regulierung dies- und jenseits des Atlantiks zu beschreiben und zu plausibilisieren. Die Gesetzgebung hinsichtlich der Äußerungsregulierung ist in den Vereinigten Staaten umstritten, dennoch bzgl. des Äußerungsregimes im Netz seit über einem Vierteljahrhundert stabil – »Section 230 is the glue that holds the Internet–as we know it today—together.«[121] Neben dem Haftungsausschluss für Inhalte Dritter, werden die Plattformbetreiber:innen durch § 230 (c) (2) (A)—(B) CDA unmittelbar zur *Content Moderation* autorisiert.[122] Inhaltliche Vorgaben macht die US-amerikanische Gesetzgebung nicht, da dies gegen den Schutz des *First Amendments* verstoßen würde. Allerdings formuliert der CDA Vorstellungen, was der Gegenstand von *Content Moderation* sein könnte:

> »No Provider or user of an interactive computer service shall be held liable on account of— (A) any action voluntarily taken in good faith to restrict access or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected.«[123]

Die Autorisierung der Plattformen, die implizierte Aufforderung gegenüber ihnen und der Haftungsausschluss sind Formen der Regulierung neuer Schule und führen zu einer unregulierten Selbstregulierung der Plattformen.

Die Gesetzgebung in Deutschland und der Europäischen Union beschreitet genau wie die US-Gesetzgebung den Weg der Regulierung neuer Schule, setzt dabei aber zugleich auf immer stärker regulierte Selbstregulierung. Um die Beschaffenheit der regulierten Selbstregulierung und die Gründe jenseits der gesetzlichen Vorgaben zur Moderation von Inhalten geht es im nachfolgenden Abschnitt.

6.2 Content Moderation

Die Gesetzgebung in Deutschland, der EU und den USA hat sich im Rahmen von Äußerungsregulierung neuer Schule dafür entschieden, den Plattformbetreiber:innen Mitverantwortung für die Regulierung bzw. Moderation von Inhalten aufzuerlegen. Dabei gibt es erhebliche transatlantische Unterschiede in der Ausgestaltung der Moderationspflichten und bei der Pflicht zur Berücksichtigung von Grundrechten.

In den USA sind die Plattformen durch den CDA dazu aufgefordert, Inhaltsmoderation zu betreiben, was sich jedoch nicht durch verpflichtende Gesetzgebung ausdrückt, sondern vielmehr in der Abwesenheit von staatlichen Sanktionen, wenn die Plattformen moderieren. In der Summe lässt sich von einer grundsätzlichen Entscheidung für die *unregulierte Selbstregulierung* und für eine *größtmögliche Staatsfreiheit* bei der Regulierung von Äußerungen auf digitalen Plattformen sprechen.

---

121 Bastian (2022). *Content Moderation Issues Online*, S. 49.

122 Vgl. Mikaelyan, Yeva (2021). *Reimagining Content Moderation: Section 230 and the Path to Industry-Government Cooperation*, in: *Loyola of Los Angeles Entertainment Law Review* 41 (2), S. 179–214, hier: S. 185.

123 § 230 (c) (2) (A) CDA.

In Deutschland und der Europäischen Union gibt es dagegen von Beginn an eine *Ad-rem*-Regulierung, die Menschenwürde und Persönlichkeitsrechte schützen soll. Der Staat darf hier generell wesentlich schneller begrenzend in die Meinungsäußerungsfreiheit eingreifen, um andere Grundrechtsgüter zu schützen. In jüngerer und jüngster Vergangenheit haben sich sowohl die deutsche als auch die europäische Legislative zur Äußerungsregulierung neuer Schule entschieden. Den Anfang machte 2017 das bereits 2021 nachgebesserte deutsche NetzDG, 2024 gefolgt vom DSA auf EU-Ebene. NetzDG und DSA verpflichten die Plattformen zu einer *regulierten Selbstregulierung*, das heißt, dass es gesetzliche Vorgaben für die Gestaltung und die Inhalte der *Content Moderation* gibt.

Bevor es um die unterschiedlichen Systeme der *Content Moderation* geht, wird auf einer theoretischen und allgemeinen Ebene auf die *Content Moderation* und ihre Bedeutung geblickt. *Content Moderation* ist auch ganz losgelöst von jeglicher staatlichen Regulierung ein wesentlicher Bestandteil digitaler Plattformen und somit der Gestaltung von Räumen der Meinungsäußerungsfreiheit. Um dies zu verdeutlichen, geht es in den nächsten Schritten darum zu zeigen, dass *Content Moderation* ein Plattformmerkmal *sui generis* ist, *warum* und *wie* Plattformen moderieren. Ein Schwerpunkt ist dabei der Blick auf die automatisierte Moderation. Hiernach geht es um die Gestaltung und den Einfluss von Gesetzgebung auf *Content Moderation* in den Vereinigten Staaten sowie in Deutschland und der EU.

