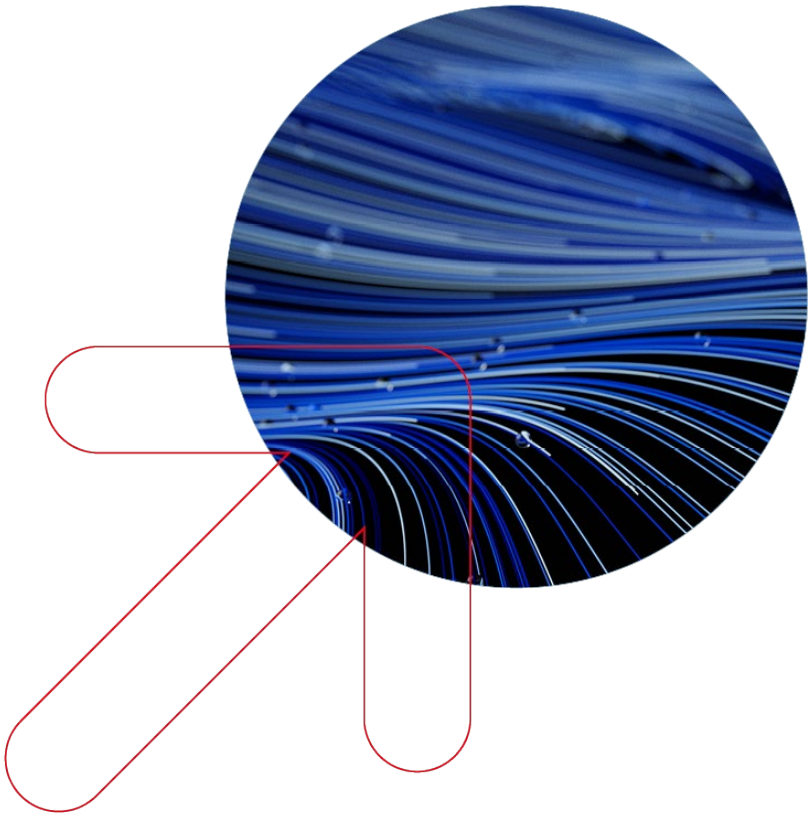
WIK • Diskussionsbeitrag

Nr. 515



Moderation von Inhalten auf Online-Plattformen

Autoren:
Serpil Taş,
Andrea Liebe
Lukas Wiewiorra

WIK
Wissenschaftliches Institut
für Infrastruktur und
Kommunikationsdienste

Bad Honnef, Dezember 2023

Impressum

WIK Wissenschaftliches Institut für
Infrastruktur und Kommunikationsdienste GmbH
Rhöndorfer Str. 68
53604 Bad Honnef
Deutschland
Tel.: +49 2224 9225-0
Fax: +49 2224 9225-63
E-Mail: info@wik.org
www.wik.org

Vertretungs- und zeichnungsberechtigte Personen

Geschäftsführerin und Direktorin	Dr. Cara Schwarz-Schilling
Direktor Abteilungsleiter Smart Cities/Smart Regions	Alex Kalevi Dieke
Direktor Abteilungsleiter Netze und Kosten	Dr. Thomas Plückebaum
Direktor Abteilungsleiter Regulierung und Wettbewerb	Dr. Bernd Sörries
Leiter der Verwaltung	Karl-Hubert Strüver
Vorsitzender des Aufsichtsrates	Dr. Thomas Solbach
Handelsregister	Amtsgericht Siegburg, HRB 7225
Steuer-Nr.	222/5751/0722
Umsatzsteueridentifikations-Nr.	DE 123 383 795

Stand: Juli 2023

ISSN 1865-8997

Bildnachweis Titel: © Robert Kneschke - stock.adobe.com

Weitere Diskussionsbeiträge finden Sie hier:
https://www.wik.org/veroeffentlichungen/diskussionsbeitraege

In den vom WIK herausgegebenen Diskussionsbeiträgen erscheinen in loser Folge Aufsätze und Vorträge von Mitarbeitern des Instituts sowie ausgewählte Zwischen- und Abschlussberichte von durchgeführten Forschungsprojekten. Mit der Herausgabe dieser Reihe bezweckt das WIK, über seine Tätigkeit zu informieren, Diskussionsanstöße zu geben, aber auch Anregungen von außen zu empfangen. Kritik und Kommentare sind deshalb jederzeit willkommen. Die in den verschiedenen Beiträgen zum Ausdruck kommenden Ansichten geben ausschließlich die Meinung der jeweiligen Autoren wieder.
WIK behält sich alle Rechte vor. Ohne ausdrückliche schriftliche Genehmigung des WIK ist es auch nicht gestattet, das Werk oder Teile daraus in irgendeiner Form (Fotokopie, Mikrofilm oder einem anderen Verfahren) zu vervielfältigen oder unter Verwendung elektronischer Systeme zu verarbeiten oder zu verbreiten.

Inhaltsverzeichnis

Abbildungsverzeichnis II

Tabellenverzeichnis II

Zusammenfassung III

Summary IV

1 Einleitung 1

2 Eine Einführung in die Moderation von Inhalten auf digitalen Plattformen 6

2.1 Anreize für die Moderation von Inhalten 7

2.1.1 Intrinsische Anreize: Unternehmensphilosophie und Gewinnmaximierung 7

2.1.2 Extrinsische Anreize: Öffentliche Kritik und regulatorische Bestimmungen 9

2.2 Konsequenzen der Moderation und die Nutzung von Empfehlungssystemen 9

3 Ein Überblick über die Herausforderungen der derzeitigen Instrumente zur Moderation von Inhalten 11

3.1 Genauigkeit und Zuverlässigkeit 13

3.1.1 Matching-Algorithmen 13

3.1.2 Prädiktive Algorithmen 16

3.1.3 Menschliche Moderation 20

3.2 Transparenz und Verantwortlichkeit 21

4 Regulatorischer und rechtlicher Rahmen für die Moderation von Inhalten 23

4.1 Förderung von Transparenz 30

4.2 Einführung einer Aufsichtsstruktur 36

4.2.1 Nationale Aufsicht: Der Koordinator für digitaler Dienste 36

4.2.2 Prüfung: Einhaltung des DSA und Identifizierung vom Problemen 38

5 Schlussfolgerung 42

Literaturverzeichnis 45

Abbildungsverzeichnis

Abbildung 3-1:	Moderationsprozess	11
Abbildung 3-2:	Beispiel - Klassifizierung, Objekterkennung und Segmentierung	16
Abbildung 4-1:	Durch den DSA adressierten Dienste	27
Abbildung 4-2:	Kaskade der DSA-Verpflichtungen	28

Tabellenverzeichnis

Tabelle 3-1:	Einstufung von Inhalten	19
Tabelle 4-1:	Die wichtigsten EU-Verordnungen und Richtlinien zu illegalen Online-Inhalten	26
Tabelle 4-2:	Relevante Verpflichtungen im DSA im Zusammenhang mit der Moderation von Inhalten	31
Tabelle 4-3:	Transparenzberichte der VLOPs bis Oktober 2023	35
Tabelle 4-4:	Instrumente des Auditing	39

Zusammenfassung

Im Hinblick auf die Verabschiedung des Digital Services Act im Jahr 2022, der einen neuen horizontalen Rechtsrahmen für die Regulierung digitaler Dienste vorsieht, analysiert dieses Diskussionspapier die Moderation von Inhalten auf Online-Plattformen. Die Moderation von Inhalten ist eine sehr komplexe und herausfordernde Aufgabe. Sie wird von zahlreichen Faktoren beeinflusst. Zum einen sind dies die Anreize der Plattformanbieter, die durch ihr Interesse an der Generierung von Gewinnen und der Steigerung der Nutzerzahlen geprägt sind. So kann es sein, dass Plattformanbieter entweder eine zu laxe oder eine zu strenge Moderationspolitik verfolgen, um dieses Ziel zu erreichen. Aber auch der regulatorische Rahmen, in dem sie agieren müssen, spielt eine Rolle. Insbesondere Haftungsregelungen können hier einen großen Einfluss haben: sind sie zu lax ausgestaltet, haben sie möglicherweise überhaupt keinen Einfluss auf die Anreize der Plattformanbieter, Inhalte zu moderieren; sind sie zu streng, kann es möglicherweise zu einer Überregulierung des Moderationsprozesses kommen, in dem mehr Inhalte als notwendig moderiert werden.

Hinzu kommen die technischen Herausforderungen, denen sich Moderationssysteme heute stellen müssen. Heutige Ansätze, die in der Regel eine Kombination aus menschlichen und maschinellen Systemen nutzen, sind bei der Identifizierung und Klassifizierung von Inhalten selten zu 100 % akkurat. Zudem ist der Prozess von außen betrachtet wenig transparent, um bei Fehlentscheidungen gegensteuern oder die Plattformanbieter für ihre Moderationsentscheidung zur Rechenschaft ziehen zu können. Denn auf Online-Plattformen, die täglich Millionen von Menschen erreichen können, kann das Entfernen unbedenklicher Inhalte, aber auch die Verbreitung von Belästigungen, sexuellen Inhalten, Hassreden, Urheberrechtsverletzungen gesellschaftlichen und individuellen Schaden anrichten. Der DSA enthält zwar eine Haftungsausschlussklausel für Online-Plattformen für von Nutzern eingestellte Inhalte, das Gesetz legt aber einige Sorgfaltspflichten in Bezug auf die Moderation von Inhalten und deren Transparenz fest. Dies kann dazu beitragen, den Moderationsprozess sichtbarer und nachvollziehbarer zu machen und dysfunktionale Prozesse zu identifizieren und gegenzusteuern. Dazu gehört die Erstellung von Transparenzberichten, die unter anderem Informationen über die moderierten Inhalte, den Einsatz automatisierter Moderationssysteme, die Qualifikation und Kenntnisse der menschlichen Moderatoren, sowie Informationen zu den aus ihrer Tätigkeit resultierenden systemischen Risiken enthalten.

Gleichzeitig wird eine Überwachungsstruktur eingerichtet und den benannten DSC und zugelassenen Forschern werden bei Bedarf zusätzliche Daten zur Verfügung gestellt, um die Einhaltung der DSA zu überprüfen und systemische Risiken, die sich aus den Aktivitäten der Plattformen ergeben, zu erkennen, zu identifizieren und zu verstehen.

Summary

In light of the adoption of the Digital Services Act in 2022, which provides for a new horizontal legal framework for the regulation of digital services, this discussion paper analyses the content moderation on online platforms. Content moderation is a very complex and challenging task. It is influenced by numerous factors. First, platform providers' incentives to moderate content are driven by the need to generate profits and increase usage. For example, platform providers may adopt moderation policies that are either too lax or too strict to achieve this goal. Second, the regulatory framework in which they have to operate also plays a role. In particular, liability rules can have a major impact: if they are too lax, they may not affect the incentives of platform providers to moderate content at all; if they are too strict, they may over-regulate the moderation process and result in more content being moderated than necessary.

On top of this, there are the technical challenges that moderation systems have to deal with today. Current approaches, which typically use a combination of human and automated systems, are rarely 100% accurate in identifying and classifying content. Furthermore, the process is not transparent enough for external parties to take action when wrong decisions are made, or to hold platform providers accountable for their moderation decisions. The removal of harmless content, as well as the spread of harassment, sexual content, hate speech and copyright infringement, can cause significant social and individual harm on online platforms that can reach millions of people every day. Although the DSA includes an exemption from liability for online platforms for content posted by users, the law does impose some duties of care in relation to content moderation and transparency. This can help to make the moderation process more transparent and accountable, and to identify and address dysfunctional processes. This includes the provision of transparency reports that include information on the content moderated, the use of automated moderation systems, the qualifications and knowledge of human moderators, and information on the systemic risks posed by their activities.

At the same time, an oversight structure will be established and additional data will be made available to designated DSCs and authorised researchers if needed to verify compliance with the DSA and to detect, identify and understand systemic risks arising from the platforms' activities.

1 Einleitung

Das Teilen von Inhalten online ist heutzutage allgegenwärtig. Täglich posten Verbraucher Informationen über wichtige Ereignisse in ihrem Leben, Nachrichten, soziale und politische Ansichten und alles, was sonst von Interesse ist (Molina & Sundar, 2022). Doch nicht alle Inhalte sind unproblematisch und harmlos und erreichen dennoch eine große Zahl von Menschen. Die Verbreitung schädlicher und illegaler Inhalte wie sexuelle Inhalte, Belästigung, Hassreden und Inhalte, die gegen Urheberrechte verstoßen ist ein großes Problem (Dias Oliva, 2020). Online-Plattformen stehen daher unter ständigen Druck, Inhalte zu moderieren (Dias Oliva, 2020; Díaz & Hecht-Felella, 2021).

Die Moderation von Inhalten ist ein sehr komplexer und herausfordernder Prozess. Aus rein funktionstechnischer Sicht sind die derzeitigen Moderationsansätze und -methoden oft nicht präzise genug. Die meisten existierenden Instrumente sind noch nicht in der Lage, eine Fehlklassifizierung von Inhalten vollständig zu vermeiden, was dazu führt, dass harmlose Inhalte als schädlich, illegal oder in irgendeiner anderen Weise als gegen geltende Regeln verstoßend eingestuft werden und umgekehrt. Einer der Gründe für dieses Problem ist, dass derzeitige automatisierte Systeme zur Klassifizierung und Identifizierung von Inhalten, deren Anwendung weit verbreitet ist, nicht in der Lage sind, den Kontext eines bestimmten Inhalts vollständig zu verstehen und entsprechend in ihre Bewertung einzubeziehen. So können diese Systeme z.B. einen Inhalt für sich genommen als moderationspflichtig einstufen. In einem breiteren Kontext kann der Inhalt jedoch zulässig sein (Shenkman et al, 2021; Duarte et al., 2017; Singh, 2019; Llansó et al., 2020,Cambridge Consultants, 2019). Studien zu Algorithmen aus der Computer Vision, die in einigen kommerziellen Systemen zur Moderation von Inhalten verwendet werden, haben gezeigt, dass sie versagen können, wenn sie mit Bildern aus der realen Welt konfrontiert werden, mit denen sie während ihrer Entwicklung und ihres Trainings nicht konfrontiert waren. Dies liegt daran, dass diese Algorithmen auf Benchmark-Datensätzen trainiert werden, die die Komplexität realer Situationen und Szenarien möglicherweise nicht vollständig repräsentieren (Qiu & Yuille, 2016). Ein damit verbundenes Problem liegt zudem darin, dass automatisierte Systeme Minderheiten weiter ausgrenzen können. Dies ist insbesondere dann der Fall, wenn die für das Training verwendeten Datensätze unvollständig sind und Minderheiten unterrepräsentieren oder historische Diskriminierungen enthalten (Duarte et al., 2017; Taş et al., 2022). In der Vergangenheit konnten bereits mehrere Fälle nachgewiesen werden, in denen automatisierte Systeme die Inhalte von und über Minderheiten falsch interpretiert und klassifiziert haben (Buolamwini & Gebru, 2018; Kayser-Bril, 2020; Blodgett & O'Connor, 2017). Dies kann dazu führen, dass die Inhalte von Minderheiten nicht ausreichend geschützt werden oder ihre Stimmen zum Schweigen gebracht werden.

Angesichts dieser Grenzen automatischer Systeme ist die Beteiligung menschlicher Moderatoren unumgänglich. So sind sie in der Regel an der Moderation von Inhalte beteiligt. Allerdings sind auch menschliche Moderatoren nicht frei von Fehlern und Vorurteilen. Sie können Zusammenhänge erfassen und natürliche Sprache interpretieren, aber nur, wenn

sie mit ihr vertraut sind. Oft werden die Moderatoren jedoch für Inhalte aus Länder und Kulturen eingesetzt, mit denen sie nicht vertraut sind, was zu Fehlinterpretationen führen kann. Auch sind die Entscheidungen trotz detaillierter Leitlinien nicht frei von Subjektivität und Auslegung. Dies kann zu Inkonsistenzen bei der Entscheidungsfindung und mitunter ebenfalls zu einer systematischen Benachteiligung von Minderheiten führen (Wilson & Land, 2021; Singh, 2019; De Streel et al., 2022; Gillespie, 2020; Cambridge Consultants, 2019; de Gregorio, 2020; Barrett, 2020; Reuber & Fischer, 2022). Darüber hinaus ist diese Arbeit mit einem hohen Maß an psychischer Belastung für die Ausführenden verbunden (Whittaker, 2020; Newton, 2029).

Sowohl menschliche Moderatoren als auch automatisierte Systeme sind notwendig um der Masse an Inhalten, die täglich von Verbrauchern geteilt werden, zu bewältigen. Beide können jedoch Fehler verursachen und inkonsistente Entscheidungen treffen. In einigen Studien wird jedoch argumentiert, dass gewinnmaximierende Plattformen nicht unbedingt ein Interesse an einer fehlerfreien oder konsistenten Moderation von Inhalten oder sogar an einer Verbesserung der automatisierten System haben (Yildirm & Zhang, 2022). Für Plattformen ist die Moderation oft ein Kompromiss. Dieser besteht darin, dass sie einige Nutzer verlieren, deren Inhalte von der Moderation betroffen sind, aber das Engagement der Nutzer stimulieren, die sich sonst durch nicht moderierte Inhalte angegriffen fühlen (Jiménez-Durán, 2022a; Jiménez-Durán; 2022b).

Lange Zeit haben die Plattformanbieter die Moderation von Inhalten selbst in die Hand genommen und eigene Richtlinien festgelegt. Diese können die moralischen Überzeugungen und die Philosophie der Plattformanbieter oder soziale Normen widerspiegeln. Sie können aber auch schlicht darauf abzielen, das Nutzererlebnis zum Ziel der Gewinnmaximierung zu gestalten (Keller & Leerssen, 2020). Da es sich immer noch um gewinnorientierte Unternehmen handelt, deren Ziel es ist, ihre Einnahmen zu erhöhen, indem sie die Interaktion der Nutzer auf und mit der Plattform stimulieren, liegt es nur nahe, dass sie ein System zur Moderation der Inhalte entwickeln, dass dieses Ziel erreicht. Ihr Ansatz führt daher möglicherweise nicht immer zu einem sozial wünschenswerten Ergebnis.

Der Moderationsprozess selbst ist relativ intransparent. Dies macht es schwierig, Anbieter für den Umgang mit Inhalten zur Verantwortung zu ziehen. Die Europäische Union (EU) hat bei der Moderation von Inhalten aus regulatorischer Sicht bisher auch eher zurückhaltend agiert (Sartor et al., 2020). Seit ihrer Verabschiedung im Jahr 2000 stützt sich die EU auf die ***Richtlinie über den elektronischen Geschäftsverkehr (Richtlinie 2000/31/EG, nachfolgend e-Commerce Richtlinie)***[1], die in den letzten 20 Jahren den wichtigsten horizontalen Rechtsrahmen für digitale Dienste bildete. Was die Moderation von Inhalten betrifft, so sind Online-Plattformen von der Haftung für die illegalen Aktivitäten ihrer Nutzer - einschließlich der von ihnen eingestellten Inhalte – gemäß dieser

1 Richtlinie 2000/31/EC des Europäischen Parlaments und des Rates vom 8. Juni 2000 über bestimmte rechtliche Aspekte der Dienste der Informationsgesellschaft, insbesondere des elektronischen Geschäftsverkehrs, im Binnenmarkt („Richtlinie über den elektronischen Geschäftsverkehr“), Amtsblatt der Europäischen Gemeinschaft, L 178, 17.07.2000, S.1-16.

Richtlinie befreit. Eine Haftung entsteht nur dann, wenn der Plattformanbieter von Inhalten, die gegen nationales oder internationales EU-Recht verstoßen, wusste und es versäumt hat, diese zu entfernen (Gellert & Wolters, 2021).

Die Regelung der Haftbarkeit - wenn sie richtig konzipiert ist - kann jedoch dazu beitragen, dass die Plattformen in die Verantwortung genommen werden und so Anreize entstehen ein bestimmtes Maß an Moderation auszuüben und verhindert wird, dass die Moderation in eine ungewollte Richtung geht.

Einige neuere Rechtsinstrumente auf EU-Ebene behalten zwar weitgehend die Klausel zum Haftungsausschluss bei, verlangen aber mehrere prozedurale Sicherungsmechanismen und proaktive Maßnahmen auf Seiten des Plattformanbieters. Dies gilt unter anderem für die ***Richtlinie über audiovisuelle Mediendienste (Richtlinie 2010/13/EU,* Richtlinie (EU) 2018/1808)**[2], die ***Richtlinie zum Urheberrecht im digitalen Binnenmarkt (Richtlinie (EU) 2019/790)***[3] oder die ***Verordnung zur Bekämpfung der Verbreitung terroristischer Online-Inhalte (Verordnung (EU) 2021/784, nachfolgend TERREG-Verordnung)***[4]. Diese Rechtsinstrumente gelten jedoch in der Regel nur für bestimmte Arten von Inhalten oder Plattformen und finden nicht immer horizontal Anwendung. Im Laufe der letzten Jahre haben einzelne Mitgliedstaaten ebenfalls Anstrengungen unternommen bzw. erfolgreich nationale Rechtsvorschriften von unterschiedlichem Umfang und Anwendungsbereich umgesetzt, wie z. B. das deutsche ***Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (nachfolgend Netzwerkdurchsetzungsgesetz (NetzDG))***[5] und das französische Gesetz ***Loi Avia***[6]. Darüber hinaus gibt es branchenweite Selbstregulierungsregeln und -empfehlungen sowie weiche Gesetze, die in erster Linie von der Europäischen Kommission (EK) initiiert wurden und verschiedene Arten illegaler und schädlicher Inhalte adressieren (De Streel et al, 2022;, Castets-Renard, 2021; Hoffman & Gasparoti, 2020). Zusätzlich zu den auf die Anbieter ausgerichteten Vorschriften sind die Mitgliedstaaten verpflichtet, gegen die Verbreitung spezifische Inhalte vorzugehen und dabei unter anderem die ***Richtlinie zur***

---

2 Richtlinie 2010/13/EU des Europäischen Parlaments und des Rates vom 10. März 2010 zur Koordinierung bestimmter Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Bereitstellung audiovisueller Mediendienste (Richtlinie über audiovisuelle Mediendienste), Amtsblatt der Europäischen Gemeinschaft, L 95 , 15.04.2010, S.1-24 in der Fassung der Richtlinie (EU) 2018/1808 des Europäischen Parlaments und des Rates vom 14. November 2018 zur Änderung der Richtlinie 2010/13/EU zur Koordinierung bestimmter Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Bereitstellung audiovisueller Mediendienste (Richtlinie über audiovisuelle Mediendienste) im Hinblick auf sich verändernde Marktgegebenheiten, Amtsblatt der Europäischen Gemeinschaft, L 303 , 28.11.2018, S. 69-92.

