutzen. Das Forum wurde 2017 von Facebook, Microsoft, Twitter/X und YouTube gegründet, um die technische Zusammenarbeit zwischen den Mitgliedsunternehmen zu fördern, relevante Forschung voranzutreiben und Wissen mit kleineren Plattformen zu teilen. In 2023 übernahm Meta den Vorsitz vom GIFCT, welches mittlerweile über 20 verschiedene Online-Plattformen als Mitglieder zählt, die sich branchenübergreifend gegen die Verbreitung von terroristischen und gewalttätigen extremistischen Inhalten im Internet engagieren. Zu den Partnern von GIFCT zählt u. a. Tech Against Terrorism (s. u.).[121]

Das GIFCT betreibt eine **Hash-Sharing-Datenbank**, in welche die Mitgliedsunternehmen terroristische, gewalttätige bzw. extremistische Inhalte, die über ihre Plattformen geteilt werden in verschiedenen Hash-Wert-Varianten bereitstellen.[122] Zudem hat Meta jüngst einen **„Hasher-Matcher-Actioner"** (HMA) auf Open-Source-Basis bereitgestellt, mit dem jede Online-Plattform die eigenen Inhalte hashen und mit den Inhalten der GIFCT-Datenbank oder anderen Hash-Datenbanken vergleichen kann.[123]

Das **Global Network on Extremism and Technology (GNET)** ist der akademische Forschungszweig des GIFCT und zielt darauf ab, die Art und Weise, wie Terroristen Technologie nutzen, besser zu verstehen.

Tech Against Terrorism

Tech Against Terrorism (TAT) ist eine vom UN Counter Terrorism Executive Directorate (UN CTED) ins Leben gerufene und unterstützte Initiative, die mit der globalen Tech-

[121] Vgl. https://gifct.org/about/, abgerufen am 22.10.23

[122] Vgl. https://gifct.org/hsdb/, abgerufen am 22.10.23

[123] Vgl. https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/, abgerufen am 23.10.23

Industrie zusammenarbeitet, um gegen die terroristische Nutzung des Internets vorzugehen und dabei die Menschenrechte zu achten. Ihr Aktionsplan stützt sich auf Öffentlichkeitsarbeit, Wissensaustausch und praktische Unterstützung. Tech Against Terrorism veröffentlich u. a. jährlich einen Bericht zu den vorherrschenden terroristischen Bedrohungen im Internet, der sich an die Öffentlichkeit richtet.[124]

Tech Against Terrorism hat die **Knowledge Sharing Platform** ins Leben gerufen. Dabei handelt es sich um eine Sammlung von Instrumenten (interaktive Tools und Ressourcen), mit denen sich Startups und kleine Tech-Unternehmen besser vor der terroristischen Ausnutzung ihrer Dienste schützen können.

Im Jahr 2018 gründeten sie ebenfalls das **Data Science Network**, das weltweit erste Netzwerk von Experten, die an der Entwicklung und dem Einsatz automatisierter Lösungen arbeiten, um die Nutzung kleinerer Technologieplattformen durch Terroristen zu bekämpfen und dabei die Menschenrechte zu achten.

Die von Tech Against Terrorism entwickelte **Terrorist Content Analytics Platform (TCAP)** wurde im November 2020 mit Unterstützung von Public Safety Canada eingeführt. Die TCAP soll die Nutzung des Internets durch Terroristen unterbinden, indem sie die schnelle und präzise Entfernung terroristischer Inhalte erleichtert. Ein Team von internen Open-Source-Intelligence-Analysten verfolgt dabei die terroristische Migration über eine Vielzahl von Technologieplattformen und meldet URLs mit terroristischen Inhalten an die TCAP. Sie verfolgen den Aufbau der weltweit größten Datenbank mit verifizierten terroristischen Inhalten, die in Echtzeit von verifizierten terroristischen Kanälen auf Messaging-Plattformen und Apps gesammelt werden (einschl. hashes die mit GIFCT geteilt werden). Des Weiteren unterstützen sie kleinere technische Plattformen zur Verbesserung der Inhaltsmoderation.[125][126]