Content Moderation als Plattformmerkmal sui generis

Content Moderation ist das zentrale Merkmal digitaler Plattformen, wie Tarleton Gillespie herausarbeitet:

> »Our thinking about platforms must change. It is not just that all platforms moderate, nor that they have to moderate, nor that they tend to disavow it while doing so. It is that moderation, far from being occasional or ancillary, is in fact an essential, constant, and definitional part of what platforms do. Moderation is the essence of platforms. It is the commodity they offer. It is their central value proposition.«[124]

Durch ihre Programmierung, Design- bzw. Code-Entscheidungen und das Erstellen von Handlungsoptionen werden Affordanzen für Äußerungen geschaffen und ihre Bemühungen, *Mission Statements*, *Community Standards*, *Netiquette Regeln* o.Ä. einzuführen, um- und durchzusetzen, sind wesentlich für das Erscheinungsbild der jeweiligen Plattform. Plattformbetreiber:innen haben ein unternehmerisches Interesse daran, dass ihre Aktivitäten für ihre Kunden (v.a. Werbetreibende) und damit mittelbar für die Nutzer:innen attraktiv sind. Daher haben sie ein Interesse daran, »störende Inhalte entfernen zu können, auch wenn diese nicht rechtswidrig sind.«[125]

Der Charakter einer Website, eines Webangebots oder einer Plattform entsteht also erst durch ihre Moderationsentscheidungen und so ist es nicht verwunderlich, dass Yifat Nahmias und Maayan Perel folgendes feststellen:

124 Gillespie (2018). *Platforms Are Not Intermediaries*, S. 201.

125 Vgl. König (2019). *Vertragliche Gestaltung der Meinungsfreiheit in sozialen Netzwerken*, S. 631.

> »Content Moderation can be defined as the organized practice of screening online content based on the characteristics of the website, its targeted audience, and jurisdictions of user-generated content to determine whether such content is appropriate.«[126]

Die Moderationsentscheidungen werden, wie das Zitat zeigt, nicht einseitig von den Plattformen getroffen, sondern hängen ebenso von externen Faktoren wie der Rechtslage und von internen Faktoren wie den Bedürfnissen der anvisierten Zielgruppe ab. Dabei zeigt eine Literaturauswertung Lucas Wrights, dass es egal ist, ob die Moderation durch gewerbliche Anbieter:innen oder durch die »Community«, also durch freiwillige Moderator:innen erfolgt.[127] Der Charakter oder *Geist der Plattform* wird durch die Moderationspraxen bestimmt.

6.2.1 Warum moderieren Plattformen?

Plattformen haben ein genuines Interesse daran, *Content Moderation* im Sinne einer aktiven Entscheidung zu betreiben, um Inhalte zu blockieren, zu löschen oder beizubehalten. Dieses Interesse entwickelte sich bereits vor dem gesetzlichen Zwang und der entsprechenden Rechtsprechung, die in vielen Jurisdiktionen der Welt besteht.

Bei Plattformen handelt es sich überwiegend um Kommunikations- und Austauschräume, welche mit dem Zweck der Gewinnerzielung allgemein verfügbar gemacht werden. Wenn Plattformen sperren oder löschen, tun sie das in der Regel, um ihre Haftungsrisiken zu verringern, andere Nutzer:innen zu schützen und um die Beziehung zu den Werbekund:innen nicht zu gefährden.[128] Illegale, unerlaubte, hasserfüllte und sonstige unerwünschte Inhalte können die Marke und den Ruf der jeweiligen Plattform beschädigen. Nutzer:innen und Kund:innen könnten sich von der Plattform zurückziehen, wenn nicht ausreichend moderiert wird. Gute und effektive Inhaltsmoderation führt dagegen zu erhöhten Interaktionsraten und aktiveren Nutzer:innen.[129] In Bezug auf die Regulierung von Äußerungen lässt sich dies ebenso feststellen und zugleich konkretisieren: »These platforms are both the architecture for publishing new speech and the architects of the institutional design that governs it.«[130]

Kate Klonick hat in einer Studie aus dem Jahr 2018 die Entwicklung der *Content Moderation* aus einer US-amerikanischen Perspektive herausgearbeitet.[131] Viele Ergebnisse sind aufgrund der globalen Ausrichtung der untersuchten Plattformen verallgemeinerbar. In ihrer Untersuchung stellt sie nicht nur die rechtlichen Entwicklungen, Problem-

126 Nahmias, Yifat & Perel, Maayan (2021). *The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations*, in: *Harvard Journal on Legislation* 58 (1), S. 145–194, hier: S. 171.

127 Vgl. Wright, Lucas (2022). *Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator*, in: *Social Media + Society* 8 (1), S. 1–11, hier: S. 3.