3 Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG, Amtsblatt der Europäischen Gemeinschaft, L 130, 17.05.2019, S. 92-125.

4 Verordnung (EU) 2021/784 des Europäischen Parlaments und des Rates vom 29. April 2021 zur Bekämpfung der Verbreitung terroristischer Online-Inhalte, Amtsblatt der Europäischen Gemeinschaft, L 172, 29.04.2021, S. 79-109.

5 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG) vom 1. September 2017 (BGBl. I S. 3352), zuletzt geändert durch Artikel 3 des Gesetzes vom 21. Juli 2022 (BGBl. I S. 1182).

6 LOI no 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet, JUSX1913052L, Journal Officiel de la République Française, 1/181, 25. Juni 2020.

Bekämpfung des sexuellen Missbrauchs und der sexuellen Ausbeutung von Kindern (Richtlinie 2011/93/EU)[7] und die ***Richtlinie zur Bekämpfung des Terrorismus (Richtlinie (EU) 2017/541)***[8] durchzusetzen (Gellert & Wolters, 2021; De Streel et al, 2022). Die Gesetzgebung zur Einschränkung der Verbreitung illegaler und schädlicher Online-Inhalte besteht also aus verschiedenen verbindlichen und nicht verbindlichen Rechtsinstrumenten. Diese Instrumente richten sich an verschiedene Parteien – an die Mitgliedstaaten und die Anbieter –, und enthalten spezifische Vorschriften für verschiedene Arten von Inhalten. Infolgedessen ist der Rechtsrahmen uneinheitlich und fragmentiert (Gellert & Wolters, 2021).

Mit dem ***Gesetz über digitale Dienste (Verordnung (EU) 2022/2065, nachfolgend Digital Services Act (DSA))***[9] hat die EU einen neuen Rechtsrahmen für Vermittlungsdienste und damit Online-Plattformen verabschiedet. Dieser Rahmen legt Verpflichtungen für Online-Plattformen in Bezug auf die Moderation von Inhalten und Transparenz fest. Er zielt darauf ab, die e-Commerce Richtlinie zu reformieren, wobei die Haftungsbefreiung weitgehend erhalten bleibt, aber prozedurale Sicherungsmechanismen und proaktive Maßnahmen vorgesehen sind, die horizontal auf alle Arten von Online-Plattformen und -Inhalten anwendbar sind.

Inspiriert von der Verabschiedung des DSA als neues horizontales Rahmenwerk, der einheitliche Regeln für den Umgang mit Inhalten festlegt, die auf einzelnen Plattformen verbreitet werden, untersucht diese Diskussionsbeitrag die Herausforderungen, die sich bei der Moderation von Inhalten ergeben und wie Regulierung diese abmildern oder beheben kann.

Der Diskussionsbeitrag ist wie folgt aufgebaut. Die Kapitel 2 und 3 befassen sich mit den Herausforderungen der Inhaltsmoderation. Da es sich um einen sozioökonomischen und zugleich technischen Prozess handelt, müssen verschiedene Facetten berücksichtigt werden. Daher wird in Kapitel 2 zunächst ein Überblick darüber gegeben, warum sich Plattformanbieter für die Moderation von Inhalten entscheiden und welche Faktoren ihre Entscheidung beeinflussen können. Die praktischen Probleme der technischen Umsetzung werden dann in Kapitel 3 betrachtet. Dabei werden insbesondere die Problembereiche "Genauigkeit und Zuverlässigkeit" sowie "Transparenz und Verantwortlichkeit" erörtert.

---

7 Richtlinie 2011/92/EU des Europäischen Parlaments und des Rates vom 13.Dezember 2011 zur Bekämpfung des sexuellen Missbrauchs und der sexuellen Ausbeutung von Kindern sowie der Kinderpornografie sowie zur Ersetzung des Rahmenbeschlusses 2004/68/JI des Rates, Amtsblatt der Europäischen Gemeinschaft, L 355, 17.12.2011, S. 1-14.

8 Richtlinie (EU) 2017/541 des Europäischen Parlaments und des Rates vom 15. März 2017 zur Terrorismusbekämpfung und zur Ersetzung des Rahmenbeschlusses 2002/475/JI des Rates und zur Änderung des Beschlusses 2005/671/JI des Rates, Amtsblatt der Europäischen Gemeinschaft, L 88, 31.03.2017, S. 6-21.

9 Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG (Gesetz über digitale Dienste), Amtsblatt der Europäischen Gemeinschaft, L 277, 27.10.2022, S.1-102.

Kapitel 4 befasst sich mit dem derzeitigen Rechtsrahmen. Dabei wird insbesondere auf den DSA eingegangen, aber auch andere Rechtsvorschriften, die die Moderation von Inhalten beeinflussen können, werden kurz beschrieben. Ein besondere Fokus liegt auf den im DSA enthaltenden Verpflichtungen, die die in den Kapiteln 2 und 3 genannten Probleme adressieren. Der Diskussionsbeitrag schließt mit einer Schlussfolgerung.

2 Eine Einführung in die Moderation von Inhalten auf digitalen Plattformen

Die Moderation von Inhalten ist ein zentraler Bestandteil von Plattformen. Gillespie (2018) beschreibt die Moderation sogar als die eigentliche Dienstleistung, die Plattformen anbieten. Plattformen sind entstanden, um das Chaos der Masse an unstrukturierten Informationen im Internet zu bewältigen und die Navigation in der Menge an Informationen zu erleichtern. Dafür kuratieren, organisieren, archivieren und moderieren sie Inhalte (Gillespie, 2018; Sander, 2020).

Obwohl praktisch alle Plattformen Inhalte moderieren, kann der Ansatz, den sie dabei verfolgen, sehr unterschiedlich sein (Roberts, 2017; Gillespie, 2018; Hubley, 2022). Da es traditionell nur sehr wenig gab, was die Plattformen einheitlich per Gesetz moderieren mussten, wurde der Prozess weitgehend von den Plattformen selbst initiiert und kontrolliert (Hubley, 2022). Während etablierte Plattformen wie Facebook, Instagram, TikTok und Co. in den letzten Jahren ihre Bemühungen um die Moderation von Inhalten stetig erhöht haben, treten einige alternative Plattformen wie BitChute, Voat, Parler und Gab bewusst als Verfechter der Meinungsfreiheit auf und moderieren so wenig wie möglich (Rauchfleisch & Kaiser, 2021; Mahl et al., 2023).

Die Praxis hat gezeigt, dass Plattformen bisher einen großen Spielraum bei der Entscheidung haben, ob und wie sie eingreifen (Liu et al., 2022; Gillespie, 2018). Über Nutzungsbedingungen oder sogenannten Community-Richtlinien definieren Plattformen, welches Verhalten und welche Inhalte bei der Nutzung ihrer Dienste erlaubt sind und welche nicht (Nieborg & Poeel; 2018; Mahl et al., 2023). Diese Regeln verbieten zumeist nicht nur offensichtlich illegale Inhalte, sondern in einigen Fällen auch verschiedene Arten von schädlichen Inhalten wie Fehlinformationen, Belästigung, Gewalt und Nacktheit, die Werbekunden abschrecken, Endnutzer vertreiben oder rechtliche Probleme verursachen können (Jiménez-Durán, 2022a; Nieborg & Poeel; 2018).

Die Plattformen haben mehrere Möglichkeiten, ihre Regeln durchzusetzen und auf Verstöße zu reagieren. Zu diesen Möglichkeiten gehören das Kennzeichnen von Inhalten mit Warnhinweisen, die Anwendung von Demonetisierungsmaßnahmen, die Sperrung von Funktionen, die Einschränkung der Sichtbarkeit von Inhalten, die Löschung von Inhalten und der Ausschluss von Nutzern von der Plattform (Gillespie, 2022; Mahl et al., 2023). Während alle diese Maßnahmen darauf abzielen, die Nutzer zu disziplinieren, zielen einige bewusst darauf ab, bestimmte Arten von Inhalten entweder zu entfernen oder ihre Sichtbarkeit und Verbreitung einzuschränken. Letztere stehen im Mittelpunkt dieses Diskussionsbeitrags.

2.1 Anreize für die Moderation von Inhalten

Die Vorgehensweise der einzelnen Plattformen bei der Moderation von Inhalten kann und sollte nicht isoliert betrachtet werden. In der Regel wird die Moderation von einer Reihe an Faktoren beeinflussen (Sander, 2020). Sander (2020) hat mindestens vier Faktoren identifiziert, die sich auf die Bemühungen von Plattformen zur Moderation von Inhalten auswirken. Diese können im Wesentlichen intrinsische und extrinsische Anreize für die Moderation von Inhalten schaffen.

2.1.1 Intrinsische Anreize: Unternehmensphilosophie und Gewinnmaximierung

Der Prozess der Inhaltsmoderation wird zum Teil durch die Unternehmensphilosophie und das Streben nach Gewinnmaximierung bestimmt. Die ***Unternehmensphilosophie*** umfasst laut Sander (2020) die übergreifenden Ziele, die Plattformen bei der Gestaltung ihrer Dienste anstreben. Sie zielt nicht nur auf die Erweiterung der Nutzerbasis der Plattform ab, sondern spiegelt vor allem die Einstellungen und Werte der Gründer und Mitarbeiter der Plattform wider (Sander, 2020). Der zweite Einflussfaktor ist laut Sander (2020) die ***Gewinnmaximierung***, da es sich bei Plattformen überwiegend um gewinnorientierte Unternehmen handelt (Sander, 2020; Hubley, 2022).Gemäß Reuber & Fischer (2022) müssen Plattformen, um erfolgreich zu sein und ihren Gewinn zu steigern, die Zahl der Nutzer erhöhen, die regelmäßig mit der Plattform interagieren. Die Moderation von Inhalten dient dem Zweck, die Nutzer dazu zu bringen, mit der Plattform zu interagieren (Hubley, 2022). Ob eine Plattform eine strenge oder laxe Politik bei der Moderation von Inhalten verfolgt, hängt unter anderem von ihrer Nutzerbasis ab. Die Nutzer sind in der Regel sehr heterogen, was ihre Demografie, ihre Vorlieben und ihre Präferenzen angeht, und neigen daher auch dazu, Inhalte unterschiedlich zu tolerieren, sich mit ihnen zu beschäftigen und sie zu veröffentlichen (Liu et al., 2022).

Je heterogener die Nutzer in Bezug auf ihre allgemeine Einstellung zur Moderation von Inhalten und zu bestimmten Arten von Inhalten sind – was insbesondere auf sehr großen Plattformen der Fall ist – desto komplexer wird die Moderation (Liu et al., 2022). Jiménez-Durán (2022a), Liu et al. (2022), und Madio & Quinn (2023) gehören zu den ersten, die den inhärenten Anreiz von gewinnmaximierenden Plattformanbietern untersuchen, bestimmte Arten von Inhalten bei einer heterogenen Nutzerbasis zu entfernen. Die Autoren zeigen, dass Plattformen – insbesondere solche, die einem werbebasierten Finanzierungsmodell unterliegen – einen Anreiz haben, Inhalte zu moderieren. Ihr Ansatz zielt jedoch eher auf den marginalen Nutzer als auf den durchschnittliche Nutzer ab. Eine geringfügige Erhöhung der Moderation muss zumindest bei einem Teil der Nutzer – z.B. bei den Teil, die sich ohne Moderation angegriffen gefühlt hätte – zu einer erhöhten Interaktion mit der Plattform führen oder neue Nutzer anziehen, während die Interaktion des Teils der Nutzer, dessen Inhalte von der Moderation betroffen sind, nicht zu stark abnimmt. Nur unter dieser Voraussetzung würde ein gewinnmaximierender Plattformanbieter eine Moderation von Inhalten vornehmen, um die Gesamtwerbeeinnahmen durch

das aus der Moderation resultierende höhere Nutzungsengagement zu steigern (Jiménez-Durán, 2022a, Jiménez-Durán, 2022b). So stellen Liu et al. (2022) fest, dass werbefinanzierte Plattformen eher dazu neigen, Inhalte mit laxen Regeln zu moderieren, um eine große und vielfältige Gruppe von Verbrauchern anzusprechen. Aus demselben Grund haben werbefinanzierte Plattformen keinen Anreiz, Systeme zu verwenden, die eine perfekte und fehlerfreie Moderation von Inhalten ermöglichen, d. h., die Inhalte nicht falsch kategorisieren oder falsch identifizieren, so dass sie auf der Plattform bleiben oder moderiert werden (Liu et al., 2022). Eine Plattform, die auf Werbeeinnahmen angewiesen ist, wird nicht von einer besseren Qualität der Moderationssysteme profitieren, da weniger präzise Systeme zu einer größeren und vielfältigeren Nutzerbasis und damit zu einer größerem Zielgruppe für Werbung führen (Yildirm & Zhang, 2022).

Auch Jiménez-Durán (2022a) kommt zu dem Schluss, dass Plattformen Inhalte nur in dem Maße moderieren, wie sie die Werbeeinnahmen erhöhen. Auch wenn das Verhalten der Werbekunden explizit modelliert wird, wie es Madio & Quinn (2023) tun, kommt ein ähnlichen Ergebnis heraus. Nach dem Modell diese Autoren ist die Platzierung von Werbeanzeigen auf Plattformen für Werbekunden weniger attraktiv, wenn diese in Verbindung mit schädlichen Inhalten angezeigt werden, da dies mit einem Reputationsverlust für die Kunden einhergehen kann. Die Autoren leiten die optimale Strategie einer Plattform ab, die lediglich ihre Gewinne steigern will: Wenn die Nutzer das Existenz schädlicher Inhalte stark bevorzugen oder wenn der marginale Reputationsverlust für Werbetreibende durch die Assoziation mit solchen Inhalten begrenzt ist, wäre es für die werbebasierte Plattform optimal, überhaupt nicht zu moderieren und schädlicher Inhalte zuzulassen.

Bei einem abonnementbasierten Einnahmemodell ist es noch weniger wahrscheinlich, dass Plattformen Inhalte moderieren. Doch wenn sie Inhalte moderieren, tun sie dies aggressiver als werbefinanzierte Plattformen. Optimiert wird hier vor allem über die Zahlungsbereitschaft und die Höhe der Abonnementgebühr (Liu et al., 2022; Yildirm & Zhang, 2022).

Wird das Ziel der Gewinnmaximierung als einzig entscheidender Faktor für die Moderation von Inhalten isoliert betrachtet, kann die Moderationspolitik der Plattformen und die Verbreitung illegaler und schädlicher Inhalte auf die Fähigkeit der Plattformen zurückgeführt werden, Einnahmen zu erzielen - unter Berücksichtigung von Aspekten wie Abonnementgebühren, Werbepolitik und Nutzerbasis. Je nach der jeweiligen Interessenslage der Werbekunden und/oder Nutzer kann dies zu einer laxeren oder strengeren Moderationspolitik führen. Dies betrifft auch die möglichen Maßnahmen, die gegen die Verbreitung bestimmter Inhalte eingesetzt werden. Da die Entfernung von Inhalten oder die Sperrung von Nutzern nicht das einzige Mittel ist, können Plattformen z. B. Empfehlungssysteme einsetzen, um zu verhindern, dass unerwünschte Inhalte bestimmten Verbrauchern gezeigt werden, oder Nutzer, die die Inhaltsmoderation als negativ empfinden, kompensieren werden, indem für sie z. B. die Intensität der Werbung verringern wird (siehe Kapitel 2.2).

2.1.2 Extrinsische Anreize: Öffentliche Kritik und regulatorische Bestimmungen

Plattformen sind nicht immun gegen Kritik und können dazu neigen, einen gewissen Kurswechsel vorzunehmen, wenn sie mit einer überwältigenden öffentlichen Reaktion auf ihren Moderationsansatz oder dessen Ergebnis konfrontiert werden (Hubley, 2022). Die ***öffentliche Kritik*** stellt somit den dritten Einflussfaktor dar, die sich auf die Bemühungen zur Moderation von Inhalten auswirkt, wie von Sander (2020) beschrieben.

Eine weiterer externer Einflussfaktor – und der vierte, der von Sander (2020) beschrieben wird –, der sich auf die Moderation von Inhalten auswirkt, ist die ***Gesetzgebung***. Staaten greifen in der Regel auf eine Kombination aus Gesetzen zurück, um die Moderationen von Inhalten auf Online-Plattformen zu beeinflussen und diese dazu zu bewegen, illegales oder unerwünschtes Nutzerverhalten zu verhindern oder einzuschränken. Hierzu gehören zum einen Gesetze, die festlegen, welche Inhalte beschränkt bzw. moderiert werden müssen, und zum anderen solche, die die Haftungsregel für die illegalen und unerwünschten Inhalte festlegen (Sander, 2020; Sartor et al., 2020). Bei ersteren werden Kategorien von Inhalten vorgeschrieben, die illegal und schädlich sind, während Gesetze zur Haftung die Bedingungen festlegen, unter denen Plattformen für solche von ihren Nutzern erzeugten Inhalte haftbar gemacht werden können (Sander, 2020). Insbesondere letzteres hat einen großen Einfluss auf die Anreize für Plattformen, Inhalte zu moderieren. Zu laxe Haftungsklauseln könnten nicht den gewünschten Effekt haben, illegale und schädliche Inhalte auf Plattformen einzudämmen. Regelungen, die eine übermäßige Haftung vorsehen, können jedoch Anreizstrukturen schaffen, die Plattformen dazu veranlassen, mehr Inhalte zu entfernen und zu beschränken als unbedingt notwendig. Angesichts des Risikos, sich schadensersatzpflichtig zu machen und Bußgeldern ausgesetzt zu sein, könnten sich Plattformen für eine übermäßige Entfernung von Inhalten entscheiden, um das Risiko von Sanktionen zu minimieren (Sartor et al., 2020; Heldt, 2018;; Castets-Renard, 2020; Digitale Gesellschaft e.V., 2020)). Dies trifft vor allem dann zu, wenn die Gewinne geringer ausfallen als die Bußgelder.

Der rechtliche Rahmen und seine Haftungsklauseln müssen daher angemessen gestaltet und angewandt werden, um zu verhindern, dass die Moderation in eine unerwünschte Richtung läuft.

2.2 Konsequenzen der Moderation und die Nutzung von Empfehlungssystemen

Wie in der Einleitung dieses Kapitels angedeutet, können Plattformanbieter auf eine Reihe von Maßnahmen zurückgreifen, um die Verbreitung unerwünschter Inhalte zu begrenzen. Das Löschen von Inhalten oder das Sperren eines Nutzer oder einer Gruppe von Nutzern von der Plattform ist die strengste Form der Inhaltsmoderation. Ersteres wird oft sofort bei Inhalten angewendet, die gegen bestehende Regeln und Gesetze

verstoßen. Letzteres, auch "**De-Plattforming**" genannt, wird in der Regel angewandt, wenn der Täter wiederholt gegen die bestehenden Regeln und Gesetze verstoßen hat.[10]

Aber nicht jede Art von Inhalt rechtfertigt eine Entfernung. Dazu gehören Inhalte, deren Verbreitung sozialen und individuellen Schaden anrichten kann, die aber technisch nicht verboten sind, weil sie weder illegal sind noch gegen die Regeln der Plattform selber verstoßen. In vielen Fällen ergreifen die Plattformen Maßnahmen, um die Sichtbarkeit solcher Inhalte zu verringern. Sie verwenden in der Regel automatisierte Systeme, um solche Inhalte zu identifizieren und diese in den algorithmischen Rankings und Empfehlungen herabzustufen. Dieser Ansatz wird oft als "**Shadow Banning**" bezeichnet, bei dem die Sichtbarkeit von Inhalten entweder ganz reduziert oder zumindest bestimmten Nutzergruppen nicht empfohlen wird (Gillespie, 2022) .

Reduzierungsstrategien wie das Shadow Banning können auch nützlich sein, wenn die Arten von Inhalten schwer zu klassifizieren sind (Gillespie, 2022). Einige der großen Plattformanbieter haben bereits Reduzierungsstrategien für grenzwertige Inhalte eingeführt. Meta kündigte 2018 seine Reduktionspolitik für die beiden Plattformen Facebook und Instagram an. Im Jahr 2019 kündigte auch YouTube seine Richtlinien zur Reduzierung an. Auch Twitter, LinkedIn, TikTok sowie Tumblr, Reddit und andere verfügen über ähnliche Strategien zur Eindämmung von Inhalten (Gillespie, 2022) .