National Center for Missing & Exploited Children

Das **National Center for Missing & Exploited Children** (NCMEC) ist die größte Kinderschutzorganisation der Vereinigten Staaten. Sie setzen sich für den Schutz von Kindern ein und schaffen wichtige Ressourcen für sie und die Menschen, die sie schützen. Dazu gehört gegen Child Sexual Abuse Material (CSAM) im Internet vorzugehen, welche praktisch in allen Online-bereichen zu finden sind.[127] Sie betreiben die **CyberTipline** - ein Online-Meldesystem für alle Arten des sexuellen Missbrauchs von Kindern im Internet. Sie reagieren jedes Jahr auf Millionen von Meldungen über sexuellen Missbrauch von Kindern im Internet. Darüber hinaus bieten sie Opfern und Überlebenden von sexuellem Kindesmissbrauch zahlreiche Ressourcen und Unterstützung an. Das NCMEC bietet auch einen Dienst namens Take It Down an, der dabei hilft, Nackt-, Teilnackt- oder sexuell eindeutige Fotos und Videos von Minderjährigen zu entfernen, indem den Bildern oder Videos ein eindeutiger Hash-Wert zugewiesen wird. Online-Plattformen können diese Hash-Werte verwenden, um diese Bilder oder Videos in ihren öffentlichen oder unverschlüsselten Diensten zu erkennen und Maßnahmen zur Entfernung dieser Inhalte zu ergreifen.

---

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

[125] Vgl. https://www.terrorismanalytics.org/about, abgerufen am 23.10.23

[126] Vgl. https://terrorismanalytics.org/about/how-it-works, abgerufen am 23.10.23

[127] Vgl. https://www.missingkids.org/theissues/csam, abgerufen am 23.10.23

Lumen-Datenbank

Die **Lumen-Datenbank** sammelt und analysiert rechtliche Beschwerden und Anträge auf Entfernung von Online-Materialien und hilft Internetnutzern, ihre Rechte zu kennen und das Gesetz zu verstehen. Anhand dieser Daten untersucht Lumen die Häufigkeit rechtlicher Bedrohungen und zeigt den Internetnutzern, woher die Entfernung von Inhalten kommt.[128]

9 Verhaltenskodizes der Branche mit Bezug zur Inhaltemoderation

Ein Verhaltenskodex (engl. „Code of Conduct") ist ein Regulierungsinstrument, das oft in Zusammenarbeit zwischen der Europäischen Kommission, Zivilgesellschaft und Unternehmen entwickelt wird. Solche Verhaltenskodizes enthalten in der Regel Selbstverpflichtungen für ihre Unterzeichner, Maßnahmen zu ergreifen, um bestimmte Ziele zu erreichen. Der Beitritt zu einem Verhaltenskodex ist freiwillig.

Diese Branchenkodizes definieren wünschenswerte Verhaltensweisen und können so Online-Plattformen dabei unterstützen, verschiedene, sonst vage Sorgfaltspflichten im Zusammenhang mit der DSA, zu konkretisieren und zu erreichen.

Wesentlich ist vor allem der im Mai 2016 zwischen der Europäischen Kommission und vier großen IT-Konzernen (Facebook, Microsoft, Twitter und YouTube) vereinbarte „Verhaltenskodex zur Bekämpfung illegaler Hassrede im Internet". Der Kodex soll sicherstellen, dass Anträge auf Entfernung von Online-Inhalten durch die Konzerne rasch bearbeitet werden. Die Konzerne hatten sich dabei verpflichtet, den Großteil dieser Anträge innerhalb von 24 Stunden zu prüfen und den Inhalt gegebenenfalls zu entfernen, dabei aber stets den Grundsatz der Meinungsfreiheit zu wahren. Bislang bekennen sich acht Unternehmen zu dem Kodex: Facebook, YouTube, Twitter, Microsoft, Instagram, Dailymotion, Snapchat und jeuxvideos.com.[129]

Ein weiteres wichtiges Abkommen ist der „Verhaltenskodex gegen Desinformation", der seit 2018 existiert und im Juni 2022 in einer neuen Fassung von 34 Akteuren unterzeichnet wurde. Die Unterzeichner haben sich hierbei auf gewissen Selbstregulierungsstandards zur Bekämpfung von Desinformation geeinigt. Dabei verpflichteten sie sich, in mehreren Bereichen tätig zu werden, wie z. B. zur Einschränkung der Verbreitung von Desinformation; der Gewährleistung der Transparenz bei politischer Werbung; zur Verbesserung der Zusammenarbeit mit Faktenprüfern und zum zu einem verbesserten Zugang zu ihren Daten für Forschende.[130]

---

[128] Vgl. https://www.lumendatabase.org/pages/about, abgerufen am 23.10.23

[129] Vgl. https://ec.europa.eu/commission/presscorner/detail/de/qanda_20_1135, abgerufen am 06.09.23

[130] Vgl. https://digital-strategy.ec.europa.eu/de/policies/code-practice-disinformation, abgerufen am 06.09.23