128 Vgl. König (2019). *Vertragliche Gestaltung der Meinungsfreiheit in sozialen Netzwerken*: S. 631.

129 Vgl. Nahmias & Perel (2021). *The Oversight of Content Moderation by AI*, S. 174; Klonick, Kate (2018). *The New Governors: The People, Rules, and Processes Governing Online Speech*, in: *Harvard Law Review* 137, S. 1598–1670, hier: S. 1627.

130 Klonick (2018). *The New Governors*, S. 1617.

131 Vgl. ebd.

und Fragestellungen sowie die wissenschaftliche Debatte dar, sondern geht auch – und das ist an dieser Stelle entscheidend – auf die Entwicklung, Etablierung und Professionalisierung der *Content Moderation* bei *YouTube*, *Facebook* und *Twitter* als dominanten Plattformen hinsichtlich des Austauschs von Inhalten und Äußerungen auf globaler Ebene ein.[132] Im Zuge ihrer Arbeit hat sie Interviews mit den Verantwortlichen *Policy-Gestalter:innen* bei den drei Plattformen geführt.

Bei allen Unterschieden zwischen den Plattformen lässt sich eine Gemeinsamkeit identifizieren. Es handelt sich bei den Verantwortlichen für aktive Gestaltung der Inhaltsmoderation um »American lawyers trained and acculturated in American free speech norms and First Amendment law.«[133] Dieser Umstand hat Wirkungen sowohl auf die Frage, *warum* Plattformen moderieren, als auch auf die Frage, *wie* sie moderieren.

Die Auswertung des Materials führt Klonick zur Identifikation dreier Kernbereichen, die Plattformen zur Moderation bewegen: »(1) an underlying belief in free speech norms; (2) a sense of corporate responsibility; and (3) the necessity of meeting users‹ norms for economic viability.«[134] Dies ist nicht nur aus dem ausgewerteten Material heraus nachzuvollziehen, sondern auch vor dem Hintergrund der Bedeutung des ersten Zusatzartikels zur US-amerikanischen Verfassung. Die Verfassung und ihre *Amendments* genießen quasi-religiösen Status, was sich auch insbesondere auf die *Free Speech* und die Redefreiheit im Internet bezieht.[135] Eine Maxime dieses Glaubens ist größtmöglicher Abstand von staatlichen Eingriffen und die Betonung von Eigenverantwortung sowohl von Individuen als auch von Unternehmen. Aus diesen Umständen heraus ist es plausibel, dass nicht nur wirtschaftliche Erwägungen, sondern auch der Glaube an die Redefreiheit und unternehmerische Verantwortung die Plattformen zum Aufbau von Moderationsregeln und -verfahren bringen. Die unternehmerische Verantwortung spiegelt sich auch in den *Mission Statements*, also den Glaubenssätzen der Plattformen wider. Schon Klonick bezog sich auf solche *Mission Statements*, als Beispiele werden die von *Meta*, *YouTube* und *Twitter* von Anfang 2023 hier herangezogen:

- »Giving people the power to build community and bring the world closer together.« (*Meta*)[136]
- »Our mission is to give everyone a voice and show them the world. We believe that everyone deserves to have a voice, and that the world is a better place when we listen, share and build community through our stories.« (*YouTube*)[137]
- »The mission we serve as Twitter, Inc. is to give everyone the power to create and share ideas and information instantly without barriers. Our business and revenue

132 Vgl. Klonick (2018). *The New Governors*, S. 1617.

133 Ebd., S. 1621.

134 Ebd., S. 1618.

135 Vgl. Franks, Mary A. (2019). *The Cult of the Constitution*, Stanford: Stanford University Press, S. 105–157 (The Cult of Free Speech) & S. 159–198 (The Cult of the Internet).

136 *Meta* (o]). Our Mission, abgerufen am 03.03.2023, von: https://about.meta.com/de/company-info/.

137 *YouTube* (o]). About YouTube, abgerufen am 02.01.2023, von: https://about.youtube/#:~:text=Our%20mission%20is%20to%20give,build%20community%20through%20our%20stories.

> will always follow that mission in ways that improve – and do not detract form – a free and global conversation.« (*Twitter*)[138]

Um diese *Missionen* zu erfüllen, müssen die Plattformen unterschiedliche Regeln setzen und sind unterschiedlichen Werten verpflichtet. Gerade *Twitter* sah sich schon immer als Vorreiter im Bereich eines sehr weitgehenden Verständnisses von *Free Speech*. Das wird deutlich, wenn zum einen das um 2011 bekanntgewordene *Bonmot* »Twitter is the free speech wing of the free speech party« und zum anderen die Übernahme der Plattform durch Elon Musk bedacht werden. Musks absolutistische Vorstellungen von Meinungsäußerungsfreiheit und sein offensives Eintreten für dieselbe zeugen in seiner Praxis als *Twitter/X-CEO*, der Einfluss auf Lösch- und Sperrentscheidungen – auch gegenüber Kritiker:innen – nimmt, jedoch eher von Willkür als von radikaler Redefreiheit.[139] Nichtsdestotrotz wird deutlich, dass die verschiedenen Plattformen auch wertebasiert und aus einem unternehmerischen Verantwortungsgefühl heraus regulierend und moderierend tätig sind.