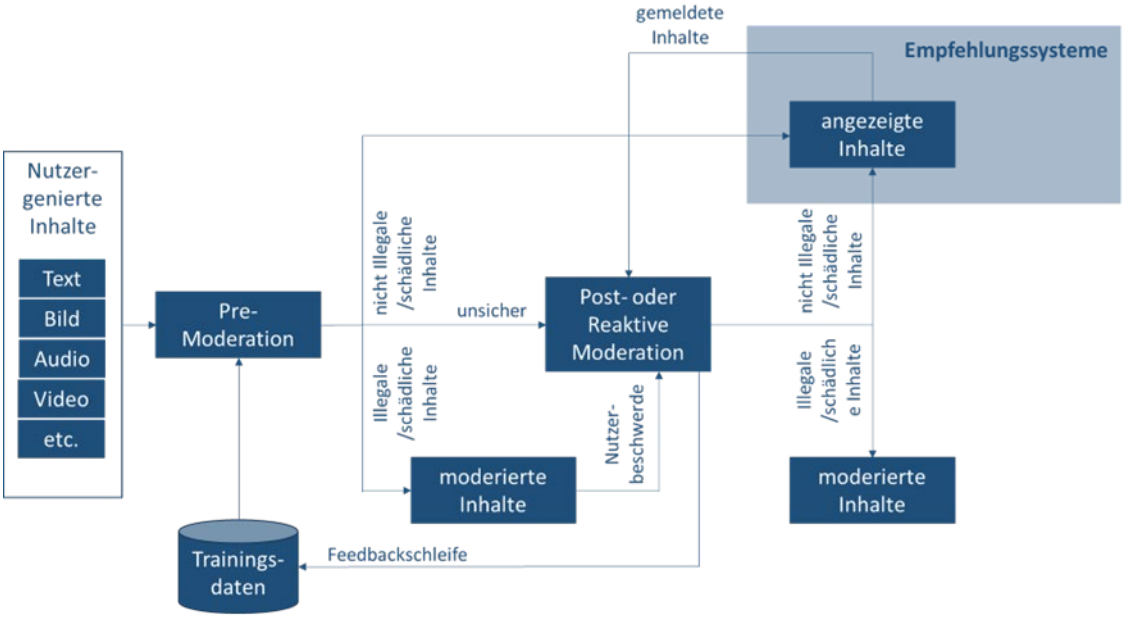
Das ist jedoch nicht die einzige Möglichkeit, wie Empfehlungssysteme bei der Moderation von Inhalten eingesetzt werden können. Empfehlungssysteme machen sich im Allgemeinen die Vorlieben und Neigungen der Nutzer zunutze und sind so konzipiert, dass sie Inhalten, die diesen Bedürfnissen entsprechen, Vorrang einräumen und andere Inhalte herabstufen, um Zufriedenheit und Engagement zu fördern. Indem sie bestimmte Arten von Inhalten vorschlagen und andere ausblenden, können Plattformen die Erfahrung jedes Nutzers individuell gestalten (Llansó et al., 2020). Durch die Nutzung dieser Eigenschaft können Plattformen nicht nur die Interaktion mit der Plattform fördern, sondern diese auch verringern. Da Nutzer von Plattformen z.B. einen Nutzenverlust erleiden, wenn ihnen zu viel Werbung angezeigt wird, kann eine Plattform die Anzeigenlast für bestimmte Verbrauchertypen erhöhen, um deren Engagement zu verringern und so die Menge an gutartigen, schädlichen oder illegalen Inhalten zu kontrollieren. Theoretisch kann eine Plattform die gleiche Prävalenz von unerwünschten Inhalten erreichen, indem sie die Werbebelastung variiert, unabhängig davon, ob sie streng oder lax moderiert (Jiménez-Durán, 2022b).

10 Der Zeitpunkt, zu dem diese Konsequenz genutzt wird, ist von Plattform zu Plattform unterschiedlich.

3 Ein Überblick über die Herausforderungen der derzeitigen Instrumente zur Moderation von Inhalten

Obwohl jede Plattform ihre eigenen Mittel und Methoden zur Moderation von Inhalten anwendet, folgen sie alle einem ähnlichen Schema (Hubley, 2022). In der Regel handelt es sich bei der Moderation von Inhalten um einen mehrstufigen soziotechnischen Prozess, an dem in der Regel sowohl automatisierte Systeme, die u. a. Deep-Learning- und Machine-Learning-Algorithmen verwenden, als auch menschliche Moderatoren beteiligt sind (Cambridge Consultants, 2019; Sartor et al., 2020). Abbildung 3-1 zeigt eine schematische Darstellung dieses mehrstufigen soziotechnischen Prozesses der Moderation von Inhalten und dessen Beziehung zum Empfehlungssystem.

Abbildung 3-1: Moderationsprozess



Quelle: Eigene Darstellung basierend auf Cambridge Consultants (2019) und Sartor et al. (2020).

Heutzutage ist die Moderation von Inhalten vor der Veröffentlichung, die gewöhnlich als ***Ex-ante-Moderation*** bezeichnet wird, allgegenwärtig. Im Wesentlichen werden Inhalte, die ein Nutzer hochladen möchte, geprüft, bevor sie veröffentlicht und für andere Nutzer sichtbar gemacht werden. Obwohl auch in dieser Phase eine manuelle Überprüfung durch menschliche Moderatoren möglich ist, übernehmen in der Regel automatisierte Systeme diese Aufgabe - vor allem bei Plattformen, die täglich von einer großen Zahl von Nutzern verwendet werden. Der Einsatz von automatisierten Systemen vor der Veröffentlichung kann besonders effektiv sein, um unzulässige Inhalte in großem Umfang zu identifizieren und zu moderieren (Gorwa et al., 2020).

Inhalte, die nicht eindeutig als gutartig oder unerwünscht eingestuft werden können, werden an die **proaktive Ex-post-Moderation** weitergeleitet, wo sie von der Plattform proaktiv erneut geprüft werden. Menschliche Moderatoren sind in dieser Stufe in der Regel an der Entscheidung beteiligt, ob diese Inhalte für die jeweilige Plattform geeignet sind und veröffentlicht werden können oder nicht. Sie arbeiten normalerweise mit detaillierten Flussdiagrammen und Entscheidungsbäumen, die sie bei der Bewertung von Inhalten befolgen müssen (Reuber & Fischer, 2022). Ein Teil der von den Ex-ante-Moderation als unsicher eingestuften Inhalte wird zunächst veröffentlicht, aber zur manuellen Überprüfung in eine Warteschlange gestellt, während andere unsichere Inhalte erst nach einer manuellen Überprüfung veröffentlicht werden. Dieser Prozess hängt stark von der Plattform und der Art des Inhalts ab.

Ebenso können die Nutzer selbst eine Post-Moderation auslösen, in diesem Fall handelt es sich um eine ***reaktive Ex-Post-Moderation*** von Inhalten. Das bedeutet, dass Personen, deren Inhalte von der Moderation betroffen sind, eine Beschwerde einreichen können, während andere Nutzer einer Plattform bereits veröffentlichte Inhalte als problematisch oder regelwidrig beanstanden können. Auch diese Fälle werden häufig von menschlichen Moderatoren geprüft. Sie können aber durch automatisierte Systeme unterstützt werden. Diese helfen ihnen dabei, Inhalte zu kategorisieren und zu priorisieren, um den Überprüfungsprozesses effizienter zu machen (Klonick, 2018; Cambridge Consultants, 2019; De Gregorio, 2020). Automatisierte Systeme könnten auch eingesetzt werden, um menschliche Moderatoren davor zu schützen, verstörende Inhalte zu sehen, um so die nachweislich hohe psychische Belastung der Arbeit zu mildern. Dazu gehören Verfahren, die Elemente in auditiven und visuellen Inhalten modifizieren - z. B. durch Unschärfe, Grauskalierung und/oder Stummschaltung – oder wichtige Informationen bereitstellen, bevor den Moderatoren die jeweiligen Inhalte angezeigt werden. (Cambridge Consultants, 2019; Innodata, 2023). Die Tätigkeit der Moderatoren hat Auswirkungen auf die Psyche. In den Medien gibt es zahlreiche Berichte über Moderatoren, die aufgrund ihrer Tätigkeit an einer posttraumatischen Belastungsstörung (PTSD) oder anderen psychischen Problemen leiden (Whittaker, 2020; Newton, 2029).

Wird der betreffende Inhalt letztlich als unangemessen und als Verstoß gegen Gesetze oder Regeln der Plattform eingestuft, gibt es für die jeweilige Plattform im Wesentlichen drei mögliche Konsequenzen[11]:

- Der Inhalt kann „versteckt“ werden (Shadow Banning).
- Der Inhalt kann, je nach Schwere des Verstoßes, gelöscht werden.
- In schwerwiegenden Fällen kann der Nutzer, der den Inhalt übermittelt hat, von der Plattform entfernt oder gesperrt werden.

11 In Kapitel 2 werden noch ein paar zusätzliche Konsequenzen aufgezählt, die ein Plattformanbieter ziehen kann, um Nutzer zu disziplinieren. Jedoch stellen diese drei, die Konsequenzen bei der direkten Moderation von Inhalten dar.

Der Moderationsprozess an sich ist jedoch nicht frei von Kritik und Bedenken, da er einige Mängel aufweist, sowohl in Bezug auf die Gesamtausführung des Prozesses seitens der Plattform als auch in Bezug auf die tatsächlichen Fähigkeiten der automatisierten Systeme und der menschlichen Moderatoren, Inhalte zu identifizieren und zu klassifizieren (Singh, 2019; Sander, 2020). In den folgenden beiden Abschnitten werden die vier am häufigsten genannten Einschränkungen des heutigen Prozesses der Moderation von Inhalten erörtert. Die folgenden Abschnitte erheben nicht den Anspruch, eine vollständige Beschreibung aller Einschränkungen darzustellen. Vielmehr sollen sie einen Überblick über einige der häufigsten Probleme geben.

3.1 Genauigkeit und Zuverlässigkeit

Ein ständiges Problem im Zusammenhang mit dem Moderationsprozess ist die Fähigkeit der verwendeten Werkzeuge, Inhalte genau und zuverlässig zu erkennen und zu moderieren. Die Verfahren für die Moderation von Online-Inhalten sind vielfältig und unterscheiden sich in Bezug auf die Art der Medieninhalte, die von der Moderation erfasst werden sollen. Plattformen können automatisierte Systeme einsetzen, die relativ einfachen Algorithmen nutzen, wie beim ***Hash-Matching*** und ***Word-Matching,*** oder auf sehr komplexe Algorithmen zurückgreifen, die ganze Szenarien (**scene *understanding***) interpretieren und Handlungen erkennen (***action recognition***) (Cambridge Consultants, 2019). Insgesamt unterscheiden Gorwa et al. (2020), Kamara et al. (2021) und Shenkman et al. (2021) zwischen Algorithmen, die einfach Inhalte mit Inhalten abgleichen, die bereits als illegal, schädlich oder anderweitig ungeeignet identifiziert wurden, und solchen, die darauf abzielen, zuvor unbekannte Inhalte zu klassifizieren.

In den folgenden Abschnitten werden ausgewählte Algorithmen und dazugehörige Systeme, die zur Moderation verwendet werden können dargestellt, um die Probleme und Vorteile des Einsatzes dieser zu veranschaulichen. Es ist zu beachten, dass es neben den hier beschriebenen Algorithmen und Systeme noch eine Vielzahl anderer gibt, die für die Moderation eingesetzt werden können.

3.1.1 Matching-Algorithmen

Matching-Algorithmen gelten als weniger komplex und ihre Verwendung ist ein einfacher Ansatz zur Moderation von Inhalten (Shenkman et al., 2021). Diese Algorithmen beinhalten in der Regel entweder die Verwendung von Zeichenketten (sogenannte Strings) für Textinhalte oder die Umwandlung von Inhalten mit auditiven und/oder visuellen Elementen in eindeutige "Hashes" oder Sequenzen von Daten, die den zugrunde liegenden Inhalt identifizieren können (Gorwa et al., 2020). Sowohl String- als auch Hash-Matching-Algorithmen prüfen, ob ein Textblock Wörter oder Phrasen enthält, die in Datensätzen mit unerwünschten textlichen Inhalten gespeichert sind, oder ob die generierten Hashes mit denen aus einer Datenbank mit auditiven und/oder visuellen Inhalten übereinstimmen, die bereits zuvor identifiziert wurden. Diese beiden Arten von Matching-Algorithmen

können in einer Vielzahl von Szenarien und für verschiedene Arten von Inhalten angewendet werden, z. B. für Inhalte mit Urheberrechtsverletzungen, missbräuchliche, gewalttätige oder extremistische Inhalte, sexuelle Inhalte, Hassreden und Belästigungen.

So beschreibt Sartor et al. (2020), dass in Bezug auf Textinhalte beleidigende, rassistische und sexuelle Inhalte erkannt werden können, indem hochgeladene Dokumente mit Listen von Schlüsselwörtern abgeglichen werden, die häufig in solchen Nachrichten vorkommen, wie Beleidigungen oder rassistische Ausdrücke. Urheberrechtsverletzungen in Textdokumenten können hingegen durch den Abgleich hochgeladener Dokumente mit einer Datenbank urheberrechtlich geschützter Textwerke erkannt werden (Sartor et al., 2020). Diese Art des Matching hat jedoch mehrere Nachteile. Der erste Nachteil besteht darin, dass die Matching-Algorithmen den Kontext der Inhalte nicht analysieren. Selbst wenn ein Treffer beim String-Matching erzielt wird, kann der Inhalt je nach Kontext, in dem er verwendet wird, akzeptabel sein. Die Moderation von Inhalten ohne Kontext kann zu so genanntem Over-Blocking führen, bei dem mehr Inhalte moderiert werden als notwendig. (Gorwa et al., 2020). Zudem können Nutzer das System leicht umgehen, indem sie einfach die Schreibweise ändern (Cambridge Consultants, 2019). Ähnliche Probleme treten bei auditiven und/oder visuellen Inhalten auf, wenn Hashes abgeglichen werden.

Generell können zwei Arten von Hash-Algorithmen unterschieden werden – das kryptografische und nicht-kryptografische Hashing. Beim kryptografischen Hashing wird ein zufälliger Hash-Wert erzeugt, der extrem empfindlich auf Änderungen reagiert. Wenn beispielsweise auch nur der Farbton eines Pixels eines Fotos oder die Länge oder das Kodierungsformat eines Videos geändert wird, wird ein anderer kryptografischer Hash erzeugt (Singh, 2019; Shenkman et al., 2021). Da selbst die kleinste Änderung in den Daten zu einem anderen Hashwert führt, können mit dieser Methode nur exakt übereinstimmende Inhalte identifiziert werden. Der Vorteil dieser Funktion ist, dass es äußerst unwahrscheinlich ist, dass ein Inhalt fälschlicherweise als ein anderer Inhalt identifiziert wird (Farid, 2021). Wie beim Abgleich von Wörtern können Nutzer das System umgehen, indem sie nur geringfügige Änderungen am Originalmaterial vornehmen.

Beim nicht-kryptografischen Hashing geht es nicht darum exakte Übereinstimmungen zu finden. Im Mittelpunkt steht die Identifizierung ähnlicher Inhalte. Damit wird der Nachteil des kryptografischen Hashings überwunden, dass nur exakte inhaltliche Übereinstimmungen gefunden werden können (Gorwa et al., 2020; Shenkman et al., 2021). Diese Methode birgt jedoch auch das Potenzial, denselben Hash-Wert für zwei Inhalte zu finden, die sich für einen menschlichen Betrachter unterscheiden, aber aufgrund von Merkmalen, die der Algorithmus berücksichtigt, ähnlich sein können, was dazu führt, dass gutartige Inhalte fälschlicherweise als illegal, schädlich oder anderweitig unangemessen eingestuft werden, oder umgekehrt (Gorwa et al., 2020; Ofcom, 2022).

Zwei bekannte Systeme, die ein nicht-kryptografisches Hash-Matching-Modell verwenden, sind PhotoDNA, das von Microsoft entwickelt wurde und inzwischen von vielen Unternehmen und Strafverfolgungsbehörden genutzt wird, und ContentID, das von Google entwickelt wurde. PhotoDNA generiert digitale Hashes und gleicht sie mit Datenbanken

ab, die Hashes von illegalem Material über sexuellen Kindesmissbrauch (CSAM) enthalten, wie z. B. die Hash-Datenbank des National Centre for Missing and Exploited Children (Singh, 2019; Gorwa et al., 2020; Farid, 2021).[12] PhotoDNA ist nachweislich resistent gegen einige verlustbehaftete Verarbeitungsschritte wie Kompression, aber weniger resistent gegen z. B. das Beschneiden einer Fotodatei. Dies veranlasst Steinebach (2023) zu der Schlussfolgerung, dass der Hash sehr gute Ergebnisse liefert, jedoch nicht die extremen Leistungen bietet, die manchmal in der öffentlichen Diskussion genannt werden (Steinebach, 2023).

ContentID dient in erster Linie dem Schutz der Eigentumsrechte der Urheber von Inhalten, indem es den Nutzern ermöglicht, digitale Hashes für ihre Inhalte zu erstellen. Die Hashes werden in einer Datenbank gespeichert und verwendet, um festzustellen, ob derselbe Inhalt erneut hochgeladen wurde (Singh 2019; Cambridge Consultants, 2019; Digitale Gesellschaft e.V., 2020; Dias Oliva, 2020). Das „Fair Use“-Prinzip bei urheberrechtlich geschütztem Material in den USA zum Beispiel, macht eine fallspezifisch Betrachtung erforderlich und erfordert ein gewisses Maß an Subjektivität bei der Durchsetzung des Urheberrechts, das einfache Matching-Algorithmen nicht bieten können. Daher kann ContentID Fälle, bei denen diese Prinzip theoretisch greift, nicht eindeutig identifizieren (Perel & Elkin-Koren, 2016; Dias Oliva, 2020). Dies wiederum kann je nach Kontext zu einer übermäßigen Sperrung von Material führen, das nicht unbedingt gegen das Urheberrecht verstößt.

Probleme können jedoch nicht nur durch die Verwendung von Matching-Algorithmen entstehen, sondern auch durch die Datenbanken, die zum Abgleich hochgeladener Inhalte verwendet werden. Besonders deutlich wird dies in der Debatte um die Erkennung terroristischer Inhalte. Die Nicht-Regierungsorganisation Global Internet Forum to Counter Terrorism (GIFCT), die 2017 von Facebook, Microsoft, Twitter und YouTube gegründet wurde und der sich inzwischen viele weitere Unternehmen angeschlossen haben, unterhält eine gemeinsame, branchenweite Hash-Datenbank für terroristische Inhalte, um zu verhindern, dass Terroristen und gewaltbereite Extremisten ihre Ideen über digitale Dienste verbreiten.[13] Die teilnehmenden Unternehmen aktualisieren die Datenbanken laufend. (Gorwa et al., 2020). Die GIFCT-Definition von Terrorismus ist jedoch nicht eindeutig und basiert lediglich auf den Richtlinien der Vereinten Nationen. Jedes Mitgliedsunternehmen definiert und dokumentiert in seinen eigenen Geschäftsbedingungen, was es unter "terroristischen Inhalten" versteht (Heller, 2019). Darüber hinaus konzentrieren einige Anbieter ihre Bemühungen zur Moderation von Inhalten auf bestimmte extremistische Gruppen (Singh, 2019; Gorwa et al., 2020). Dies kann zu einer Verzerrung der in den Datenbanken vertretenen Inhalte führen, die nicht für alle Arten von Plattformen oder Regionen gleichermaßen gilt (Heller, 2019).

---

12 Andere Hash-Technologien zur Erkennung von CSAM sind CSAI Match von YouTube, Content Safety API von Google, NeuralHash von Apple und PDQ von Meta (Gorwa et al., 2020; Shenkman et al., 2021; Ofcom, 2022).

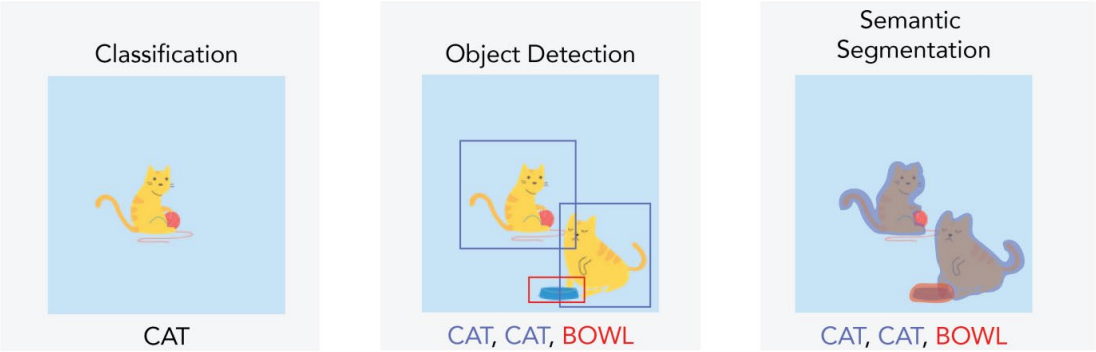
13 Siehe https://gifct.org/ [letzter Zugriff: 29.03.2023].

3.1.2 Prädiktive Algorithmen

Im Gegensatz zu den bisher beschriebenen Algorithmen bewerten prädiktive Algorithmen neue Inhalte, für die es keine vorherige Klassifizierung gibt. Diese Algorithmen basieren auf den Konzepten des Deep- und Machine-Learning. Ein solcher Algorithmus ist z.B. das ***Natural Language Processing (NLP)***. NLP wird vor allem dann eingesetzt, wenn das Erkennen einzelner Wortmuster, wie es bei einfachen Matching-Modell der Fall wäre, nicht ausreicht, um Inhalte eindeutig als gutartig oder unangemessen zu klassifizieren (Sartor et al., 2020). Ziel von NLP ist es, die Sprache so zu verstehen und zu verarbeiten, wie sie von Menschen verwendet wird, indem linguistische Studien genutzt werden und die Syntax, Semantik und Pragmatik von Ausdrücken in der natürlichen Sprache untersucht werden (Sartor et al, 2020).

Für visuelle Inhalte umfassen prädiktive Algorithmen unter anderem Algorithmen aus dem Bereich der ***Computer Vision*** wie ***Klassifizierung (classification)***, ***Objekterkennung (object detection)***, ***Segmentierung (segmentation)***, ***Szenenverständnis (scene understanding)*** und ***Handlungserkennung (action recognition)***. Bei den ersten drei geht es darum, ein oder mehrere Objekte oder Merkmale wie Nacktheit, Gewalt, Waffen, Drogen, Alkohol etc. in einem Bild zu klassifizieren oder sie insgesamt mit unterschiedlicher Präzision zu lokalisieren (Shenkman et al., 2021).

Abbildung 3-2: Beispiel - Klassifizierung, Objekterkennung und Segmentierung



Quelle: Ausschnitt aus Walia (2022).

Das ***Szeneverständnis*** ist sogar noch komplexer, da es eine Reihe verschiedener Algorithmen verwendet, darunter Algorithmen zur Klassifizierung, Objekterkennung und Segmentierung, aber auch zur Tiefenschätzung, Posenschätzung und/oder Stimmungsanalyse (Shenkman et al., 2021). Zusammen werden diese Algorithmen nicht nur zur Identifizierung der einzelnen Figuren/Objekte in einem Bild oder Video verwendet, sondern auch zu deren Analyse im Kontext ihrer Beziehung zu anderen Objekten im Bild/Video (Llansó et al., 2020). Algorithmen zur ***Handlungserkennung*** analysieren hingegen physische Anhaltspunkte, um zu erkennen, welche Handlungen in Bildern/Videos ausgeführt

werden (Shenkman et al., 2021). Algorithmen aus dem Bereich der ***Computer Vision*** werden in kommerziellen Content-Moderationssystemen wie Amazon Rekognition Content Moderation, Google Cloud's Visionoder Azure AI Content Safety verwendet (Shenkman et al., 2021).

Trotz ihres Anspruchs haben prädiktive Algorithmen jedoch eine Reihe von Einschränkungen, die ihre Anwendung erschweren. Grundsätzlich hängt die Genauigkeit dieser Algorithmen bei der Identifizierung von Inhalten in hohem Maße davon ab, wie sie konzipiert und trainiert werden – einschließlich der Qualität der zu diesem Zweck verwendeten Daten (Singh, 2019). So hat sich beispielsweise wiederholt gezeigt, dass sie in Situationen versagen, auf die sie während der Entwurfs- und Trainingsphase nie gestoßen sind, und dass sie Schwierigkeiten haben, mit Änderungen umzugehen, die sogar auf natürliche Weise auftreten können (Shenkman et al., 2021). Qiu & Yuille (2016) zeigen zum Beispiel, wie die Genauigkeit eines Algorithmus, der für die Erkennung von Möbeln wie Sofas trainiert wurde, variiert, wenn er Bilder von Sofas aus verschiedenen Blickwinkeln untersuchen muss. Die Genauigkeit, mit der das Sofa erkannt wird, ist bei Blickwinkeln geringer, die im Trainingsdatensatz unterrepräsentiert waren. Positiv ausgedrückt: Der Algorithmus bevorzugt Sofas aus Blickwinkeln, die im Benchmark-Datensatz gut vertreten sind, und ist in der Lage, diese genauer zu erkennen (Yuille & Liu, 2019). Das Versagen von Algorithmen mit realen Bildern und natürlicher Diversifikation wird auch durch die Analyse der Leistung realer kommerzieller Content-Moderationssysteme dokumentiert. Aldahoul et al. (2023) untersuchten die Content-Moderationssysteme von AWS und Microsoft Azure, die Algorithmen aus der Computer Vision einbeziehen. Die Autoren wendeten die Systeme auf drei Datensätze an, die herkömmliche Fotos, Cartoon-Bilder und Skizzenbilder umfassen, um ihre Leistung bei der Erkennung von sexuellen Aktivitäten, Nacktheit und Pornografie zu bewerten. Die Autoren fanden heraus, dass Azure- und AWS-Moderatoren bei Cartoon-Bildern mit einer durchschnittlichen Genauigkeit von 97 % gleich gut abschnitten. AWS ist besser bei der Erkennung von Nacktheit und Pornografie in Skizzenbildern und erreicht eine durchschnittliche Genauigkeit von 93 %. Microsoft Azure erkennt Nacktheit und Pornografie in herkömmlichen Fotos mit einer durchschnittlichen Genauigkeit von 87 % und ist damit besser als AWS. Beide Systeme stehen also vor der größten Herausforderung, wenn es um die Erkennung und Klassifizierung echter Bilder geht.

Ein ähnliches Problem ergibt sich, wenn Algorithmen voreingenommene Entscheidungen in Bezug auf einzelne soziodemografische Gruppen treffen. Buolamwini und Gebru (2018) führen Experimente mit kommerziellen Gesichtserkennungssystemen durch und zeigen, dass die Gesichter von Menschen mit einem helleren Hautton im Allgemeinen von diesem System besser erkannt werden als die von Menschen mit einem dunkleren Hautton. Snow (2018) berichtet über einen Test, bei dem die Gesichtserkennungssoftware Amazon Rekognition verwendet wurde, um Bilder von Mitgliedern des US-Kongresses mit einer Datenbank von Verbrecherfotos zu vergleichen. Nicht-weiße Mitglieder wurden überproportional häufig als Kriminelle identifiziert. Kayser-Bril (2020)

berichtete über ein Experiment von AlgorithmWatch, das zeigte, dass eine (inzwischen verbesserte und aktualisierte) Version von Googles Cloud Vision ein Thermometer in einem Bild als "Waffe" identifizierte, wenn es von einer dunkelhäutigen Hand gehalten wurde, während das Thermometer als "elektronisches Gerät" identifiziert wurde, wenn es in einem ähnlichen Bild von einer hellhäutigen Hand gehalten wurde. Blodgett und O'Connor (2017) wendeten vier verschiedene NLP-Systeme von bekannten Herstellern auf Nachrichten an, die im sozialen Netzwerk Twitter gepostet wurden. Die Autoren stellten Unterschiede in der Genauigkeit der Texterkennung fest, wobei Textnachrichten, die von weiblichen oder afroamerikanischen Autoren geschrieben wurden, weniger genau waren als solche, die von weißen männlichen Autoren verfasst wurden. Auch dies kann passieren, wenn bestimmte Gruppen im Datensatz unterrepräsentiert sind. Selbst wenn der Trainingsdatensatz vollständig oder sogar repräsentativ ist, aber soziale Ungleichheiten oder Diskriminierung in der Vergangenheit bereits in den Daten vorhanden sind und in das Training des automatisierten Systems einfließen, kann es zu verzerrten Ergebnissen führen. Damit steigt auch das Risiko einer weiteren Marginalisierung von Minderheiten durch unverhältnismäßige Fehlinterpretation von Inhalten von und über sie (Duarte et al., 2017; Taş et al., 2022). Es besteht die Gefahr, dass solche verzerrten Entscheidungen systemimmanent werden, da die Entscheidungen als neue Daten wieder in die Algorithmen eingespeist werden (Cofone, 2019; Tas & Wiewiorra, 2022).

Ein weiteres Einfallstor für Verzerrungen ist die Programmierphase von Algorithmen selbst. Studien zeigen, dass sich Vorurteile von Programmierern auf automatisierte Systeme übertragen können. Selbst ein theoretisch unvoreingenommener Programmierer ist möglicherweise nicht in der Lage, alle möglichen eintretenden Verzerrungen vollständig zu berücksichtigen und zu korrigieren, da er oder sie sich nur auf Aspekte konzentrieren kann, die ihm oder ihr bekannt und vertraut sind, was dazu führt, dass automatisierte Systeme falsch kalibriert sind. Es hat sich zum Beispiel gezeigt, dass unterschiedliche Programmierteams dem entgegenwirken können (Tas & Wiewiorra, 2022).

Auch die korrekte Anwendung kann entscheidend sein. So kann es zu verzerrten Ergebnissen kommen, wenn ein automatisiertes System, das für eine bestimmte Aufgabe entwickelt wurde, in einem ungewohnten Kontext angewendet wird. Viele Algorithmen zur Handlungserkennung sind beispielsweise hochspezialisiert. Diese Instrumente können nur bestimmte Faktoren berücksichtigen oder sie gleichzeitig abdecken (Shenkman et al., 2021). Das Gleiche gilt für NLP. NLP ist domänenspezifisch, d. h. es kann sich nur auf eine bestimmte Art von Inhalten konzentrieren, für die es ausgebildet wurde. Außerdem erfordert NLP klare, konsistente Definitionen der Art der zu identifizierenden Inhalte (Duarte et al., 2017; Singh, 2019; Llansó et al., 2020).

Auch der Kontext ist für prädiktive Algorithmen immer noch schwer zu erkennen. Die Interpretation von nutzergenerierten Inhalten ist in der Regel mit Mehrdeutigkeit und subjektiven Einschätzungen behaftet. So kann beispielsweise derselbe Inhalt, der von einer Person in einem bestimmten Kontext geteilt wird, in einem anderen Kontext eine völlig andere Bedeutung haben. Der Zweck des Inhalts muss ebenfalls berücksichtigt werden,

einschließlich seines künstlerischen, wissenschaftlichen und pädagogischen Werts. So sind Algorithmen aus der Computer Vision zwar in der Lage, Nacktheit in einem Bild zu erkennen, aber nicht in der Lage zu beurteilen, ob diese Nacktheit im Kontext des künstlerischen Ausdrucks oder des Missbrauchs steht, so Shenkman et al. (2021). Ähnlich ist heute NLP noch immer nicht in der Lage, alle Nuancen und kontextuellen Elemente der menschlichen Sprache zu erfassen. Dieses Problem wird noch verstärkt, wenn NLP-Systeme, die für ein bestimmtes Objektiv entwickelt wurden, für unterschiedliche Arten von Inhalten, Sprachen und Kontexten verwendet werden (Duarte et al., 2017; Singh, 2019; Llansó et al., 2020).

Dies sind nur einige der Einschränkungen von automatisierten Systemen und ihrer Algorithmen, die zu einer falschen Klassifizierung von Inhalten führen können. Diese befinden sich jedoch noch in der Entwicklungsphase und können aktuell noch nicht ihr volles Potenzial ausschöpfen. Laut dem GIFCT (2021) ist beispielsweise die Verfügbarkeit von Systemen, die Algorithmen zum Verstehen von Szenen verwenden, gering. Diese befindet sich noch im Entwicklungsstadium. Das Gleiche gilt für die Handlungserkennung oder Systeme zur Verarbeitung natürlicher Sprache (Shenkman et al., 2021). Infolgedessen sind Fehler möglich. Dies kann sowohl zu falsch-negativen als auch zu falsch-positiven Klassifizierungen führen, bei denen Inhalte, die hätten entfernt werden müssen, weiterhin für die Öffentlichkeit sichtbar sind, oder bei denen gutartige Inhalte versehentlich entfernt werden.

Es muss jedoch darauf hingewiesen werden, das diese Algorithmen und dazugehörigen Systeme sich stetig weiter entwickeln und dadurch wahrscheinlich heute schon besser als gestern funktionieren. Dennoch ist festzuhalten, dass selbst wenn Genauigkeitsraten von 99,5 % zu erreichen werden können, diese immer noch zu falsch positiven oder negativen Ergebnissen hervorrufen, von denen Millionen von Menschen betroffen wären (Heller, 2019).

Tabelle 3-1: Einstufung von Inhalten

	Klassifizierung als nicht schädlicher / illegaler Inhalt	**Klassifizierung als schädlicher / illegaler Inhalt**
Schädliche/ illegale Inhalte	Falsche Klassifizierung | FN (illegaler/schädlicher Inhalt wird nicht erkannt/nicht moderiert)	Richtige Klassifizierung | TP (illegaler/schädlicher Inhalt wird erkannt/moderiert)
Nicht schädliche /illegal Inhalte	Richtige Klassifizierung | TN (nicht illegaler/schädlicher Inhalt wird erkannt/nicht moderiert)	Falsche Klassifizierung | FP (nicht illegaler/schädlicher Inhalt wird nicht erkannt/moderiert)

Quelle: Eigene Darstellung. FN=False-Negative; TP=True-Positive; TN=True-Negative; FP=False-Positive.

Ein unbestreitbarer Vorteil automatischer Inhaltsmoderationssysteme ist ihre Skalierbarkeit. Diese Systeme sind in der Lage, eine große Menge an Inhalten in bemerkenswert kurzer Zeit zu bewerten. Die gleiche Effizienz könnte mit der Moderation durch Menschen nicht erreicht werden. Statista (2023) berichtet, dass jede Minute 1,7 Millionen Inhalte auf

Facebook gepostet und etwa 500 Stunden Video auf YouTube hochgeladen werden. Diese Zahlen sind nur ein Bruchteil der Inhalte, die von Verbrauchern auf verschiedenen Plattformen hochgeladen werden. Die Menge der Inhalte, die menschlicher Moderatoren in einem bestimmten Zeitraum moderieren können, ist aufgrund ihrer kognitiven Fähigkeiten viel begrenzter. Heldt (2018) berichtet beispielsweise von einer offiziellen Erklärung von YouTube, dass durch den Einsatz automatisierter Inhaltsmoderationssysteme fünfmal mehr Videos mit terroristischen Inhalten vor dem Hochladen gelöscht werden konnten als ohne. Das Unternehmen schätzt, dass diese Systeme 180.000 menschliche Moderatoren mit jeweils 40 Stunden pro Woche ersetzen könnten (Heldt, 2018). Es ist also nicht nur eine Frage der Zeit, sondern auch der Kosten, wenn es um den Einsatz von menschliche Moderatoren geht.

Ein weiterer Vorteil automatisierter Systeme ist die Konsistenz ihrer Entscheidungen, da die Regeln und Standards für die Moderation von Inhalten fest in ihnen verankert sind. Das bedeutet aber auch, dass sie weniger sensibel und anpassungsfähig an den jeweiligen Kontext sind, in dem der Inhalt steht (Jhaver et al., 2019; Jiang et al., 2023).

3.1.3 Menschliche Moderation

Obwohl die Verwendung von automatisierten System hinsichtlich der Menge der zu moderierenden Inhalte effektiv ist, kann es dazu führen, dass Inhalte nicht immer korrekt klassifiziert werden. Aus diesem Grund wird oft argumentiert, dass menschliche Moderatoren weiterhin notwendig sind. Ihnen wird zugeschrieben, dass sie in der Lage sind, verschiedene Nuancen in der Sprache zu erkennen, wie z. B. subtile Anspielungen und den Gesamtzusammenhang, in den der Inhalt gestellt wird, die von automatisierten Systemen nicht richtig oder nur unvollständig berücksichtigt werden (Singh, 2019; Cambridge Consultans, 2019; De Streel et al., 2022).

In der Realität haben die Moderatoren jedoch in der Regel etwa eine Minute Zeit, um die Inhalte zu bewerten. Da die Moderatoren also auch große Mengen an Inhalten relativ schnell prüfen müssen, können ihre Entscheidungen inkonsistent und fehlerhaft sein. Trotz der Listen und Richtlinien für die manuelle Moderation, die die Moderatoren befolgen müssen, unterliegen die Inhalte den Interpretationen der Moderatoren und werden möglicherweise nicht von jedem einzelnen Moderator einheitlich behandelt. Außerdem sind menschliche Moderatoren nicht frei von Voreingenommenheit und Vorurteilen, die die Bewertung von Inhalten beeinflussen. Wenn Moderatoren mit der Moderation von Inhalten aus fremden Regionen und Kulturkreisen betraut werden, sind sie zudem nur begrenzt in der Lage, sprachliche und inhaltliche Nuancen zu verstehen. [14]

Es ist daher nicht auszuschließen, dass bei der menschlichen Moderation Fehler auftreten. Das zeigt zum Beispiel eine Untersuchung von ProPublica. Tobin et al. (2017)

14 Dieser Abschnitt basiert auf Wilson & Land (2021), Singh, (2019). De Streel et al. (2022), Gillespie (2020), Cambridge Consultants (2019), De Gregorio (2020), Barrett (2020), Reuber & Fischer (2022).

analysierten 900 Beiträge, die von Nutzern als potenzielle Verstöße gegen die Community-Richtlinien von Facebook markiert worden waren, und wandten sich mit einer Stichprobe von 49 Vorgängen an das Unternehmen, um eine Erklärung zu erhalten. Das Unternehmen gab zu, dass die Moderatoren bei der Überprüfung von fast der Hälfte der Beiträge Fehler gemacht hatten.

Des Weiteren können die Nutzer am Moderationsprozess beteiligt werden. Die Plattformen ermöglichen den Nutzern, Inhalte zu melden, die sie für illegal oder unangemessen halten. Der Vorteil der Einbeziehung der Nutzer in die Inhaltsmoderation besteht darin, dass die Nutzer das Gefühl haben, die Plattform mitzugestalten. Ein Nachteil der Einbeziehung der Benutzer in die Inhaltsmoderation besteht darin, dass die Benutzer enttäuscht sein können, wenn ihre Beschwerden nicht zu den erhofften Moderationsergebniss führen. Eine Konsequenz wäre, dass sie sich dadurch von der Plattform abwenden (Reuber & Fischer, 2022).

3.2 Transparenz und Verantwortlichkeit

Abgesehen von der Tatsache, dass die Moderation von Inhalten nicht immer fehlerfrei ist und zu falsch positiven, negativen und inkonsistenten Entscheidungen führt - die allesamt Menschenrechtsverletzungen darstellen können - mangelt es auch an Transparenz des Moderationsprozesses, was es schwierig macht, die Plattformen zur Verantwortung zu ziehen. Für Außenstehende sind die Entscheidungen nicht immer nachvollziehbar. Laut Díaz & Hecht-Felella (2021) erfolgen öffentlichkeitswirksame Moderationen häufig ad hoc vorgenommen, wobei die Maßnahmen mit neuen Regeln begründet werden, die an verschiedenen Stellen, von den Blogs des Unternehmens über die Twitter-Accounts des Unternehmens bis hin zu Websites Dritter, angekündigt werden. Dadurch wirkt die das Vorgehen unzusammenhängend und unklar, was es sowohl den Nutzern erschwert, die Regeln zu verstehen oder einzuhalten, als auch externen Gruppen, die Plattformen für ihre Entscheidungen verantwortlich zu machen (Díaz & Hecht-Felella, 2021).

Das Transparenzprobleme besteht auch hinsichtlich der automatisierte Systeme, deren genaue Funktionsweise nicht einsehbar ist (Black-Box-Problematik).

Ein gewisses Maß an Transparenz bei der Moderation von Inhalten ist jedoch unerlässlich, damit sowohl die Nutzer als auch die verantwortlichen Fachleute ausreichend über die Mechanismen informiert sind. Denn dadurch kann ein grundlegendes Maß an Verantwortlichkeit gewährleistet werden und einer dysfunktionalen Ausführung der Moderation entgegengewirkt werden (Elkin-Koren & Perel, M., 2020; Perel & Elkin-Koren, 2017). Perel, & Elkin-Koren (2017)stellen fest, dass die Undurchsichtigkeit von automatisierten Systemen auf zwei Gegebenheiten zurückzuführen ist. So resultiert der Mangel an Transparenz nicht nur aus dem Bestreben der Unternehmen, ihre Systeme zu schützen und aus kommerziellen Gründen geheim zu halten und zu verhindern, dass sie von Dritten ausgenutzt oder umgangen werden, sondern auch aus dem begrenzten technischen

Wissen der Öffentlichkeit, um diese Systeme zu erfassen und zu verstehen (Burrell, 2016; Tas & Wiewiorra, 2022). Selbst die Entwickler wissen oft nur teilweise, wie ihre automatisierten Systeme insgesamt funktionieren. Das liegt daran, dass die meisten verwendeten Algorithmen aus einer Kombinationen von Codes bestehen, die von mehreren Entwicklern gemeinsam entworfen wurden und eine große Anzahl von Rechenverfahren und Variablen enthalten, die zusammenarbeiten. Dadurch ist die zugrunde liegende Syntax selbst für hochqualifizierte Personen schwer zu entziffern (Swart, 2021, Tas et al, 2022). Die zweite Gegebenheit ist ihre Lernfähigkeit. Aufgrund der Lernfähigkeit wäre es selbst dann, wenn einzelne isolierte Schritte nachvollziehbar wären, schwierig zu erkennen, warum der Code tut, was er tut, ohne zu verstehen, wie er entstanden ist und welche Erfahrungen er auf seinem Weg gemacht hat. Die jeweiligen Algorithmen der Systeme schlicht und kommentarlose bereitzustellen, führt daher nicht zwangsläufig dazu, dass diese verstanden und kontrolliert werden können (Perel, & Elkin-Koren, 2017). So ist primäre die sinnvolle Bereitstellung von relevanten Informationen erforderlich(Suzor et al., 2019).

In den letzten Jahren haben die Plattformen ihre Transparenzmaßnahmen aufgrund des zunehmenden Drucks seitens verschiedener Interessengruppen verfeinert. So legen sie beispielsweise die Menge der entfernten Inhalte sowie den Umfang und das Volumen der bei ihnen eingehenden Anträge auf Inhaltsmoderation offen (Suzor et al., 2019, Singh, 2019). Nach Ansicht von Suzor et al. (2019) reichen jedoch aggregierte Daten allein nicht aus, um die detaillierte Analyse zu ermöglichen, die erforderlich ist, um Plattformen zur Verantwortung zu ziehen. Sie können zwar einen Überblick über den Prozess der Inhaltsmoderation geben und grobe Problembereiche aufzeigen, bieten aber nicht die nötige Tiefe für eine umfassende Prüfung. Keller & Leerssen, (2020) stellen ebenfalls fest, dass "aggregated data in transparency reports only shows the platforms' own assessment, and not the merits of the underlying cases. That means researchers can't evaluate the accuracy of takedown decisions, or spot any trends of inconsistent enforcement" (Keller & Leerssen, 2020, S.228). In der Vergangenheit wurden Informationen über einzelne Entscheidungen oft nicht in einer Weise präsentiert, die für Forscher und andere zugänglich war, um die Moderation auf einer allgemeinen Ebene zu verstehen (Suzor et al., 2019). Quintais et al. (2020) fügen hinzu, dass " transparency reports in their current form […] largely focus on the removal of content (and accounts) rather than other (often called "softer") forms of moderation", wie z. B. Shadow-Banning (Quintais et al, 2020, S. 37f.) Ein weiteres Problem in Bezug auf die Rechenschaftslegung ist die begrenzte Möglichkeit, gegen Entscheidungen der Plattformen Einspruch bzw. Beschwerde einzulegen (Hubley, 2022).

4 Regulatorischer und rechtlicher Rahmen für die Moderation von Inhalten

Da sich dieser Diskussionsbeitrag auf Online-Plattformen konzentriert, werden im Folgenden die relevanten Gesetzestexte und Regeln diskutiert, die insbesondere Online-Plattformen betreffen.