Die bedeutendste Motivation für Plattformen zu moderieren ist ökonomischer Natur. Nutzer:innen erwarten von Plattformen, zu einem gewissen Grad moderierend tätig zu werden. Dabei ist sowohl zu wenig als auch zu viel Moderation dem Geschäftsinteresse der Plattformen – möglichst lange Verweildauer und möglichst viele Interaktionen der Nutzer:innen, um Daten sammeln und Werbung verkaufen zu können – abträglich. Wer möglichst treffend die Erwartungen von Nutzer:innen erfüllt, kann wirtschaftlich erfolgreich sein. Dies erklärt teilweise auch, warum *YouTube* oder *Facebook* ökonomisch äußerst erfolgreich sind und *Twitter/X*, mit seinen schwachen Moderationsregeln, nicht.[140]

Anknüpfend an einen Essay David Posts aus dem Jahr 1995[141] zeigt Klonick, dass der Wettbewerb der äußerungsbasierten Plattformen auf einem »market for rules« stattfindet, bei dem Nutzer:innen jene Plattformen nutzen, deren Moderationsregeln ihnen am meisten zusagen. Posts Idee ergänzt Klonick um zwei empirische Erkenntnisse, die

---

138 *Twitter, Inc.* (o]). Investor Relation FAQ: Whats Twitter's mission statement?, abgerufenen am 02.01.2023, von: https://investor.twitterinc.com/contact/faq/default.aspx#:~:text=back%20to%20top-,What%20is%20Twitter%27s%20mission%20statement%3F,a%20free%20and%20global%20conversation.

139 Vgl. Perrino, Nico (01.12.2022). Free speech culture, Elon Musk, and Twitter, *The Fire*, abgerufen am 02.01.2023, von: https://www.thefire.org/news/free-speech-culture-elon-musk-and-twitter; Malik, Nesrine (28.11.2022). Elon Musk's Twitter is fast proving that free speech at all costs is a dangerous fantasy, *The Guardian*, abgerufen am 02.01.2023, von: https://www.theguardian.com/commentisfree/2022/nov/28/elon-musk-twitter-free-speech-donald-trump-kanye-west; Gensing, Patrick (26.04.2022). Twitter-Übernahme durch Musk: Meinungsfreiheit über alles?, *Tagesschau.de*, abgerufen am 02.01.2023, von: https://www.tagesschau.de/ausland/amerika/musk-twitter-109.html; Wang, Tony (19.10.2020). Twitter thread, *twitter.com*, abgerufen am 02.01.2023, von: https://twitter.com/tonyw/status/1318198195302801409?lang=de; Halliday, Josh (22.03.2012). Changing Media Summit: Twitter's Tony Wang: ›We are the free speech wing of the free speech party‹, *The Guardian*, abgerufen am 02.01.2023, von: https://amp.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

140 Vgl. Klonick (2018). *The New Governors*, S. 1627–1629.

141 Vgl. Post, David G. (1995). *Anarchy, State, and the Internet: An Essay on Law-Making in Cyberspace*, in: *Journal of Online Law*, article 3, par. 1–44.

Post in den 90ern nicht voraussehen konnte. Zum einen üben Nutzer:innen – und darüber hinaus Politik, Zivilgesellschaft und Öffentlichkeit – dahingehend Druck auf die Plattformen aus, ihre Regeln zu ändern. Zum anderen führt die Plattformisierung dazu, dass eher komplementäre monopolistische Angebote entstanden sind als substituierbare Plattformen mit vergleichbaren Funktionen.[142] Ferner verstärken Netzwerk- und soziale Gruppeneffekte die Bindung von Nutzer:innen an bestimmte Plattformen. Dennoch trägt die Metapher vom *Markt der Regelsätze* ein Stück weit, denn sie erklärt, warum einige Plattformen erfolgreich sind und andere nicht. Gerade in der Folge von Skandalen oder öffentlich diskutierten Entwicklungen auf spezifischen Plattformen und in bestimmten Altersgruppen wird von Nutzer:innenbewegungen von Plattformen zu anderen Plattformen berichtet.[143]

Plattformen haben also viele Gründe zu moderieren. Sie tun es, weil es ein Teil ihres Wesens ist, weil sie sich an Werte und unternehmerische Verantwortung gebunden sehen und weil Kund:innen und Nutzer:innen die Moderation von Inhalten erwarten, also weil es handfeste ökonomische Gründe für eine Moderationspraxis gibt.