Seit ihrer Verabschiedung im Jahr 2000 bildet die ***e-Commerce Richtlinie*** den Rechtsrahmen für digitale Dienste in der EU, mit Vorschriften, die horizontal für alle Dienste der Informationsgesellschaft und damit auch für Online-Plattformen gelten (De Streel et al., 2020; Gellert & Wolters, 2021; Schwemer, 2022). Im Hinblick auf Inhalte, die Nutzer auf Online-Plattformen einstellen, ist insbesondere Art. 14 Abs. 1 lit. a der e-Commerce Richtlinie relevant. Dieser betrifft die Haftung für Online-Inhalte. Danach sind Hosting-Provider, zu denen rechtlich auch Plattformanbieter gehören, von der Haftung befreit, wenn sie keine Kenntnis davon haben, dass sie rechtswidrige Inhalte hosten und damit verbreiten. Nach Art. 14 Abs. 1 lit. b der e-Commerce Richtlinie gilt die Ausnahme auch dann, wenn der Anbieter Kenntnis von einem rechtswidrigen Inhalt hat, sofern er unverzüglich tätig wird, um den Inhalt zu entfernen oder den Zugang zu ihm zu sperren (Madiega, 2020; Gellert & Wolters, 2021). Somit haftet der jeweilige Plattformanbieter zwar nicht für rechtswidrige Inhalte, von denen er keine Kenntnis hat, ist aber dennoch verpflichtet, diese nach Kenntnisnahme zu moderieren (Gellert & Wolters, 2021). In Art. 15 Abs. 1 der e-Commerce Richtline heißt es weiter, dass Anbieter von den Mitgliedsstaaten nicht verpflichtet werden dürfen, "die von ihnen übermittelten oder gespeicherten Informationen zu überwachen oder aktiv nach Umständen zu forschen, die auf eine rechtswidrige Tätigkeit hinweisen".

Einige neuere Rechtsinstrumente behalten zwar die Haftungsausschlussklauseln der e-Commerce Richtline weitgehend bei, legen Plattformanbietern verschiedene prozedurale und Transparenzverpflichtungen auf in Bezug auf die Moderation von Inhalten auf. Dies ist sinnvoll, da die derzeitigen Prozesse zur Moderation von Inhalten auf Plattformen bei weitem nicht perfekt sind, unabhängig von den Bemühungen der Plattformanbieter, mit bestimmten Arten von Inhalten umzugehen, wie in den vorherigen Kapiteln geteigt wurde.

Auf europäischer Ebene verpflichtet so zum Beispiel die ***Richtlinie über audiovisuelle Mediendienste*** Video-Sharing-Plattform-Dienstes dazu, aktiv Maßnahmen zu ergreifen und Inhalte zu moderieren, um Minderjährige vor Inhalten zu schützen, die ihre Entwicklung beeinträchtigen können (Art. 28b Abs. 1 lit. a, Richtlinie über audiovisuelle Mediendienste), die Allgemeinheit vor Inhalten zu schützen, die zu Gewalt oder Hass aufstacheln (Art. 28b Abs. 1 liz. b, Richtlinie über audiovisuelle Mediendienste) oder die Allgemeinheit vor der Verbreitung eine Straftat im Zusammenhang mit Terrorismus, Kinderpornografie, Rassismus und Fremdenfeindlichkeit darstellt (Art. 28b Abs. 1 lit. c, Richtlinie über audiovisuelle Mediendienste) (Sartor et al., 2020). In Art. 28b Abs. 3 der Richtlinie über audiovisuelle Mediendienste wird eine Reihe von Maßnahmen aufgeführt, die Videoplattformen ergreifen müssen. Zu diesen Maßnahmen gehören die Einrichtung von

Mechanismen für die Meldung und Kennzeichnung von Inhalten, Systeme, die erklären, wie auf diese Meldungen oder Hinweise reagiert wurde, Systeme, mit denen die Nutzer die bereitgestellten Inhalte bewerten können, und Verfahren für die Bearbeitung und Beantwortung von Nutzerbeschwerden (Gellert & Wolters, 2021). In Art. 28b Abs. 3 der Richtlinie über audiovisuelle Mediendienste heißt es weiter, dass jede Maßnahme hinsichtlich "der Art der fraglichen Inhalte, des Schadens, den sie anrichten können, der Merkmale der zu schützenden Personenkategorie sowie der betroffenen Rechte und berechtigten Interessen" angemessen sein muss. Die Maßnahmen sollten auch "verhältnismäßig sein und der Größe des Video-Sharing-Plattform-Dienstes und die Art des angebotenen Dienstes s Rechnung tragen". In Art. 28b Abs. 6 der Richtlinie über audiovisuelle Mediendienste heißt es jedoch auch, dass die aufzuerlegenden Maßnahmen unbeschadet der Art. 14 und 15 der e-Commerce Richtlinie gelten (Sartor et al., 2020).

Die ***Richtlinie zum Urheberrecht im digitalen Binnenmarkt*** hingegen weicht ausdrücklich von Art.14 der e-Commerce Richtline ab (Sartor et al., 2020; Hoffmann & Gasparoti, 2020). Art. 17 Abs. 4 der Richtlinie zum Urheberrecht im digitalen Binnenmarkt sieht für die Haftungsbefreiung eine "Best-Effort"-Verpflichtung vor. Danach haften Plattformen, die das Teilen von Inhalten ermöglichen, für die Bereitstellung von urheberrechtsverletzendem Material, wenn sie sich nicht nach besten Kräften bemühen, die Erlaubnis einzuholen, den Inhalt nach Benachrichtigung durch den Urheberrechtsinhaber zu entfernen und sicherzustellen, dass das Material nicht erneut hochgeladen werden kann (Sartor et al., 2020; Gellert & Wolters, 2021).

Schließlich erlegt auch die die ***TERREG-Verordnung*** den Anbietern von Hosting-Diensten, inkl. Online-Plattformen, Sorgfaltspflichten auf, "um die öffentliche Verbreitung terroristischer Inhalte durch ihre Dienste zu bekämpfen" (Art. 1 Abs. 1 lit a, TERREG-Verordnung). Die Verordnung sieht eine Reihe von Maßnahmen für vor. Nach dieser Verordnung sind Plattformanbieter verpflichtet, terroristische Inhalte zu entfernen oder den Zugang zu terroristischen Inhalten in allen Mitgliedstaaten so schnell wie möglich, in jedem Fall aber innerhalb einer Stunde nach Eingang der Entfernungsanordnung bei einer zuständigen Behörde, zu sperren (Art. 3 Abs. 3, TERREG-Verordnung). Darüber hinaus müssen die Anbieter unter anderem Mechanismen einrichten, um terroristische Inhalte zu ermitteln und rasch zu entfernen oder den Zugang zu ihnen zu sperren, den Nutzern die Möglichkeit geben, Inhalte zu melden und zu kennzeichnen, das Bewusstsein für terroristische Inhalte in ihren Diensten schärfen (Art. 5 Abs. 2, TERREG-Verordnung). Zudem müssen Beschwerdemechanismen eingerichtet werden (Art. 10, TERREG-Verordnung). Außerdem müssen sie bestimmte Transparenzverpflichtungen erfüllen, wie z. B. die Berichterstattung über Maßnahmen zur Überprüfung und Sperrung des Zugangs zu terroristischen Inhalten, Maßnahmen gegen das erneute Hochladen terroristischer Inhalte, Informationen über die Anzahl der gesperrten oder deaktivierten Inhalte sowie die Anzahl und das Ergebnis von Beschwerden und behördlichen oder gerichtlichen Überprüfungsverfahren (Art. 7 Abs. 3, TERREG-Verordnung) (Gellert & Wolters, 2021). Diese

Verordnung berührt auch nicht Art. 15 der e-Commerce Richtlinie (Erwägungsgrund 25 & Art. 5 Abs. 8, TERREG-Verordnung).

Darüber hinaus haben mehrere Mitgliedstaaten bereits einige Versuche oder erfolgreiche Umsetzungen nationaler Rechtsvorschriften mit unterschiedlichem Umfang und Anwendungsbereich unternommen, wie z. B. das ***NetzDG*** in Deutschland oder das französische Gesetz "***Loi Avia***". Es gibt auch einige EU-Richtlinien wie die ***Richtlinie zur Bekämpfung des sexuellen Missbrauchs und der sexuellen Ausbeutung von Kindern*** und die ***Richtlinie zur Bekämpfung des Terrorismus***, die insbesondre an die Mitgliedstaaten adressiert sind und sie dazu verpflichten, gegen die Verbreitung von missbräuchlichen und terroristischen Inhalten vorzugehen (De Streel et al., 2020).

Tabelle 4-1: Die wichtigsten EU-Verordnungen und Richtlinien zu illegalen Online-Inhalten

Anwendungsbereich	Hartes Recht	Weiches Recht
Anwendbar auf alle Arten von Online-Inhalten und anwendbar auf alle Arten von Online-Plattformen	***Richtlinie 2000/31/EC*** des Europäischen Parlaments und des Rates vom 8. Juni 2000 über bestimmte rechtliche Aspekte der Dienste der Informationsgesellschaft, insbesondere des elektronischen Geschäftsverkehrs, im Binnenmarkt („Richtlinie über den elektronischen Geschäftsverkehr“) ***Verordnung (EU) 2022/2065*** des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG (Gesetz über digitale Dienste)	***Mitteilung COM(2017) 555 final*** der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen vom 28. September 2017 zum Umgang mit illegaler Online-Inhalten ***Empfehlung (EU) 2018/334*** der Kommission vom 1. März 2018 für wirksame Maßnahmen im Umgang mit illegalen Online-Inhalten
Regeln für Video-Sharing-Plattformen	***Richtlinie 2010/13/EU*** des Europäischen Parlaments und des Rates vom 10. März 2010 zur Koordinierung bestimmter Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Bereitstellung audiovisueller Mediendienste (Richtlinie über audiovisuelle Mediendienste) in der Fassung der Richtlinie (EU) 2018/ des Europäischen Parlaments und des Rates vom 14. November 2018 zur Änderung der Richtlinie 2010/13/EU zur Koordinierung bestimmter Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Bereitstellung audiovisueller Mediendienste (Richtlinie über audiovisuelle Mediendienste) im Hinblick auf sich verändernde Marktgegebenheiten	
Arten von Inhalten	**Hartes Recht**	**Selbst-Regulierung**
Terroristische Inhalte	***Richtlinie (EU) 2017/541*** des Europäischen Parlaments und des Rates vom 15. März 2017 zur Ter-rorismusbekämpfung und zur Ersetzung des Rahmenbeschlusses 2002/475/JI des Rates und zur Änderung des Beschlusses 2005/671/JI des Rates ***Verordnung (EU) 2021/784*** des Europäischen Parlaments und des Rates vom 29.April 2021 zur Bekämpfung der Verbreitung terroristischer Online-Inhalte	**EU Internet Forum (2015)**
Material über sexuellen Missbrauch von Kindern	***Richtlinie 2011/92/EU*** des Europäischen Parlaments und des Rates vom 13.Dezember 2011 zur Bekämpfung des sexuellen Missbrauchs und der sexuellen Ausbeutung von Kindern sowie der Kinderpornografie sowie zur Ersetzung des Rahmenbeschlusses 2004/68/JI des Rates	**Bündnis für einen besseren Online-Minderheitenschutz (2017)**
Illegale Hassreden	***Rahmenbeschluss 2008/913/JI*** des Rastes vom 28. November 2008 zur strafrechtlichen Bekämpfung bestimmter Formen und Ausdrucksweisen von Rassismus und Fremdenfeindlichkeit	**Verhaltenskodex für illegale Hassreden im Internet (2016)**
Verletzung der Eigentums-/Urheberrechte	***Richtlinie (EU) 2019/790*** des Europäischen Parlaments und des Rates vom 17.April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG ***Richtlinie 2004/48/EG*** des Europäischen Parlaments und des Rates vom 29. April 2004 zur Durchsetzung der Rechte des geistigen Eigentums	**Gemeinsame Absichtserklärung über gefälschte Waren im Internet (2011, überarbeitet 2016)**

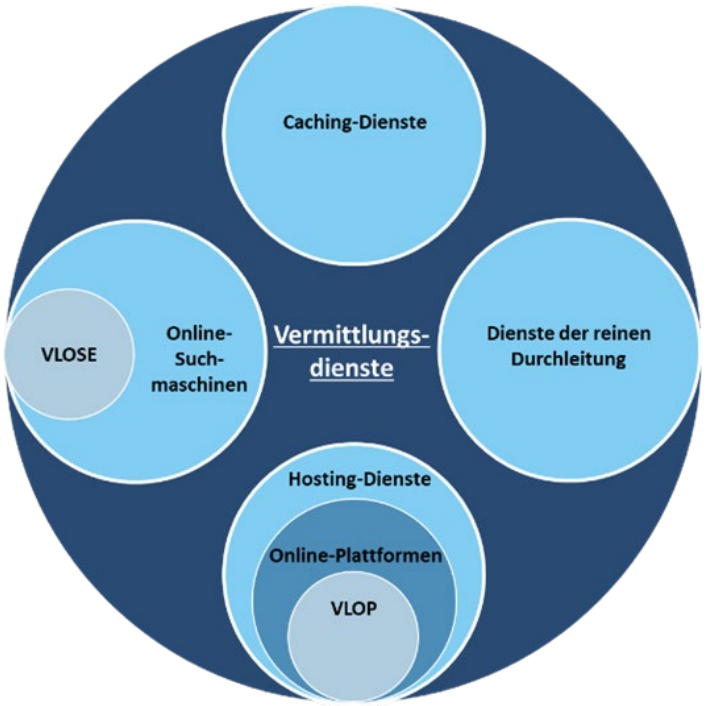
Quelle: Eigene Darstellung basierend auf De Streel et al. (2020) und Europäisches Parlament (2021). Die Tabelle hat keinen Vollständigkeitsanspruch.

Dieser Flickenteppich aus verbindlichen und nicht verbindlichen Instrumenten, die sich an verschiedene Parteien, Mitgliedstaaten und Diensteanbieter, richten und spezifische Regeln für verschiedene Arten von Inhalten vorsehen und es den Mitgliedstaaten überlassen, ihre eigenen nationalen Regeln mit unterschiedlichem Umsetzungsgrad anzuwenden, führte zu einem uneinheitlichen und fragmentierten Rechtsrahmen innerhalb des europäischen Binnenmarkts (Gellert & Wolters, 2021; De Streel et al., 2022).

Der DSA, der 2022 verabschiedet wurde, soll die Rechtslage für Online-Plattformen und andere Vermittlungsdienste in der EU harmonisieren und vereinfachen (Erwägungsgrund 4, DSA). Mit der Einführung eines neuen horizontalen Rechtsrahmens wird vor allem die e-Commerce Richtlinie reformiert. Es ist wichtig zu beachten, dass dieser Rahmen mit den vertikalen EU-Rechtsvorschriften koexistieren wird, die nur für bestimmte Arten von Inhalten und Diensten gelten (Wilman, 2022; Erwägungsgründe 10 & 11, DSA).

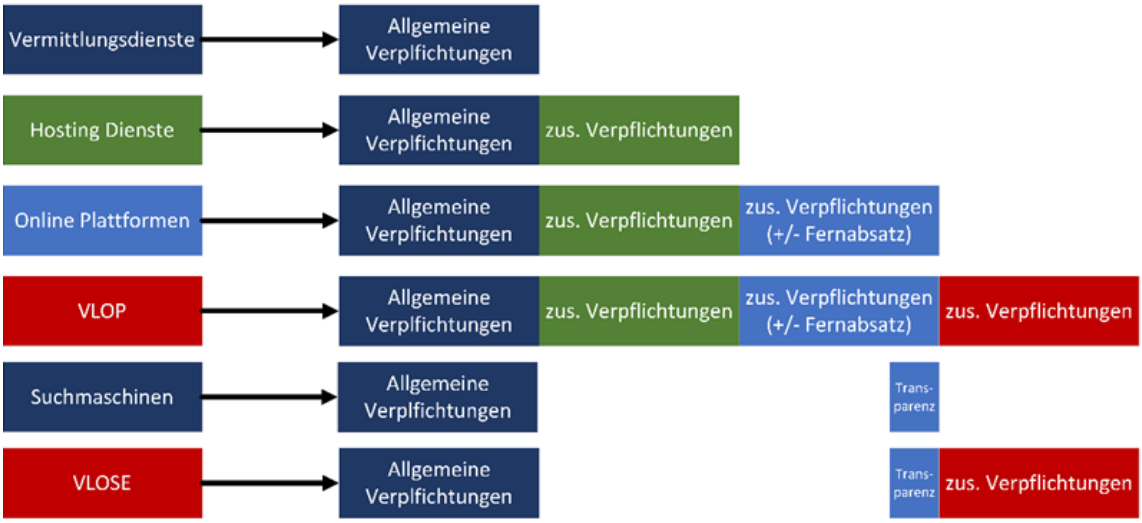
Der DSA gilt für alle in der EU tätigen Vermittlungsdienste, einschließlich Online-Plattformen, unabhängig davon, wo sie niedergelassen sind. Das Gesetz verfolgt jedoch einen asymmetrischen Regulierungsansatz. Das bedeutet, dass die Verpflichtungen, die gelten, von der Art der erbrachten Vermittlungsleistung sowie von deren Größe und Reichweite des Dienstes abhängen (G'sell, 2023; Husovec, 2023). Die Abbildungen veranschaulichen die Arten von Vermittlungsdiensten, die unter den DSA fallen (Abbildung 4-1), und die abgestufte Anwendung der Vorschriften auf diese Dienste (Abbildung 4-2).

Abbildung 4-1: Durch den DSA adressierten Dienste



Quelle: Eigene Darstellung basierend auf Quintais et al. (2022). VLOP: Very Large Online Platforms (zu Deutsch: sehr große Online-Plattformen); VLOSE: Very Large Online Search Engines (zu Deutsch: sehr große Online-Suchmachinen).

Abbildung 4-2: Kaskade der DSA-Verpflichtungen



Quelle: Eigene Darstellung basierend auf Schmid & Koehler (2022). VLOP: Very Large Online Platforms (zu Deutsch: sehr große Online-Plattformen); VLOSE: Very Large Online Search Engines (zu Deutsch: sehr große Online-Suchmaschinen).

Während alle Online-Plattformen die allgemeinen und einige zusätzliche Verpflichtungen erfüllen müssen, gelten einige Verpflichtungen nur für sehr große Online-Plattformen (im Folgenden: VLOPs) mit mehr als 450 Millionen Nutzern und/oder Plattformen mit mehr als 50 Mittarbeiter oder einem Umsatz von mehr als 10 Millionen Euro beschränkt (Husovec, 2023). Durch „rebalancing responsibilities in the online ecosystem according to the size of the players, the DSA ensures that the regulatory costs of these new rules are proportionate" (Europäische Kommission, 2023).

Außerdem können VLOPs aufgrund ihrer großen Reichweite gesellschaftliche Risiken schaffen, deren Ausmaß und Folgen sich von denen kleinerer Plattformen deutlich unterscheiden. Anbieter solcher VLOPs sollten nach dem DSA dem höchsten Standard der Sorgfaltspflicht unterliegen, der in einem angemessenen Verhältnis zu ihren gesellschaftlichen Auswirkungen steht (Erwägungsgründe 76, DSA). In Erwägungsgrund 79 des DSA heißt es: "Im Rahmen dieser Verordnung sollten Anbieter sehr großer Online-Plattformen und sehr großer Online-Suchmaschinen daher prüfen, welche systemischen Risiken mit der Gestaltung, Funktionsweise und Nutzung ihrer Dienste sowie mit einem möglichen Missbrauch durch die Nutzer verbunden sind, und sollten unter Achtung der Grundrechte angemessene Gegenmaßnahmen treffen" (Erwägungsgrund 79, DSA). Die vier identifizierten Kategorien von systemischen Risiken sind

- Risiken, „die durch Verbreitung rechtswidriger Inhalte entstehen können, darunter die Verbreitung von Darstellungen von sexuellem Missbrauch von Kindern oder von rechtswidriger Hassrede oder andere Arten von Missbrauch ihrer Dienste für Straftaten sowie rechtswidrige Tätigkeiten wie ein nach Unions- oder nationalem

Recht untersagter Verkauf von Waren oder Dienstleistungen, wie z. B. gefährlicher oder gefälschter Güter oder rechtswidrig gehandelter Tiere" (Erwägungsgrund 80, DSA);

- Risiken in Bezug auf" tatsächlichen oder absehbaren Auswirkungen des Dienstes auf die Ausübung der durch die Charta der Grundrechte geschützten Grundrechte, einschließlich, aber nicht beschränkt auf Menschenwürde, Freiheit der Meinungsäußerung und Informationsfreiheit, einschließlich der Freiheit und des Pluralismus der Medien, Recht auf Achtung des Privatlebens, Datenschutz, Recht auf Nichtdiskriminierung, Rechte des Kindes und Verbraucherschutz" (Erwägungsgrund 81, DSA);
- Risiken in Bezug auf "die tatsächlichen oder absehbaren negativen Auswirkungen auf demokratische Prozesse, die gesellschaftliche Debatte und Wahlprozesse sowie auf die öffentliche Sicherheit" (Erwägungsgrund 82, DSA); und
- Risiken in Bezug auf "tatsächlichen oder absehbaren negativen Auswirkungen auf den Schutz der öffentlichen Gesundheit oder von Minderjährigen und schwerwiegenden negativen Folgen für das körperliche und geistige Wohlbefinden einer Person oder in Bezug auf geschlechtsspezifische Gewalt" (Erwägungsgrund 83, DSA).

Ähnlich wie die anderen, neueren Regulierungen behält auch der DSA die ursprüngliche Klausel zum Haftungsausschluss der e-Commerce Richtlinie für Online-Inhalte weiterhin bei.

Dass der Haftungsausschluss erhalten bleibt, kann u.a. damit begründet werden, dass die Auslegung der Haftungsbestimmungen dem nationalen Recht unterliegt. Auf diese Weise vereinheitlicht der DSA den Ausschluss der Haftung von Online-Plattformen für rechtswidrige Inhalte, die von den Nutzern übermittelt werden, und verhindert, dass in den einzelnen Mitgliedstaaten in diesem Zusammenhang unterschiedliche Haftungsregelungen zur Anwendung kommen (Buiten, 2021). Zudem wird dadurch das in Kapitel 2.1.2 dargestellte Risiko vermindert, dass durch eine auferlegte Haftung für die Anbieter Anreize geschaffen werden, übermäßig Inhalte zu moderieren oder gar die übermittelten Inhalte strukturell zu überwachen. So befreit Art. 6 Abs. 1 des DSA Anbieter von Hosting-Diensten – darunter auch Online-Plattformen – von der Haftung, sofern sie illegale Inhalte unverzüglich entfernen, sobald sie von ihnen Kenntnis erlangen (Buiten, 2021; G'sell, 2023). Diese Ausschlussregelung gilt jedoch nicht, wenn der Anbieter "sich nicht darauf beschränkt, die Dienstleistungen auf neutrale Weise und durch die bloße technische und automatische Verarbeitung der vom Nutzer bereitgestellten Informationen zu erbringen", sondern eine aktive Rolle spielt, die ihm Kenntnis oder Kontrolle über den Inhalt verschafft, wie in Erwägungsgrund 18 des DSA dargelegt (G'sell, 2023). Durch die Nutzung von Moderations- und Empfehlungssystemen könnten den meisten Plattformen eine solche aktive Rolle zugestanden werden. Jedoch ist dies im Zusammenhang mit der „Guter-Samariter-Klausel" des Artikels 7 des DSA zu interpretieren (Buiten, 2021). Nach dieser können für Anbieter die „Haftungsausschlüsse auch dann in Betracht, wenn sie auf

Eigeninitiative nach Treu und Glauben und sorgfältig freiwillige Untersuchungen durchführen oder andere Maßnahmen zur Erkennung, Feststellung und Entfernung rechtswidriger Inhalte oder zur Sperrung des Zugangs zu rechtswidrigen Inhalten treffen“ (Artikel 7, DSA). Gleichwohl müssen sie dabei objektiv, nicht diskriminierend und verhältnismäßig Vorgehen und sicherstellen, dass bei der Verwendung „automatisierter Werkzeuge zur Durchführung solcher Maßnahmen die betreffende Technologie ausreichend zuverlässig ist, um die Fehlerquote so weit wie möglich zu begrenzen“ (Erwägungsgrund 26, DSA). Eine Verpflichtung zur Überwachung von oder aktiven Nachforschung nach rechtswidrige Tätigkeiten gibt es jedoch nicht (Art. 8, DSA).

In den Artikeln 10 und 11 werden darüber hinaus Regeln für die Anordnungen eingeführt, die von den nationalen Justiz- oder Verwaltungsbehörden an die Anbieter gerichtet werden können. Diese Anordnungen können die Anbieter verpflichten, mit den Justiz- oder Verwaltungsbehörden der Mitgliedstaaten zusammenzuarbeiten, um gegen bestimmte Fälle illegaler Inhalte vorzugehen (Buiten, 2021).

Doch der DSA behält nicht nur die Haftungsbefreiung bei, sondern führt ebenfalls einige Sorgfaltspflichten zu Inhaltsmoderation und ihrer Transparenz ein (G’sell, 2023). Wie die Diskussion im vorangegangenen Kapitel gezeigt hat, führen die Moderationsbemühungen von Plattformanbietern nicht immer zu den gesellschaftlich wünschenswerten Ergebnissen, sei es aufgrund ihrer eigenen Motivation zur Gewinnmaximierung oder ungewollt aufgrund der Grenzen der derzeitigen Moderationsansätze. Ein undurchsichtiger Moderationsprozess macht es schwierig, einzelne Entscheidungen nachzuvollziehen und mögliche Fehlinterpretationen aufzudecken. Dies ist besonders problematisch, weil Moderationsentscheidungen großen persönlichen oder gesellschaftlichen Schaden anrichten können.

In den nachfolgenden Abschnitten wird daher betrachtet, welche der zentrale Regulierungsmaßnahme im Bereich der Inhaltsmoderation die oben genannten Problemfelder in ihren Verpflichtungen aufgreift.

4.1 Förderung von Transparenz

Je nach Größe der Plattform enthält der DSA unterschiedliche Sorgfaltspflichten, die Transparenz über die Moderation von Inhalten schaffen sollen. Für sehr große Online-Plattformen gelten die meisten Verpflichtungen (Tabelle 4-2).Diese Verpflichtungen lassen sich in vier Kategorien zusammenfassen.

Tabelle 4-2: Relevante Verpflichtungen im DSA im Zusammenhang mit der Moderation von Inhalten

	Vermittlungs-dienste	Hosting Dienste	Online Plattformen	VLOPs & VLOSEs
Allgemeine Geschäftsbedingungen	Art. 14	Art. 14	Art. 14	Art. 14
Verpflichtungen zur Transparenz-berichterstattung	Art. 15	Art. 15	Art. 15 Art. 24	Art. 15 Art. 24 Art. 42
Melde- und Beschwerdesysteme		Art. 16 Art. 17	Art. 16 Art. 17 Art. 20 Art. 22 Art. 23	Art. 16 Art. 17 Art. 20 Art. 22 Art. 23
Risikobewertung und -minderung				Art. 34 Art. 35

Quelle: Eigene Darstellung. VLOP: Very Large Online Platforms (zu Deutsch: sehr große Online-Plattformen); VLOSE: Very Large Online Search Engines (zu Deutsch: sehr große Online-Suchmachinen).

Verpflichtungen, die die allgemeinen Geschäftsbedingung betreffen

Dazu gehört die Verpflichtung der Plattformanbieter, die Geschäftsbedingungen in klarer und benutzerfreundlicher Sprache zu erläutern und sie in einer leicht zugänglichen und maschinenlesbarer Form öffentlich zur Verfügung zu stellen. Die Geschäftsbedingung soll Angaben zu „ allen Leitlinien erfahren, Maßnahmen und Werkzeuge, die zur Moderation von Inhalten eingesetzt werden […] sowie zu den Verfahrensregeln für ihr internes Beschwerdemanagementsystem“ beinhalten (Art. 14 Abs. 1, DSA). Über Änderungen in den Bedingungen sind die Nutzer zu informieren (Art.14 Abs., DSA).

Verpflichtungen, die die Einrichtung von Melde- und Beschwerdesysteme betreffen

Dazu gehören Regelungen zur Ausgestaltung der Meldeverfahren - einschließlich der erforderlichen Angaben der meldenden Personen oder Stellen zu den von ihnen als rechtswidrig angesehenen Inhalten (Art. 16, DSA). Darüber hinaus sieht Art. 17 DSA vor, dass der von einer Moderation betroffene Nutzer eine Begründung für jede Entscheidung erhalten muss, die eine Beschränkung von Inhalten, die Beschränkung oder Schließung von Nutzerkonten oder die Beschränkung von Geldzahlungen als Maßnahme gegen die Verbreitung rechtswidriger oder unerwünschter Inhalte nach sich zieht. Eine solche Begründung sollte Informationen über die Tatsachen und Umstände enthalten, auf denen die Entscheidung beruht, einschließlich Informationen über die Grundlage der Entscheidung ("Rechtsvorschriften" oder "Allgemeine Geschäftsbedingungen") sowie Informationen über die Rechtsbehelfe, die dem Nutzer zur Verfügung stehen. Schließlich regelt Art. 20 des DSA die Ausgestaltung von Beschwerdemanagementsystemen, die Nutzern und meldenden Personen bzw. Einrichtungen die Möglichkeit einräumen, sich bei den Plattformanbietern über ihre Entscheidung zu beschweren. Dabei soll das Beschwerdesystem leicht zugänglich und benutzerfreundlich gestaltet sein und die Online-Plattform sollen

diese zeitnah, sorgfältig und frei von Willkür bearbeiten. Sollte es zu einer Fehlentscheidung gekommen sein, muss diese Rückgängig gemacht werden (Art. 20 Abs. 4, DSA). Zu Identifizierung von rechtswidrigen Inhalten sollen zudem vertrauenswürde Hinweisgeber eingesetzt werden. Diese sind DSC-zertifizierte Stellen im Rahmen der DSA, die Meldungen einrichten können, die vorrangig behandelt werden. Ein vertrauenswürdig Hinweisgeber kann eine öffentliche Einrichtung, eine NGO, aber auch eine private oder halböffentliche Einrichtung sein. Eine solche Stelle muss über Fachwissen und Kompetenz zur Aufdeckung illegaler Inhalte, Unabhängigkeit, transparente Finanzierungsstrukturen und einen Jahresbericht über ihre Tätigkeit verfügen. Sie müssen jährliche Berichte über die von ihnen eingereichten Meldungen erstellen, die auch die von den Plattformanbietern getroffenen Entscheidungen und Maßnahmen enthalten (Art. 22, DSA). Art. 23 des DSA enthält allgemeine Sicherheitsmaßnahmen gegen einen möglichen Missbrauch des Melde- und Beschwerdesystems.

Da sich der DSA noch in der Umsetzung befindet, lassen sich noch keine genauen Implikationen dieser Regelungen ableiten. Jedoch könnten sie dazu beitragen, dass das Vorgehen gegen Fehlentscheidungen – insbesondere für privat Personen – zu einem gewissen Grad vereinfacht wird. Das Einbinden von vertrauenswürdigen Hinweisgeber, kann eine zusätzliche Ebene der Kontrolle ermöglichen. Sie haben eine relativ starke Position im System, da ihre Berichte von Online-Plattformen bevorzugt behandelt und von der DSC verwendet werden. Daher ist aber auch ein gewisses Maß an Transparenz und Überwachung ihrer Aufgaben wünschenswert und sollte in die Berichterstattungspflichten der Trusted Flagger aufgenommen werden. Es ist klar, dass die Auswahl, Zertifizierung und ggf. Überwachung ihrer Arbeit ein neuralgischer Punkt ist.

Verpflichtungen, die die Riskobewertung und -minderung betreffen.

Im Rahmen der Risikobewertung (Art. 34, DSA) müssen VLOPs mindestens einmal jährlich alle wesentlichen systemischen Risiken ermitteln, die sich aus dem Betrieb ihrer Dienste ergeben. Die Risikobewertung umfasst Risiken im Zusammenhang mit der Verbreitung illegaler Inhalte, negativen Auswirkungen auf die Grundrechte und absichtlicher Manipulation ihres Dienstes. Bei der Risikobewertung berücksichtigen die VLOPs insbesondere, wie sich ihr Gesamtdienst, einschließlich der Systeme zur Moderation von Inhalten und der Empfehlungssysteme, auf eines der Systemrisiken auswirkt. Basierend darauf haben VLOPs Maßnahmen zu ergreifen, um Risiken zu minimieren. Dies beinhaltet unter anderem die Anpassung der Geschäftsbedingung und die Verfahren zur Moderation (Art. 35, DSA).

Verpflichtungen, die die Transparenzberichterstattung und den Zugang zu Daten betreffen

Alle Plattformanbieter, mit Ausnahme derjenigen, bei denen es sich um Kleinst- oder Kleinunternehmen handelt, müssen mindestens einmal jährlich Transparenzberichte veröffentlichen, in denen sie über die Aktivitäten der Inhaltsmoderation Auskunft geben (Art.

15, DSA). Diese Berichte müssen insbesondere Informationen über die von den zuständigen Behörden erhaltenen Anfragen enthalten, aufgeschlüsselt nach der Art der illegalen Inhalte und nach der Art der Benachrichtigung. Darüber hinaus müssen Plattformanbieter Informationen über die von ihnen initiierte Moderation zur Verfügung stellen, einschließlich Angaben zu den eingesetzten automatisierten Systemen und menschlichen Moderatoren sowie zu den getroffenen Maßnahmen. Zusätzlich zu den in Art. 15 des DSA geforderten Daten müssen die Berichte auch Informationen über Streitigkeiten, Kontosperrungen im Rahmen der Maßnahmen nach Art. 23 des DSA enthalten. Mindestens halbjährlich müssen die Online-Plattformen auch die durchschnittliche Anzahl der monatlich aktiven Nutzer in jedem Mitgliedsstaat veröffentlichen (Art. 24, DSA).

Die ersten Transparenzberichte im Rahmen des DSA wurden von den VLOPs bereits Ende Oktober 2023 vorgelegt. Gemäß Art. 42 des DSA müssen VLOPs mindestens alle sechs Monate einen Transparenzbericht in maschinenlesbarem Format vorlegen. Die restlichen Online-Plattformen nach Art. 15 und 24 des DSA nur jährlich. Die Transparenzberichte muss entsprechen Art. 15, 24 und 42 folgende Informationen enthalten[15]:

- Anzahl der von den Behörden der Mitgliedstaaten eingegangenen Anfragen, aufgeschlüsselt nach der Art der illegalen Inhalte, dem Mitgliedstaat, der die Anfrage gestellt hat, und der für die Antwort benötigten Zeit;
- Anzahl der eingereichten Beschwerden, aufgeschlüsselt nach der Art des mutmaßlich illegalen Inhalts, Anzahl der von vertrauenswürdigen Flaggenanbietern eingereichten Beschwerden, etwaige aufgrund der Beschwerden ergriffene Maßnahmen, wobei zu unterscheiden ist, ob die Maßnahmen auf der Grundlage des Gesetzes oder der Geschäftsbedingungen des Anbieters ergriffen wurden, Anzahl der mit automatisierten Mitteln bearbeiteten Beschwerden und die für die Beantwortung benötigte Zeit;
- Informationen über die von den Anbietern auf eigene Initiative durchgeführte Inhaltsmoderation, einschließlich des Einsatzes automatisierter Werkzeuge, der Maßnahmen zur Schulung und Unterstützung der mit der Inhaltsmoderation betrauten Personen sowie der Anzahl und Art der ergriffenen Maßnahmen;
- Anzahl der eingegangenen Beschwerden, die Grundlage für diese Beschwerden, die aufgrund dieser Beschwerden getroffenen Entscheidungen, die für diese Entscheidungen durchschnittlich benötigte Zeit und die Anzahl der Fälle, in denen diese Entscheidungen rückgängig gemacht wurden;
- Einsatz automatisierter Mittel für die Moderation von Inhalten,
- Anzahl der Streitfälle, die den außergerichtlichen Streitbeilegungsstellen vorgelegt wurden, die Ergebnisse der Streitbeilegung und die durchschnittlich benötigte Zeit für den Abschluss der Streitbeilegungsverfahren sowie der Anteil der

15 Siehe zu der Liste Liebe & Wiewiorra (2023).

Streitfälle, in denen der Anbieter der Online-Plattform die Entscheidungen der Stelle umgesetzt hat;

- Anzahl der gemäß Artikel 23 des DSA verhängten Aussetzungen, wobei zwischen Aussetzungen wegen der Bereitstellung offensichtlich illegaler Inhalte, der Einreichung offensichtlich unbegründeter Meldungen und der Einreichung offensichtlich unbegründeter Beschwerden zu unterscheiden ist.
- Angabe der Humanressourcen, die der Anbieter sehr großer Online-Plattformen für die Moderation von Inhalten einsetzt, aufgeschlüsselt nach den jeweiligen Amtssprachen der Mitgliedstaaten,
- Angabe der Qualifikationen und der Sprachkenntnisse der Personen, die diese Tätigkeiten ausführen, sowie der Schulung und Unterstützung, die diese Mitarbeiter erhalten;
- Angabe der Genauigkeitsindikatoren und der damit verbundenen Informationen gemäß Artikel 15 Absatz 1 Buchstabe e des DSA, aufgeschlüsselt nach den einzelnen Amtssprachen der Mitgliedstaaten.

Diese Informationen können Moderationsprozesse und -entscheidungen transparenter machen. Zwar handelt es sich auch hier meisten um aggregierte Daten und weniger um Informationen zu einzelnen Entscheidungen, dennoch können diese wie in Kapitel 3.2 beschrieben, zumindest einen Überblick über die jeweiligen Moderationsprozesse und -entscheidungen bieten. Soweit eine Vergleichbarkeit in der Darstellung der Daten gegeben ist, kann damit die Moderation von Inhalten über die Zeit hinweg zu beobachtet und Veränderungen identifiziert werden. Gleichzeitig können unter der gleichen Prämisse Plattformen ähnlichen Charakters miteinander vergleichen werden, um regulatorische Eingriffe abzuwägen. Schließlich können die Informationen für ein Benchmarking der durchgeführten Inhaltsmoderation verwendet werden (Liebe & Wiewiorra, 2023).

Darüber hinaus sind Indikatoren enthalten, die für die Verbesserung der Qualität und die Weiterentwicklung von Systemen und Techniken der Inhaltsmoderation wichtig sind. Dabei handelt es sich um qualitative Angaben zu den Techniken und Methoden sowie um Informationen wie Kapazitäten und Qualifikationen des eingesetzten Personals (Liebe & Wiewiorra, 2023).

Die folgende Tabelle gibt einen Überblick über die Transparenzberichte der VLOPs bis Oktober 2023.

Tabelle 4-3: Transparenzberichte der VLOPs bis Oktober 2023

	Unternehmen	Dienst	Link zum Bericht
Social Media	Alphabet	YouTube (joint report with the other alphabet services)	https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf
	Meta	Facebook	https://transparency.fb.com/sr/dsa-transparency-report-oct2023-facebook/
	Meta	Instagram	https://transparency.fb.com/sr/dsa-transparency-report-oct2023-instagram/
	Bytedance	TikTok	https://sf16-va.tiktokcdn.com/obj/eden-va2/fssl-reh7uulsn/DSA%20Report%20October%202023/DSA%20draft%20Transparency%20report%20-%2025%20October%202023.pdf
	Microsoft	LinkedIn	https://content.linkedin.com/content/dam/help/linkedin/en-us/October-2023-LinkedIn-DSA-Transparency-Report10.pdf
	Snap	Snapchat	https://values.snap.com/privacy/transparency?lang=en-US
	Pinterest	Pinterest	https://policy.pinterest.com/en/digital-services-act-transparency-report
	Twitter/X	Twitter/X	https://transparency.twitter.com/dsa-transparency-report.html
App Stores	Alphabet	Google App Store /Google Play	Gemeinsamer Bericht mit den anderen Diensten von Alphabet
	Apple	Apple App Store	https://www.apple.com/legal/dsa/transparency/eu/app-store/2310/
Wiki	Wikimedia	Wikipedia	https://foundation.wikimedia.org/wiki/Legal:Supplemental_Transparency_Report_for_August-September_2023
Online Marktplätze	Amazon	Amazon Marketplace	https://assets.aboutamazon.com/cd/28/4d02dd2e41ec8c6d1bc341e9d919/amazon-eu-store-transparency-report-jan-june-2023.pdf
	Alphabet	Google Shopping (joint report with the other alphabet services)	Gemeinsamer Bericht mit den anderen Diensten von Alphabet
	Alibaba	AliExpress	https://files.alicdn.com/tpsservice/0475f95eec8798f4d6e9937a08e77c38.pdf?spm=a2g0o.tm1000005123.1782285040.1.51e96f3d1CSSYv&aecmd=true
	Booking.com	Booking.com	https://r-xx.bstatic.com/data/mobile/dsa_transparency_report_bf3fdc24.pdf
	Zalando	Zalando	https://mosaic02.ztat.net/cnt/contentful-apps/uploads/a74cdebf-cfc7-46dd-8853-13afed1e41aa.pdf
Maps	Alphabet	Google Maps (joint report with the other alphabet services)	Gemeinsamer Bericht mit den anderen Diensten von Alphabet

Quelle: Angepasste Version der Tabelle von Liebe & Wiewiorra (2023).

Art. 24 des DSA schreibt vor, dass Entscheidungen und Begründungen gemäß Art. 17 Abs. 1 des DSA zur Inhaltsmoderation der Kommission gemeldet werden müssen. Die Kommission hat die Aufgabe, eine öffentlich zugängliche Datenbank für diese Informationen einzurichten und zu pflegen. Diese Transparenzdatenbank wurde bereits eingeführt und enthält Daten seit Ende September 2023. Das System ist über die Website https://transparency.dsa.ec.europa.eu/ zugänglich.

Artikel 17 Absatz 3 des DSA spezifiziert die zu meldenden Informationen wie folgt:

- Tatsachen und Umstände, auf die sich die Entscheidung stützt,
- Angaben über den Einsatz automatisierter Mittel bei der Entscheidung,

- Hinweis auf den geltend gemachten Rechtsgrund,
- Hinweis auf den geltend gemachten vertraglichen Grund,
- klare und benutzerfreundliche Informationen über die Rechtsbehelfsmöglichkeiten, die dem Dienstleistungsempfänger gegen die Entscheidung zur Verfügung stehen.

Insbesondere die Transparenzberichte (Art. 15 Abs.1 und Art. 24 Abs. 1) und die Datenbank (Art. 24 Abs. 5, die Entscheidungen über die Entfernung von Inhalten enthält) tragen Sichtbarkeit der Moderationsprozesse bei.

4.2 Einführung einer Aufsichtsstruktur

Darüber hinaus wird unter dem DSA eine Aufsichtsstruktur eingerichtet, um insbesondere die Einhaltung der Vorschriften des DSA zu gewährleisten. Daran werden sowohl die EU-Kommission als auch nationale Behörden beteiligt sein. Sollten den Verpflichtungen des DSA nicht nachgegangen werden, kann dies Sanktionen nach sich ziehen (Art. 52, DSA).

4.2.1 Nationale Aufsicht: Der Koordinator für digitaler Dienste

Der DSA überlässt es den Mitgliedstaaten, eine oder mehrere Behörden mit der Überwachung und Durchsetzung des DSA zu betrauen. (Art.49 Abs. 1, DSA). So muss in jedem Mitgliedstaat ein Koordinator für digitale Dienste (zu Englisch: Digital Service Coordinator, DSC) benannt werden (Art. 49 Absatz 2, DSA). Der Entwurf des Digitale Dienste Gesetzes[16] in Deutschland sieht in Art. 1 eine unabhängige Stelle bei der Bundesnetzagentur (BNetzA) als Koordinator für digitale Dienste in Deutschland vor. Die Frist für das Inkrafttreten ist der Februar 2024.