6.2.2 Wie moderieren Plattformen?

Nachdem dargestellt wurde, *warum* Plattformen moderieren, geht es nun um die Frage, *wie* sie es tun. Aus dem *wie* lässt sich ableiten, wie die Rechte von Nutzer:innen Eingang in den Moderationsprozess finden und welche Probleme für Meinungsäußerungsfreiheit und Persönlichkeitsrechte zugleich in ihm entstehen (können).

Die zentralen Werkzeuge des *Content Managements* von Online-Intermediären sind »das Löschen und Sortieren von Inhalten und Verlinkungen Dritter sowie der Ausschluss von Nutzern.«[144] Mit Gillespie lassen sich drei Ansatzpunkte identifizieren, welche die Plattformen zur *Content Moderation* nutzen: *Erstens* moderieren sie mittels Sperrungen, Entfernungen und Filtern. *Zweitens* empfehlen sie durch *Newsfeeds*, Hitlisten und personalisierte Vorschläge Inhalte für die Nutzer:innen und *drittens* kuratieren sie Kraft von Angeboten auf Startseiten und mit »featured content«.[145]

Die Plattformen verwenden diese drei Hebel, um das Nutzer:innenverhalten dynamisch und aktiv zu lenken. Sie erzeugen durch die verschiedenen Arten von Moderation für jede:n Nutzer:in »the ›right‹ feed[], the ›right‹ social exchanges, and the ›right‹ kind of

142 Vgl. Klonick (2018). The *New Governors*, S. 1629–1630.

143 Vgl. nur: Reuter, Markus (19.12.2022). Mastodon: Der Twitter-Exodus in Zahlen, *Netzpolitik.org*, abgerufen am 02.01.2023, von: https://netzpolitik.org/2022/mastodon-der-twitter-exodus-in-zahlen/#netzpolitik-pw; Hart, Robert (03.02.2022). Facebook Loses Daily Active Users For The First Time: Here's Where They're Going, *Forbes*, abgerufen am 02.01.2023, von: https://www.forbes.com/sites/roberthart/2022/02/03/facebook-loses-daily-active-users-for-the-first-time--heres-where-theyre-going/?sh=20ec44961e6d; Gehm, Florian (25.11.2019). Nutzerschwund: Dieser Effekt bedeutet den endgültigen Abstieg von Facebook, *Welt*, abgerufen am 02.01.2023, von: https://www.welt.de/wirtschaft/article203803684/Facebook-Netzwerk-verliert-mehr-Nutzer-als-erwartet.html.

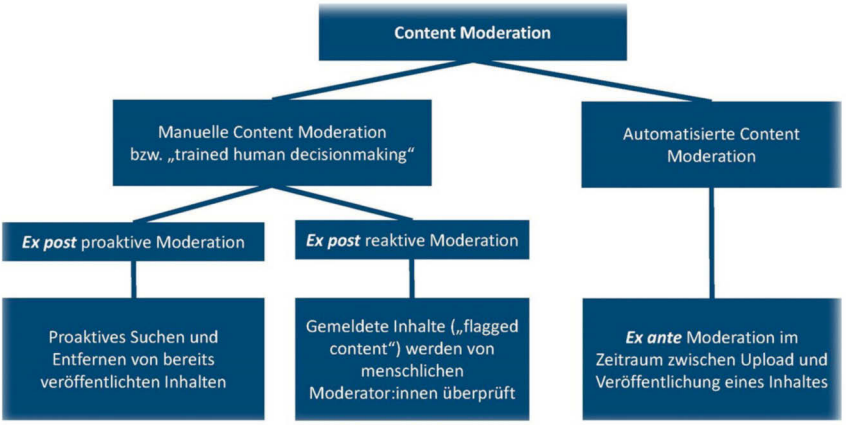
144 Schiedermair, Stephanie & Weil, Johannes (2022). *Online-Intermediäre als Träger der Meinungsfreiheit*, in: *Die Öffentliche Verwaltung (DÖV)*, S. 305–314, hier: S. 305.