Der DSA berührt Bereiche wie Verbraucherschutz, Medienregulierung und Datenschutz, für die die meisten Mitgliedstaaten eigene Einrichtungen haben. Um diese Behörden zu koordinieren und den Austausch mit der Kommission zu gewährleisten, muss es in jedem Mitgliedstaat eine einzige Stelle auf nationaler Ebene geben, die als Koordinator fungiert. Darüber hinaus muss der DSC auch selbst wichtige Überwachungsaufgaben wahrnehmen: Er prüft Forschungsanträge für den Datenzugang zu Plattformen im Namen aller Nutzer, muss den Zugang zu Plattformdaten sicherstellen und außergerichtliche Streitbeilegungsstellen akkreditieren. Darüber hinaus ist jeder DSC Teil des Europäischen Gremiums für digitale Dienste, in dem alle nationalen DSCs und die Kommission

16 Entwurf eines Gesetzes zur Durchführung der Verordnung (EU) 2022/2065 des Europäischen Parlaments und des Rates vom 19. Oktober 2022 über einen Binnenmarkt für digitale Dienste und zur Änderung der Richtlinie 2000/31/EG sowie zur Durchführung der Verordnung (EU) 2019/1150 des Europäischen Parlaments und des Rates vom 20. Juni 2019 zur Förderung von Fairness und Transparenz für gewerbliche Nutzer von Online-Vermittlungsdiensten und zur Änderung weiterer Gesetze, Referentenentwurf des Bundesministeriums für Digitales und Verkehr, Bearbeitungstand: 01.08.2023. https://bmdv.bund.de/SharedDocs/DE/Anlage/Gesetze/Gesetze-20/gesetz-durchfuehrung-verordnung-binnenmarkt-digitale-dienste.pdf?__blob=publicationFile [Letzter Zugriff: 22.12.2023].

vertreten sind. Der Ausschuss hat hauptsächlich eine beratende Funktion, kann aber auch Verfahren einleiten (Jaursch, 2022).

Der Erfolg und damit die Stärke und Durchsetzungskraft der neuen europäischen Plattformaufsicht hängt nicht nur von der Kommission ab, sondern in hohem Maße auch von der Ausgestaltung der DSCs. Jaursch (2022) hebt einige Herausforderungen hervor. Mehr als jedes andere Gesetz zuvor legt der DSA großen Wert auf Datenanalysen und Berichtspflichten. Eine Fülle von Daten wird generiert. Obwohl die deutschen Behörden Erfahrungen mit Markt- und Datenanalysen haben, müssen diese Strukturen im DSC deutlich ausgebaut und auf die Plattformen zugeschnitten werden. So sollte der DSC zu einer zentralen Stelle für Plattformforschung werden, der eigene Studien durchführt und Forschung in Auftrag gibt. Darüber hinaus erfordert der DSA im Vergleich zu anderen EU-Regelungen eine sehr viel stärkere Zusammenarbeit zwischen sehr unterschiedlichen Behörden wie Datenschutzbehörden, Kartellbehörden oder Medienaufsichtsbehörden. Der DSC muss nun mit all diesen Behörden gleichzeitig zusammenarbeiten. Die Einrichtung fallbezogener interdisziplinärer Projektteams erscheint sinnvoll (Jaursch, 2022).

Es ist klar, dass neben der Auswahl der zuständigen Behörden und des DSC der rasche Aufbau funktionierender Strukturen eine besondere Herausforderung sein wird. Zu diesem Zweck muss die Bundesregierung ein Gesetz zur Benennung des DSC auf den Weg bringen, in dem die wesentlichen Elemente festgelegt werden (Jaursch, 2022).

- *Unabhängigkeit*: Der DSC sollte eine unabhängige Einrichtung sein, die keinen Weisungen unterliegt. Er sollte über eine angemessene Personal- und Mittelausstattung verfügen. Seine Arbeit sollte für die Industrie und die Öffentlichkeit transparent sein, und es sollten geeignete Kontaktstellen eingerichtet werden.
- *Arbeitsmethoden*: Die Arbeitsweise muss fallbezogen und je nach Thema flexibel sein und sich durch einen engen Austausch zwischen den zuständigen Behörden auszeichnen. Dies erfordert geeignete Informationssysteme.
- *Fachwissen*: Für die Arbeit in Teams sind interdisziplinäre und interkulturelle Fähigkeiten erforderlich. Die Zusammenarbeit ist sowohl auf nationaler Ebene zwischen den Behörden als auch auf EU-Ebene mit der Kommission und dem Verwaltungsrat gemeint. Um Fachwissen aufzubauen, sollten Kontakte mit externen Experten und der wissenschaftlichen Gemeinschaft gepflegt und ein kontinuierlicher Austausch angestrebt werden.
- *Datenanalyse und Forschung*: Der DSC sollte als zentrale und koordinierende Stelle für alle DSA-bezogenen Datenanalysen angesehen werden, und die entsprechenden Aktivitäten sollten hier zusammenlaufen. Dementsprechend besteht die Notwendigkeit, Kapazitäten aufzubauen.

4.2.2 Prüfung: Einhaltung des DSA und Identifizierung vom Problemen

Prüfungen (nachfolgend: Audits oder Auditing) können von verschiedenen Parteien durchgeführt werden und sich auf unterschiedliche Quellen stützen. Der DSA berücksichtigt zwei verschiedene Arten von Prüfungen.

Bei der ersten Art von Audits wird geprüft, ob die Online-Plattformen der Verpflichtungen des DSA nachkommt. Die zweite Art befasst sich mit der Evaluierung systemischer Risikien, die durch die Praktiken von Online-Plattformen wie ihren Moderationsansätze für Online-Inhalte entstehen können. Hierfür sieht der DSA sieht verschiedene Audits und Bewertungen von VLOPs vor. Einige davon sollen von den Plattformen selbst durchgeführt werden, während andere von externen, beauftragten und unabhängigen Stellen bewertet werden sollen. Dementsprechend wird unterschieden zwischen First-Party-Audits durch die VLOPs (Art. 34 & 35,DSA), Second-Party-Audits durch von den VLOPs beauftragte unabhängige Organisationen (Art. 37, DSA) und Third-Party-Audits durch den DSC und die Europäische Kommission (Art. 40, DSA) (Meßmer & Degeling, 2023).

Einige dieser Prüfungen und Bewertungen sind obligatorisch, andere freiwillig, für einige gibt es Durchführungs- oder delegierte Rechtsakte, und einige sollten Teil eines Verhaltenskodex sein. Daher wird eine breite Palette unterschiedlicher Perspektiven berücksichtigt werden. Vor diesem Hintergrund wird eine Vielzahl von Akteuren mit der Analyse, Bewertung und Prüfung der Inhaltsmoderation betraut werden.

Die Prüfungen können einerseits rein auf der Basis der von den Anbietern zur Verfügung gestellten Daten erfolgen, andererseits können sie, insbesondere bei Zweit- und Drittprüfungen, auch auf zusätzlich gewonnenen und genutzten Daten beruhen. Um die Auswirkungen der Digitalisierung im Allgemeinen und die Auswirkungen weitreichender gesetzgeberischer Initiativen im Bereich der Digitalpolitik auf Wirtschaft und Gesellschaft zu analysieren und mögliche Handlungs- und Gestaltungsfelder zu identifizieren, bedarf es einer verlässlichen, umfassenden und konsistenten Datenbasis, die Ausgangspunkt für fundierte, evidenzbasierte Analysen und planvolle Entscheidungshilfen sein kann. Der Aufbau solcher geeigneter Datenbanken trägt zu einem evidenzbasierten Monitoring und Auditing bei, erfordert aber Zeit und Kontinuität.

Die Europäische Kommission hat zu diesem Zweck das Europäische Zentrum für Algorithmische Transparenz (Englisch: **European Centre for Algorithmic Transparency,** ECAT) eingerichtet. Es ist bei der Gemeinsamen Forschungsstelle (Englisch: Joint Research Centre, JRC) angesiedelt. [17] Die EU-Beobachtungsstelle für die Online-Plattform-Ökonomie soll Trends und Daten in der Online-Plattform-Ökonomie beobachten und analysieren und hat daher eine flankierende Funktion.[18]

17 https://algorithmic-transparency.ec.europa.eu/index_en [letzter Zugriff 17.07.2023].
18 https://digital-strategy.ec.europa.eu/de/policies/eu-observatory-online-platform-economy [letzter Zugriff 17.07.2023].

Eine systematische und strukturierte Erfassung von internetbasierten Dienstleistungen, sowohl auf der Unternehmens- als auch auf der Dienstleistungsebene, hat in der EU und in Deutschland bisher nicht stattgefunden. Genau dies ist notwendig, um eine umfassende Informationsbasis und ein fundiertes Verständnis von Internetplattformen und dem bestehenden Ökosystem zu entwickeln. Das WIK hat in den letzten Jahren mit dem Projekt DOTT (Datenbank für OTT-Dienste) begonnen, sich dieser Thematik anzunehmen und eine Datenbank aufzubauen, die sich auf verschiedene Segmente von Online-Diensten konzentriert. Das Projekt wird mit einem stärkeren Fokus auf DSA-bezogene Indikatoren fortgesetzt.[19]

Insbesondere für die Untersuchung der Moderationsprozesse und der Frage, ob sich aus ihnen Risiken ergeben, lassen sich grundsätzlich mehrere Instrumente einsetzen. Erstens sind sie unabhängig davon, ob sie für First-Party-, Second-Party- oder Third-Party-Audits verwendet werden. In der nachstehenden Tabelle 4-4 sind einige der gebräuchlichsten Prüfungstechniken für algorithmische Systeme aufgeführt.

Tabelle 4-4: Instrumente des Auditing

Audit Methode	Beschreibung	Ziel	Herausforderungen
Code Audit	Auditoren haben direkten Zugriff auf die Codebasis des zugrundeliegenden Systems oder auf "Pseudocode"-Beschreibungen in Klartext, was der Code tut.	Verstehen der Absichten von Algorithmen; im Falle des maschinellen Lernens nützlich, um zu verstehen, welche Ziele optimiert werden.	Codebasen können riesig sein - einzelne Ingenieure in großen Unternehmen verstehen selten, wie alle Teile der Plattform funktionieren. Es ist schwer, Auswirkungen/Ergebnisse durch den Code zu erkennen. Bedenken hinsichtlich IP und Sicherheit.
Nutzerbefragung	Die Prüfer führen eine Umfrage durch und/oder befragen die Benutzer, um beschreibende Daten über die Erfahrungen der Benutzer auf der Plattform zu sammeln.	Sammeln von Informationen über die Nutzererfahrungen auf einer Plattform - um ein grobes Bild der Arten von problematischem Verhalten zu zeichnen, das dann weiter untersucht werden kann.	Anfällig für die üblichen sozialwissenschaftlichen Bedenken bei Umfragen - Druck, auf eine bestimmte Art und Weise zu antworten, unzuverlässiges menschliches Gedächtnis und die Schwierigkeit, den Ergebnissen einen kausalen Zusammenhang zuzuordnen.
Scraping Audit	Auditoren sammeln Daten direkt von einer Plattform, in der Regel durch das Schreiben von Code, der automatisch auf eine Webseite klickt oder durch sie scrollt, um Daten von Interesse zu sammeln (z.B. Text, den Benutzer posten).	Verstehen von Inhalten, wie sie auf der Plattform dargestellt werden - insbesondere das Treffen von beschreibenden Aussagen (z.B. "dieser Anteil der Suchergebnisse enthielt diesen Begriff") oder der Vergleich von Ergebnissen für verschiedene Gruppen oder Begriffe.	Erfordert die Entwicklung eines benutzerdefinierten Tools für jede Social-Media-Plattform, was sehr flüchtig sein kann, da kleine (legitime) Änderungen am Layout einer Website das Programm zerstören können.

19 https://dott.wik.org/.

Audit Methode	Beschreibung	Ziel	Herausforderungen
API Audit	Prüfer greifen über eine von der Plattform bereitgestellte programmatische Schnittstelle auf Daten zu, die es ihnen ermöglicht, Computerprogramme zu schreiben, um Informationen an eine Plattform zu senden und von ihr zu empfangen. Eine API könnte es beispielsweise einem Benutzer ermöglichen, einen Suchbegriff zu senden und die Anzahl der Beiträge zurückzubekommen, die diesem Suchbegriff entsprechen.	Leichterer programmatischer Zugriff auf Daten als bei einer Scraping-Prüfung - dies ermöglicht eine einfachere Automatisierung der Erfassung für beschreibende Aussagen oder vergleichende Arbeiten.	Öffentlich verfügbare APIs liefern einer Regulierungsbehörde möglicherweise nicht die Daten, die sie benötigt. Mit ihren Befugnissen zum Sammeln von Informationen könnte sie eine Plattform dazu zwingen, Zugang zu weiteren APIs oder sogar zu einer benutzerdefinierten API zu gewähren, aber das kann zusätzliche technische Arbeit der Plattformen erfordern.
Sock puppet Audit	Die Prüfer verwenden Computerprogramme, um sich als Nutzer der Plattform auszugeben (diese Programme werden "Sockenpuppen" genannt). Die Daten, die von der Plattform als Reaktion auf die programmierten Benutzer erzeugt werden, werden aufgezeichnet und ausgewertet.	Verstehen, was ein bestimmtes Benutzerprofil oder eine Gruppe von Benutzerprofilen auf einer Plattform erleben kann.	Sockenpuppen geben sich nur als Benutzer aus - sie sind keine echten Benutzer und daher bestenfalls ein Stellvertreter für individuelle Benutzeraktivitäten und Erfahrungen.
Crowd sourced Audit	Ein Crowd-sourced Audit (manchmal auch als "mystery shopper" bekannt) nutzt echte Nutzer, die Informationen von der Plattform sammeln, während sie sie nutzen - entweder durch manuelle Erfahrungsberichte oder durch automatisierte Mittel wie eine Browsererweiterung	Beobachten Sie, welche Inhalte die Nutzer auf einer Plattform erleben und ob verschiedene Nutzerprofile unterschiedliche Inhalte erleben.	Erfordert einen individuellen Ansatz zur Datenerfassung für jede zu prüfende Medienplattform, der oft auf Web-Scraping-Techniken beruht; bisher nur auf Desktop und nicht auf Mobilgeräten demonstriert, so dass die Ergebnisse verzerrt oder mobile Erfahrungen übersehen werden können.

Quelle: Ada Lovelace Institute (2021).

KIVI, das System der Landesmedienanstalten NRW, kann im weitesten Sinne als Scanning Audit Tool verstanden werden. Es dient der Erkennung von Rechtsverletzungen im Internet und basiert auf kommerziellen Content-Moderationssystemen. So verwendet KIVI Amazon Recognition, um pornografische Inhalte zu identifizieren (Meineck et al., 2022, Kleinz, 2022). Es liegt nahe, Systeme die für die Moderation verwendet werden, auch für das Auditing einzusetzen. Auch wenn sie auf die Moderation von Inhalten ausgerichtet sind, ist der Umfang der Überwachung und Kontrolle sehr ähnlich. Die folgende Liste gibt einen kurzen Überblick über die derzeit auf dem Markt befindlichen Systeme.

- *Hive-Moderation*[20]
- *Amazon Rekognition*[21]

20 https://hivemoderation.com/ [letzter Zugriff 28.03.2023].
21 https://aws.amazon.com/de/rekognition/ [letzter Zugriff 28.03.2023].

- *WebPurify:*[22]
- *Sightengine*[23]
- *Alibaba Cloud*[24]
- *Azure Content Moderator*[25]

Bei der Nutzung dieser Marktlösungen zu Prüf- und Überwachungszwecken werden jedoch Ansätze verwendet, die auch von den Plattformen genutzt werden. Einerseits kann argumentiert werden, dass diese Ansätze dem Stand der Technik entsprechen und es daher keinen Sinn macht, parallel zusätzliche Lösungen zu entwickeln, die ausschließlich für Auditing verwendet werden. Andererseits muss aber auch die kritische Frage gestellt werden, ob die Plattformen oder Systeme sich selbst auditieren und der zu erwartende Erkenntnisgewinn begrenzt ist. Es ist jedoch zu erkennen, dass die Lösungen verschiedene Prüfungsinstrumente abdecken und von unterschiedlichen Anbietern angeboten werden. Dementsprechend sollte es möglich sein, für viele Bereich andere Instrumente für das Monitoring zu verwenden, als bei der Moderation eingesetzt wurden. Aber nicht nur die Betrachtung von Input und Output kann zum Verständnis der Systeme beitragen, sondern auch die Betrachtung der Trainingsdaten (Singh, 2019).

Unabhängig davon, wer die Audits durchführt, ist Audit-Washing ein großes Problem. Audit-Washing ist eine Bedrohung für die Integrität von Audits und bedeutet, dass sich Unternehmen durch die Durchführung eines Audits selbst reinwaschen, wenn dieses nicht bestimmten Standards und Kontrollen unterliegt. Ein Bericht des GMF kommt zu dem Schluss, dass schlecht konzipierte oder durchgeführte Prüfungen im besten Fall bedeutungslos sind und im schlimmsten Fall von dem Schaden, den sie mindern sollen, ablenken oder ihn sogar entschuldigen können. Um diesen Risiken zu begegnen, werden in dem Papier die folgenden Punkte genannt, die beachtet werden müssen, damit das Instrument zuverlässig ist. Das "Wer" der Prüfungen umfasst die Person oder Organisation, die die Prüfung durchführt, mit klar definierten Qualifikationen, Datenzugriffsbedingungen und internen Prüfungsleitplanken. Das "Was" umfasst die Art und den Zweck der Prüfung, einschließlich ihrer Einordnung in ein größeres sozio-technisches System. Das "Warum" umfasst die Ziele der Prüfung, seien es eng definierte Rechtsnormen oder umfassendere ethische Ziele, die für den Vergleich von Prüfungen wesentlich sind. Das "Wie" schließlich umfasst eine klare Formulierung von Prüfungsstandards, die eine wichtige Grundlage für die Entwicklung von Prüfungszertifizierungsmechanismen und den Schutz vor Prüfungswäsche bilden. (Goodman & Tréhu, 2022)

22 https://www.webpurify.com/ [letzter Zugriff 28.03.2023].
23 https://sightengine.com/ [letzter Zugriff 28.03.2023].
24 https://www.alibabacloud.com/help/en/content-moderation/latest/product-introduction [letzter Zugriff 28.03.2023].
25 https://azure.microsoft.com/de-de/pricing/details/cognitive-services/content-moderator/ [letzter Zugriff 28.03.2023].

5 Schlussfolgerung

Heutige Moderationsprozesse sind nicht frei von Ungenauigkeit und Intransparenz.

Die derzeitigen Fähigkeiten automatisierter Systeme, die zur Moderation von Inhalten eingesetzt werden, tragen wesentlich zu ersterem bei. So ist es für viele Algorithmen, die von diesen Systemen verwendet werden, nach wie vor schwierig, eine Fehlklassifizierung von Inhalten vollständig zu vermeiden. Dies hat zur Folge, dass harmlose Inhalte als schädlich, illegal oder anderweitig regelwidrig eingestuft und moderiert werden und umgekehrt. Dies ist auf verschiedene Faktoren zurückzuführen: 1) den Entwicklungs- und Trainingsprozess, einschließlich der Qualität der Trainingsdaten; 2) die mangelnde Fähigkeit der Systeme, den Kontext zu interpretieren; und 3) die Möglichkeiten, die Systeme zu umgehen.

Die grundsätzliche Fehleranfälligkeit führt dazu, dass menschliche Eingriffe durch Moderatoren notwendig bleiben. Dies geht jedoch mit einem hohen Ressourcenaufwand für die Plattformanbieter und einer psychischen Belastung der ausführenden Moderatoren einher. Hinzu kommt, dass es auch bei ihnen immer wieder zu Fehlentscheidungen und Inkonsistenzen kommt.

Das Ergebnis der Moderation ist jedoch nicht ausschließlich auf die Fähigkeiten der eingesetzten Instrumente zurückzuführen, sondern auch auf die Moderationspolitik der Plattformen. Wissenschaftliche Studien zeigen, dass insbesondere webbasierte Plattformen, die ausschließlich auf Gewinnmaximierung ausgerichtet sind, eine laxe Moderationspolitik verfolgen. Jede Moderationsentscheidung ist eine Abwägung zwischen dem Verlust des Engagements einiger Nutzer, die von der Moderation betroffen sind, und der Zunahme des Engagements der Nutzer, die die Moderation gutheißen würden. Je heterogener die Präferenzen der Nutzer eines Dienstes in Bezug auf Inhaltsmoderation und schädliche oder illegale Inhalte sind, desto schwieriger wird es, eine einheitliche Politik zu entwickeln. Ebenso hätten sie kein Interesse in bessere Systeme zu investieren, um stets fehlerfreie Ergebnisse zu erzielen.

Die Intransparenz wird auch durch zwei Faktoren begünstigt. Zum einen sind die Moderationsprozesse intransparent, weil die Plattformen Geschäftsgeheimnisse und kommerzielle Interessen wahren. Dies betrifft insbesondere automatisierte Systeme, deren Intransparenz sie davor schützen soll, von Dritten umgangen oder ausgenutzt zu werden. Der zweite Faktor ist die mangelnde Fähigkeit von Außenstehenden, diese automatisierten Systeme in ihrer Gesamtheit zu verstehen. Selbst die Entwickler, die an den Systemen arbeiten, sind manchmal nicht in der Lage, die Funktionsweise der Systeme vollständig zu erfassen. Aus diesem Grund haben sich einige Wissenschaftler dafür ausgesprochen, dass eine reine Veröffentlichung des Codes von automatisierten Systemen nicht ausreicht, um Transparenz zu schaffen. Vielmehr seien zusätzliche relevante Informationen und Erläuterungen erforderlich.

Aber nicht nur die Funktionsweise der automatisierten Systeme ist intransparent, sondern auch die zugrunde liegenden Moderationsrichtlinien der Plattformanbieter. So beschreiben Díaz & Hecht-Felella (2021), dass Plattformen ihr Handeln mit ständig neuen, punktuell verkündeten Regeln begründen.