145 Vgl. Gillespie (2018). *Platforms Are Not Intermediaries*, S. 202.

community.«[146] *Richtig* kann die Förderung eines gesunden, legalen und ethischen Verhaltens der Plattformnutzer:innen als Ziel der Moderation bedeuten, jedoch auch immer das, was Werbeeinnahmen steigert, Interaktion sowie Verweildauer erhöht und die Extraktion von Daten begünstigt.[147]

Die praktische Umsetzung von *Content Moderation* ist von Plattform zu Plattform unterschiedlich. Insbesondere die großen Plattformen haben ausgefeilte, mehrstufige Moderationssysteme entwickelt. Allgemein lassen sich die Moderationsmethoden zeitlich (*ex ante* oder *ex post*), ihrem Modus nach (proaktiv oder reaktiv) und schließlich nach ihren ausführenden Instanzen (automatisiert oder manuell) differenzieren.[148] Idealtypisch lässt sich das System der Inhaltsmoderation mit Klonick wie folgt abbilden (siehe Abb. 11):

Abb. 11: System der Content Moderation, (Eigene Darstellung in Anlehnung an Klonick, 2018)[149]

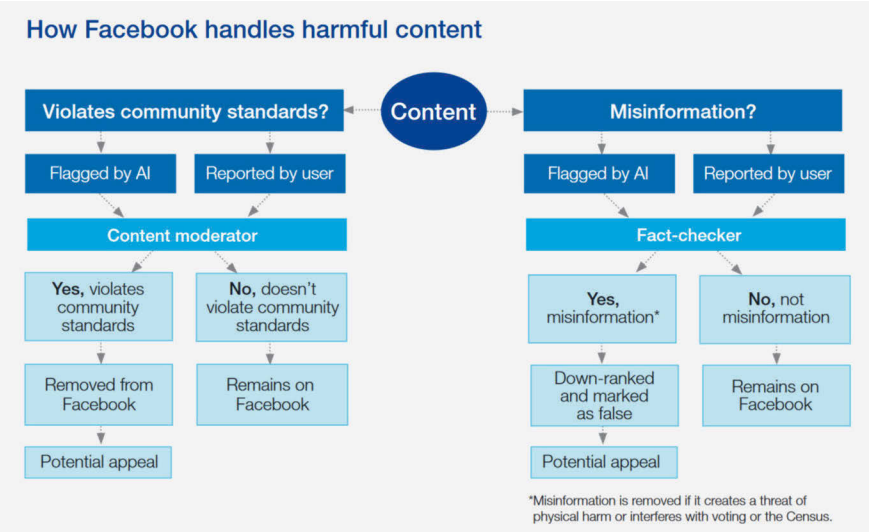


Die Differenzierung in manuelle und automatisierte Moderation darf nicht als komplett trennscharf verstanden werden. Wie in den untenstehenden Ausführungen zur automatisierten *Content Moderation* deutlich wird, ist diese auf menschliche Zuarbeit und auf die Festlegung zu moderierender Variablen angewiesen. Zudem sind es menschliche Moderator:innen, die Entscheidungen automatisierter Systeme gegebenenfalls im Nachhinein überprüfen. Auch die manuelle Moderation, die Klonick auch als »trained human decisionmaking«[150] bezeichnet, ist nicht losgelöst von automatisierten Hilfsmitteln. Algorithmen helfen bei der Suche nach problematischem Material und von Nutzer:innen oder Dritten gemeldete Inhalte (»flagged content«) werden in hochgradig formalisierten und softwaregestützten Prozessen bearbeitet. Wie gut Klonicks allgemei-

146 Gillespie (2018). *Platforms Are Not Intermediaries*, S. 202.
147 Vgl. ebd.
148 Vgl. Klonick (2018). *The New Governors*, S. 1635–1639.
149 Vgl. ebd.
150 Ebd., S. 1635.

ne Beschreibung des Moderationsprozesses zutrifft, lässt sich am Beispiel von *Facebooks Ex-post*-Moderationsprozess in der folgenden Übersicht ablesen (siehe Abb. 12):

Abb. 12: Moderationsprozess bei Facebook (Barrett, 2020).[151]



Obgleich die Bedeutung automatisierter Moderationssysteme weiter zunimmt, ist die *reaktive ex post Content Moderation* weiterhin die bedeutsamste Moderationstechnik:[152] »[A]lmost all user-generated content that is published is reviewed *reactively*, that is, through ex post flagging by other users and review by human content moderators against internal guidelines [Herv.i.O.; Anm. P.B.].«[153] Nur in bestimmten Bereichen, wie bei der Moderation von (Kinder-)Pornografie, im Bereich von kommerziellen Urheber:innenrechten oder bei Terrorpropaganda, sind sonstige Moderationsmethoden von größerer Bedeutung. Alle anderen Moderationsentscheidungen treffen qualifizierte menschliche Moderator:innen.

Im Zuge dieser Moderationsmethode kommt den Plattformnutzer:innen eine besondere Rolle zu, denn ihre Meldungen sind Grundlage effektiver Moderation. Einerseits ist so eine Moderation ohne menschliche Überwachung möglich und andererseits wird das Moderationssystem durch die Nutzer:innenbeteiligung mit Legitimation versehen.[154] Plattformmoderator:innen beschäftigen sich zum Großteil nur mit durch Nutzer:innen oder Dritte beanstandeten Inhalten, wobei sich die Plattformen in ihren

151 Barrett, Paul M. (2020). Who Moderates the Social Media Giants: A Call to End Outsourcing, *NYU Stern Center for Business and Human Rights*, abgerufen am 04.01.2023, von: https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf, S. 2.