Diese Mischung aus Fehleranfälligkeit und Intransparenz erschwert es, Moderationsentscheidungen nachzuvollziehen und zu differenzieren, ob der rein technische Moderationsprozess oder die Moderationspolitik oder eine Kombination aus beiden, die sich gegenseitig bedingen können, für die getroffenen Moderationsentscheidungen verantwortlich sind, ob daraus Risiken für die Gesellschaft und den Einzelnen entstehen, die gemeinsam gemindert werden müssen, und wo die entsprechenden Verantwortlichkeiten liegen. Dies ist insbesondere bei sehr großen Plattformen problematisch, da diese eine große Anzahl von Konsumenten erreichen und sowohl falsch negative als auch falsch positive Entscheidungen zu schwerwiegenden individuellen oder gesellschaftlichen Schäden führen können.

Dem trägt der DSA insbesondere durch zahlreiche Transparenzpflichten Rechnung, die die Prozesse nachvollziehbarer machen können. Diese Pflichten betreffen unter anderem die Allgemeinen Geschäftsbedingungen. Diese müssen laut DSA klare Informationen über die Moderation von Inhalten enthalten und Änderungen müssen den Nutzern mitgeteilt werden. Eine Situation, in der Moderationsentscheidungen aufgrund von AGB-Änderungen uneinheitlich über verschiedene Kanäle kommuniziert werden, wird dadurch vermieden. Die Einrichtung eines Beschwerde- und Meldesystems beseitigt ein weiteres Problem der Rechenschaftslegung, das sich aus den bisher eingeschränkten Möglichkeiten, insbesondere als Privatperson gegen die Plattformen vorzugehen, ergibt. Schließlich gibt es unter dem DSA Transparenzberichte und eine Transparenzdatenbank. Diese werden sowohl auf aggregierter als auch auf individueller Ebene Informationen und Daten zu Entscheidungen liefern, die zur Nachvollziehung von Entscheidungen, zur Übersicht über die Moderation und zur Identifizierung von Unregelmäßigkeiten genutzt werden können.

Und obwohl der DSA keine Haftung für illegale Inhalte vorsieht und damit verhindert, dass Plattformen zu einer übermäßigen Moderation veranlasst werden, sieht der DSA vor, dass Plattformanbieter - insbesondere VLOPs - in der Verantwortung stehen, eine Risikobewertung durchzuführen und Maßnahmen zu ergreifen, um systemische Risiken zu mindern.

Darüber hinaus erlaubt der DSA der nationalen Datenschutzbehörde und befugten Forschern, Informationen und Daten einzuholen, die es ihnen ermöglichen, systemische Risiken zu identifizieren oder zu verstehen, die sich aus der Tätigkeit und damit auch aus der Moderationstätigkeit der Plattformanbieter ergeben. Hier stellt sich die Frage, wie umfangreich diese Daten sein werden, insbesondere vor dem Hintergrund, dass der Schutz personenbezogener Daten, vertraulicher Informationen und Geschäfts-

geheimnisse gewahrt werden muss, und welche Art der Prüfung möglich und sinnvoll sein könnte.

Wissenschaftler machen sich seit geraumer Zeit Gedanken, wie die Fähigkeiten von Algorithmen allgemein geprüft werden können. Hier gibt es unterschiedliche Ansätze, die von der Untersuchung der Trainingsdaten, über die Betrachtung von einzelnen Entscheidungen, der Analyse von Codes oder die Nutzung von Befragungen und Erfahrungsberichten. Da es sich bei der Moderation von Inhalten jedoch um einen soziotechnischen Prozess handelt, muss dieser auch als Ganzes betrachtet werden.

So hat die Landesmedienanstalt NRW zum Beispiel KIVI entwickelt, was als Scanning Audit-Tool verstanden werden kann. Es dient der Erkennung von Rechtsverletzungen im Internet und verwendet Amazon Recognition - ein System, das von vielen kleinen und großen Unternehmen zur Moderation von Inhalten eingesetzt werden kann und wird. Es liegt nahe, Systeme, die für die Moderation eingesetzt werden, auch für das Auditing zu verwenden, da im Prinzip die gleiche Leistung gefordert ist. Auf der einen Seite kann argumentiert werden, dass diese Systeme, wie z.B. Amazon Rekognition, dem aktuellen Stand der Technik entsprechen und es daher keinen Sinn macht, parallel zusätzliche Lösungen zu entwickeln, die ausschließlich für das Auditing eingesetzt werden. Es stellt sich jedoch die Frage, was erreicht werden soll, wenn solche Systeme, die zur Moderation eingesetzt werden, auch zur Identifikation von Fehlklassifikationen verwendet werden. Wenn mit KIVI eine Plattform auditiert wird, die bereits Amazon Recognition einsetzt, ist davon auszugehen, dass falsch positive Ergebnisse gar nicht erst sichtbar werden. Theoretisch sollte es auch keine falsch negativen Ergebnisse geben. Es sei denn, sie wurden in einem anderen Prozessschritt autorisiert. Oder es handelt sich um natürliche Abweichungen, denn im Idealfall sollten automatisierte Systeme bei jeder Moderation mitlernen, so dass ein Inhalt, der heute nicht moderationswürdig ist, es morgen sein wird. Die Praxis zeigt auch, dass die Entwicklung vor allem von prädiktiven Algorithmen, die in automatisierten Moderationssystemen eingesetzt werden, ständig vorangetrieben und verbessert wird.

Literaturverzeichnis

Ada Lovelace Institute (2021). Technical methods for the regulatory inspection of algorithmic systems in social media platforms. https://www.adalovelaceinstitute.org/report/ [Letzter Zugriff: 13.07.2023].

Aldahoul, N; Karim, H. A.; Momo, M. A. & Sy, M. A. G. (2023). Evaluation of Content Moderation Software for Nudity and Pornography Detection in Various Scenarios. MECON Multimedia University Engineering Conference, Malaysia. https://www.researchgate.net/publication/371805170_Evaluation_of_Content_Moderation_Software_for_Nudity_and_Pornography_Detection_in_Various_Scenarios [Letzter Zugriff: 30.10.2023].

Blodgett, S. & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. arXiv. https://arxiv.org/abs/1707.00061. [Letzter Zugriff: 25.03.2023].

Buiten, M. C. (2021). The Digital Services Act: From Intermediary Liability to Platform Regulation, JIPITEC, 12 (5),361- 370.

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability and Transparency, New York, NY, USA. Proceedings of Machine Learning Research: PMLR, 81, S. 77-91.

Burrell, Jenna (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. In: Big Data & Society, 3 (1), S. 1-12.

Barrett, P. M (2020). Who Moderates the Social Media Giants? A Call to End Outsourcing. NYU Stern. Center for Business and Human Rights.

Cambridge Consultants. 2019. Use of AI in Online Content Moderation. A Report produced on Behalf of Ofcom. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf [Letzter Zugriff: 07.02.2023].

Castets-Renard, C. (2020). Algorithmic content moderation on social media in EU law: Illusion of perfect enforcement. Journal of Law, Technology & Policy, 2020 (2), 283-323.

Cofone, Ignacio N. (2019). Algorithmic Discrimination Is an Information Problem. Hastings Law Journal, 70 (6), S. 1389-1444.

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. Computer Law & Security Review, 36, 105374.

De Streel, A.; Defreyne, E.; Jacquemin, H.; Ledger, M.; Michel, A.; Innesti, A.; Goubet, M.; Ustowski, D. (2020). Online Platforms' Moderation of Illegal Content Online. Law, Practices and Options for Reform. A study requested by the IMCO committee, Policy Department for Economic, Scientific and Quality of Life Policies, Directorate-General for Internal Policies, European Parliament.

Dias Olivia, T. (2020). Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression, Human Rights Law Review, 20 (4), 607-640.

Díaz, Á & Hecht-Felella, L. (2021). Double Standards in Social Media Content Moderation. Brennan Center for Justice at NYU School of Law. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation [Letzter Zugriff. 14.09.2023].

Digitale Gesellschaft e.V. (2020). Was sind Uploadfilter? https://digitalegesellschaft.de/wp-content/uploads/2020/12/DigitaleGesellschaft-Uploadfilter-Interaktiv-V04-NEU-111.pdf [Letzter Zugriff: 04.09.2023].

Duarte, N.; Llansó, E. & Loup, A. (2017). Mixed Messages? The Limits of Automated Social Media Content Analysis. cdt. https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf [Letzter Zugriff: 03.07.2023].

Elkin-Koren, N. & Perel, M. (2020). Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law, In: Frosio, G (hrsg.). Oxford Handbook of Online Intermediary Liability, Kapitel 34, 669-678.

Europäische Kommission (2023a). Questions and Answers: Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 [Letzter Zugriff: 10.07.2023].

Europäisches Parlament (2021). Liability of online platforms. Panel for the Future of Science and Technology. EPRS | European Parliamentary Research Service Scientific Foresight Unit (STOA). https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)656318 [Letzter Zugriff: 22.12.2023].

Farid, H. (2021). An Overview of Perceptual Hashing. Journal of Online Trust and Safety, 1(1), 1-22.

Gellert, R. & Wolters, P. (2021). The revision of the European framework for the liability and responsibilities of hosting service providers Towards a better limitation of the dissemination of illegal content. Radbound University. https://repository.ubn.ru.nl/bitstream/handle/2066/234104/234104.pdf?sequence=1&isAllowed=y [Letzter Zugriff 05.09.2023].

G'sell, F. (2023). The Digital Services Act: a General Assessment. In: von Ungern-Sternberg, A. (hrsg). Content Regulation in the European Union – The Digital Services Act. Schriften des IRDT – Trier Studies on Digital Law, 1, Verein für Recht und Digitalisierung e.V., Institute for Digital Law (IRDT), Deutschland: Trier.

GIFCT (2021). GIFCT Technical Approaches Working Group – Gap Analysis and Recommendations for deploying technical solutions to tackle the terrorist use of the internet. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf [Letzter Zugriff 05.09.2023].

Gillespie, T. (2018). Custodians of the Internet - Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.

Gillespie, T. (2020). Content Moderation, AI, and the question of scale. Big Data & Society, 7(2), 1-5.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. Social Media & Society, 8(3), 1-13.

Goodman, E.P. & Tréhu, J. (2022): AI Audit-Washing and Accountability, GMF Policy Paper, https://www.gmfus.org/sites/default/files/2022-11/Goodman%20%26%20Trehu%20-%20Algorithmic%20Auditing%20-%20paper.pdf [Letzter Zugriff 17.07.2023].

Gorwa, R.; Binns, R. & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7 (1), pp. 1-15.

Heldt, A. (2018). Intelligente Upload-Filter: Bedrohung für die Meinungsfreiheit? In Mohabbat Kar, R.; Thapa, B. E. P. & Parycek, P. (Hrsg.), (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft, 392-416.

Heller, B. (2019). Combating Terrorist-Related Content Through AI and Information Sharing. Transatlantic Working Group. https://www.ivir.nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf [Letzter Zugriff: 04.09.2023].

Hoffmann, A. & Gasparoti, A. (2020). Liability for illegal content online. cep. https://www.cep.eu/fileadmin/user_upload/hayek-stiftung.de/cepStudy_Liability_for_illegal_content_online.pdf [Letzter Zugriff: 14.09.2023].

Hubley, H. (2022). Bad Speech, Good Evidence: Content Moderation in the Context of Open-Source Investigations. International Criminal Law Review, 22, 989 -1015.

Husovec, M. (2023). The DSA's Scope Briefly Explained. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4365029 [Letzter Zugriff: 23.12.2023].

Innodata (2023). The Ethics of Content Moderation: Who Protects the Protectors? https://innodata.com/the-ethics-of-content-moderation/ [Letzter Zugriff: 08.09.2023].

Jaursch, J. (2022). Wie die deutsche Plattformaufsicht aufgebaut sein sollte Empfehlungen für einen starken „Digital Services Coordinator", Stiftung Neue Verantwortung. https://www.stiftung-nv.de/sites/default/files/snv_empfehlungen_fur_einen_starken_dsc.pdf [Letzter Zugriff: 13.07.2022].

Jiménez-Durán, R (2022a). The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. New Working Paper Series, No. 324, University of Chicago Booth School of Business, Stigler Center for the Study of the Economy and the State.

Jiménez-Durán, R (2022b). The Economics of Content Moderation on Social Media. ProMarket. https://www.promarket.org/2022/11/10/the-economics-of-content-moderation-on-social-media/ [Letzter Zugriff: 23.12.2023]

Jhaver, S.; Birman, I.; Gilbert, E. & Bruckman, A.(2019). Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. ACM Trans. Comput.-Hum. Interact. 26 (5), Artikel 31.

Jiang, J.A.; Nie, P.; Brubaker, J.R. & Fiesler, C. (2023). A Trade-off-centered Framework of Content Moderation. ACM Trans. Comput.-Hum. Interact. 30, 1, Artikel 3.

Kamara, S.; Knodel, M.; Llansó, E.; Nojeim, G.; Qin, L.; Thakur, D. & Vogus, C. (2021). Outside Looking In – Approaches to Content Moderation in End-to-End Encrypted Systems. cdt – Center of Democracy & Technology. https://iapp.org/media/pdf/resource_center/CDT_Report_Outside_Looking_In_Approaches_to_Content_Moderation_in_End_to_End_Encrypted_Systems.pdf [Letzter Zugriff: 23.06.2023].

Kayser-Brill, N. (2020). Google apologizes after its Vision AI produced racist results. AlgorithmWatch. https://algorithmwatch.org/en/google-vision-racism/ [Letzter Aufruf 20.11.2023].

Keller, S. & Leerssen, P. (2020) Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation. In: Persil, N. & Tucker J.A. (hrsg.). Social Media and Democracy: The State of the Field, Prospects for Reform. SSRC Anxieties of Democracy. Cambridge University Press, Cambridge, 220-251.

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. Harvard Law Review, 131 (6), 1599-1670.

Liu, Y.; Yildirim, P & Zhang, Z. J. (2022). Implications of Revenue Models and Technology for Content Moderation Strategies. Marketing Science, 41(4), 831-847.

Llansó, E.; van Hoboken, J.; Leerssen, P. & Harambam, J. (2020). Artificial Intelligence, Content Moderation, and Freedom of Expression. The Bellagio Session. 3th session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. https://www.ivir.nl/publicaties/download/TWG_Bellagio_papers_March_2020_full.pdf [Letzter Zugriff: 21.02.2023].

Liebe, A. & Wiewiorra, L. (2023). PLASMA Insight No. 1 - Data availability and data volume under the Digital Services Act. WIK. Bad Honnef. https://plasma.wik.org/static/PDF/PLASMA_Insight_No1_final.pdf [Letzter Zugriff: 24.12.2023].

Madiga, T. (2020). Reform of the EU liability regime for online intermediaries – Background on the forthcoming digital services act. EPRS – European Parliamentary Research Service. European Parliament. https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA(2020)649404_EN.pdf [Letzter Zugriff: 14.09.2023].

Madio, L. & Quinn, M. (2023). Content Moderation and Advertising in Social Media Platforms. University of Padova, Marco Fanno Working Papers, No. 297. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551103 [Letzter Zugriff: 03.07.2023].

Mahl, D.; Zeng, J. & Schäfer, M.S. (2023). Conceptualizing platformed conspiracism: Analytical framework and empirical case study of BitChute and Gab. New Media & Society, 0(0), 1-20.

Meineck, S.; Schmid, T. & Janus, P. (2022). KI in der Medienaufsicht: Was leistet das Tool KIVI?. https://www.bpb.de/lernen/digitale-bildung/werkstatt/513732/ki-in-der-medienaufsicht-was-leistet-das-tool-kivi/ [Letzter Zugriff: 29.03.2023].

Meßmer, A-K.; Degeling, M. (2023): Auditing Recommender Systems: Putting the DSA into practice with a risk-scenario-based approach.

https://www.stiftung-nv.de/sites/default/files/snv_auditing-recommender-systems_messmer_degeling.pdf [Letzter Zugriff 17.07.2023].

Molina, M. D. & Sundar, S. S. (2022). When AI moderates online content: effects of human collaboration and interactive transparency on user trust. Journal of Computer-Mediated Communication, 27 (4), 1-12.

Newton, C. (2029). The Trauma Floor. The secret lives of Facebook moderators in America. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [Letzter Zugriff: 19.12.2023].

Nieborg, D. B. & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. New Mesia & Society, 20(11), 4275-4292.

Ofcom (2022). Overview of Perceptual Hashing Technology. https://www.ofcom.org.uk/__data/assets/pdf_file/0036/247977/Perceptual-hashing-technology.pdf [Letzter Zugriff: 04.09.2023].

Perel, M. & Elkin-Koren, N. (2017). Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. Florida Law Review, 69 (1), 181-221.

Qiu, W. & Yuille, A. L. (2016). UnrealCV: Connecting Computer Vision to Unreal Engine. Computer Vision – ECCV 2016 Workshops. https://link.springer.com/chapter/10.1007/978-3-319-49409-8_75#citeas [Letzter Zugriff: 30.10.2023].

Quintais, J.P., Péter Mezei, István Harkai, João Carlos Magalhães, Christian Katzenbach, Sebastian Felix Schwemer, and Thomas Riis (2020). Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis", reCreating Europe Report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4210278 [Letzter Zugriff 06.07.2023].

Rauchfleisch, A. & Kaiser, J. (2021). Deplatforming the far-right: An analysis ofYouTube and BitChute. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867818 [Letzter Zugriff: 20.23.2023].

Reuber, A. R. & Fischer, E. (2022). Relying on the engagement of others: A review of the governance choices facing social media platform start-ups. International Small Business Journal: Researching Entrepreneurship, 40 (1), 3-22.

Roberts, S.T. (2017). Content Moderation. In: Schintler, L.A. &McNeely, C.L. (hrsg). Encyclopedia of Big Data. Schweiz: Cham.

Sander, B. (2020). Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation. Fordham International Law Journal, 43 (4), 939-1006.

Sartor G.; Sartor, G. & Loreggia, A. (2020). The impact of algorithms for online content filtering or moderation – Upload filters. Policy Department for Citizens' Rights and Constitutional Affairs, Directorate-General for Internal Policies, European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf [Letzter Zugriff: 07.02.2023].

Schmid, G. & Köhler, P. (2022). Der Digital Services Act – ein Überblick. https://www.taylorwessing.com/de/insights-and-events/insights/2022/11/digital-services-act-ein-ueberblick [Letzter Zugriff 12.07.2023].

Schwemer, S. F. (2022). Digital Services Act: A reform of the e-Commerce Directive and much more. prepared for A Savin, Research Handbook on EU Internet Law (2022). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213014 [Letzter Zugriff 22.12.2023].

Shenkman, C.; Thakur, D. & Llansó, E. (2021). Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis. cdt. https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf [Letzter Zugriff: 03.07.2023].

Singh, S. (2019). Everything in Moderation – An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. Open Technology Institute. New America. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/ [Letzter Zugriff: 17.02.2013].

Snow, Jacob (2018). Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. ACLU. https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falselymatched-28 [Letzter Zugriff: 21.03.2023].

Steinebach, M. (2023). An Analysis of PhotoDNA. Proceedings of the 18th International Conference on Availability, Reliability and Security, 44, 1–8. https://dl.acm.org/doi/abs/10.1145/3600160.3605048 [Letzter Zugriff: 03.07.2023].

Statista (2023). Media usage in an internet minute as of April 2022 . https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/ [Letzter Zugriff: 19.12.2023].

Swart, Joëlle (2021). Experiencing Algorithms: How Young People Understand, Feel About, and Engage With Algorithmic News Selection on Social Media. Social Media & Society, 7 (2), S. 2056305121100.

Suzor, N.P.; West, S.M. & Quodling, A. (2019). International Journal of Communication, 13, 1526-1543.

Taş, S. & Wiewiorra, L. (2022). Nachvollziehbarkeit und Kontrolle algorithmischer Entscheidungen und Systeme. WIK Kurzstudie. WIK. https://www.wik.org/fileadmin/files/_migrated/news_files/WIK_Kurzstudie_Algorithmen.pdf [Letzter Zugriff: 03.07.2023].

Tobin, A.; Varner, M. & Angwin, J. (2017) Facebook's uneven enforcement of Hate Speech rules allows vile posts to stay up. ProPublica. https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes [Letzter Zugriff: 14.09.2023].

Whittaker, Z. (2020). Facebook to pay $52 million to content moderators suffering from PTSD. TechCrunch. https://techcrunch.com/2020/05/12/facebook-moderators-ptsd-settlement/?guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABarYKyK9Tffyf_B5xhKMbI9-bNfX4iqWr0-p8K5ShmB_d5e1-zaR4NHJzxIAZOFX2cYGzNDmLacKmxJaWAPilnUZvD49dwNcrlhBPQdKEEn3lbmZKuXYoKjhKxO-

MsR6QN2R7OJxxSTSX4E4gtKuuReR56JCfizi2CQRZLMfxIC&guccounter=2 [Letzter Zugriff: 19.12.2023].

Wilman, F. (2022). The Digital Services Act (DSA) – An Overview. SSRN. http://dx.doi.org/10.2139/ssrn.4304586 [Letzter Zugriff: 22.12.2023].

Wilson, R.A. & Land, M. (2021). Hate Speech on Social Media: Content Moderation in Context. Connecticut Law Review, 52(3), 1029-1242.

Yildirim, P & Zhan, Z. J. (2022). How Social Media Firms Moderate Their Content. Knowledge at Wharton. https://knowledge.wharton.upenn.edu/article/social-media-firms-moderate-content/#:~:text=John%20Zhang%2C%20and%20Wharton%20doctoral,the%20latter%20when%20it%20does [Letzter Zugriff: 14.09.2023].

Yuille, A. L. & Liu, C. (2019). Limitations of Deep Learning for Vision, and How We Might Fix Them. The Gradient. https://thegradient.pub/the-limitations-of-visual-deep-learning-and-how-we-might-fix-them/ [Letzter Zugriff: 30.10.2023].

ISSN 1865-8997