152 Vgl. Klonick (2018). *The New Governors*, S. 1635–1636.

153 Ebd., S. 1638.

154 Vgl. ebd.

Meldeformularen zugleich dem Wissen und den analytischen Fähigkeiten der Nutzer:innen bedienen, indem diese den beanstandeten Inhalt einer Kategorie zuordnen müssen, um ihn melden zu können.[155] Diese Vorsortierung erleichtert den Moderationsprozess und bietet zugleich Ansatzpunkte für automatisierte *Machine-Learning-Moderationsmethoden*, da solcherlei kategorisierte Inhalte in Kombination mit den im Anschluss getroffenen Moderationsentscheidungen menschlicher Moderator:innen exzellente Trainingsdatensätze erzeugen. Zugleich hilft die Kategorisierung den Moderator:innen auch, bestimmte Inhalte im Rahmen der Moderation priorisiert zu bearbeiten.[156]

Eine Studie rund um Wissenschaftler:innen der *University of Texas* in Arlington verglich 2022 die Moderationskategorien von 14 Plattformen[157] in den Vereinigten Staaten und identifizierte die wesentlichen Arten moderierter Inhalte. Insgesamt konnten 15 Haupt- und 25 Unterkategorien bestimmt werden, die sich von Plattform zu Plattform z.T. erheblich unterscheiden können.[158] Die Kategorien sind in der Übersicht in Abbildung 13 aufgeführt:

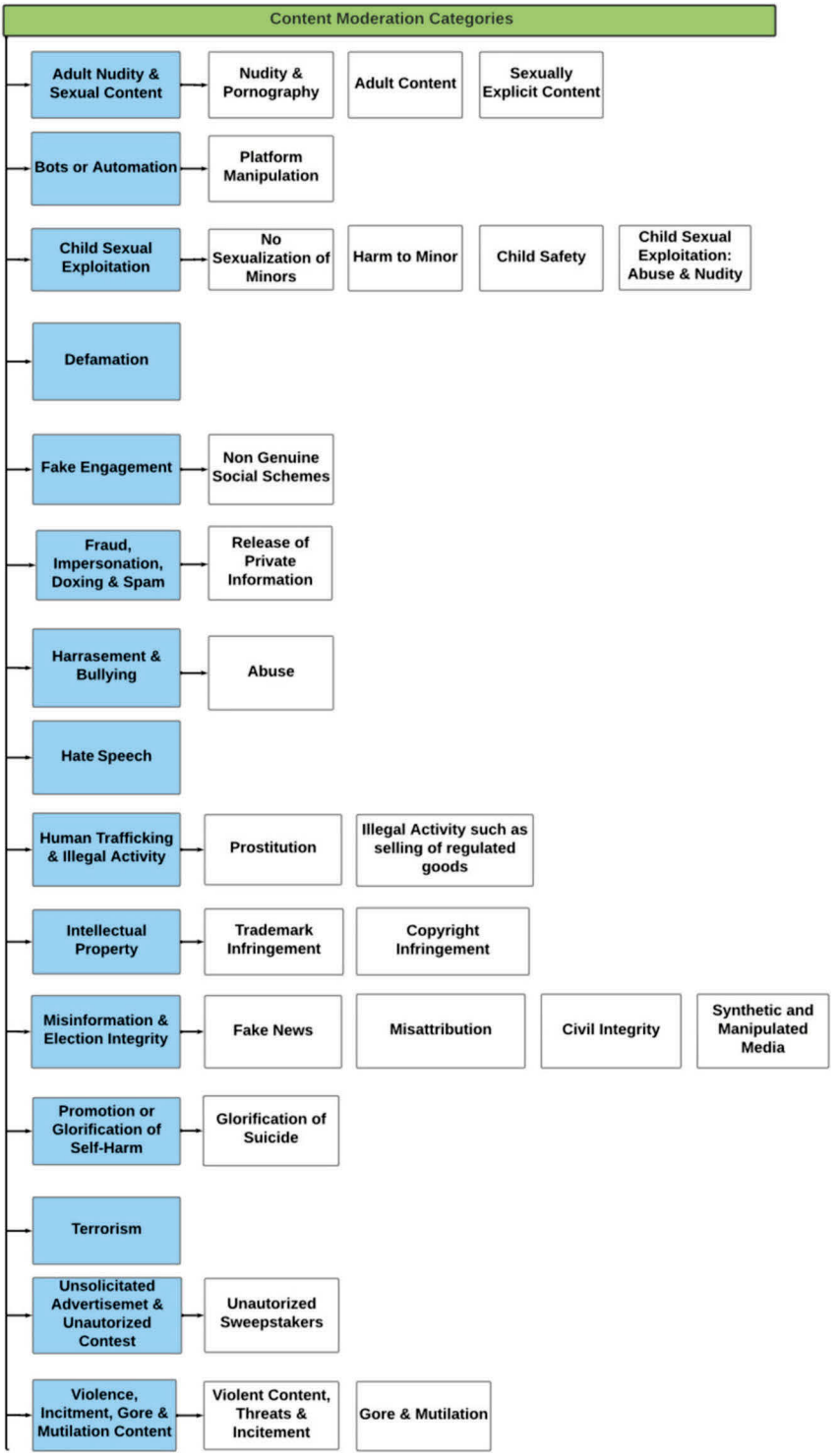
155 Vgl. Klonick (2018). *The New Governors*, S. 1639.

156 Vgl. ebd.

157 *Parler*, *Gab*, *MeWe*, *4chan*, *Rumble*, *Tumblr*, *Snapchat*, *TikTok*, *Reddit*, *YouTube*, *Facebook*, *Instagram*, *Twitter* & *Twitch*.

158 Vgl. Singhal, Mohit; Ling, Chen; Paudel, Pujan; Thota, Poojitha; Kumarswamy, Nihal; Stringhini, Gianluca & Nilizadeh, Shirin (2022). *SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and research to Practice*, Pre-Print, abgerufen am 03.01.2023, von: https://arxiv.org/abs/2206.14855, S. 7–9.

Abb. 13: Content Moderation-Kategorien 14 US-amerikanischer Plattformen (Gem. einer Analyse von Singhal et al., 2022).[159]



159 Singhal et al. (2022). *SoK*, S. 8.

Betrachtet man die Kategorien und Unterkategorien, so fällt es leicht, eine Verbindung zu den in dieser Arbeit aufgezeigten invektiven Online-Konstellationen und den herausgearbeiteten Problemachsen herzustellen. *Content Moderation* nimmt eine Vielzahl von Themen in den Blick, für die sie im Bereich der unregulierten Selbstregulierung in den USA unterschiedliche Umgangsformen entwickelt hat, die wiederum von Plattform zu Plattform unterschiedlich sind. Nicht jede Plattform moderiert jede Kategorie von Inhalten und Äußerungen.[160]

Jedoch, so die Studie von Singhal et al., sind die Moderationsprozesse in Bezug auf jene Inhalte, für die es in den Vereinigten Staaten gesetzliche Vorgaben zu beachten gilt, sprich für die Bereiche »*Child Sexual Exploitation*, *Violence*, *Intellectual Property*, and *Terrorism* [Herv.i.O.],« von eher einheitlichem Vorgehen über die unterschiedlichen Plattformen hinweg gekennzeichnet.[161]

Menschliche Moderator:innen stehen immer noch im Zentrum des Moderationsprozesses. Die z.T. komplexen Entscheidungen, die sie treffen müssen, wenn sie über das Verbleiben oder das Löschen von Inhalten und gegebenenfalls über Sanktionen gegenüber Nutzer:innenprofilen bestimmen, können automatisierte Systeme bislang nicht in einem den Ansprüchen von Nutzer:innen, Gesellschaft und Plattformbetreiber:innen genügendem Maße gewährleisten. Ihre Entscheidungen treffen die Moderator:innen anhand von internen und ständig angepassten Moderationsrichtlinien. Sie basieren auf den *Community Guidelines* bzw. *Community Standards* der jeweiligen Plattformen, die öffentlich einsehbar sind und während des Prozesses der Registrierung bestätigt werden müssen. Demnach sind die Grundlagen für die Moderationsentscheidungen für die Nutzer:innen in Grundzügen transparent. Es bleibt jedoch zu vermuten, dass die internen Moderationsregeln wesentlich umfangreicher und detaillierter sind. Trotz alledem müssen Moderator:innen Abwägungsentscheidungen darüber treffen, welche Inhalte mit den Plattformregeln konform sind und welche nicht. In einigen Jurisdiktionen müssen dabei auch gesetzliche Vorgaben beachtet werden, die von den Unternehmensregeln abweichen können. Dies macht die Aufgabe zusätzlich anspruchsvoll und weist den Moderator:innen eine besondere Rolle zu:

> »Content moderators act in a capacity very similar to that of a judge: moderators are trained to exercise professional judgement concerning the application of a platform's internal rules and, in applying these rules, moderators are expected to use legal concepts like relevance, reason through example and analogy, and apply multifactor tests.«[162]

In der Praxis bedeutet das umfangreiche Abwägungsentscheidungen, die einerseits Folgen für die Äußerungsfreiheiten der Nutzer:innen haben können, etwa wenn (zu) vieles gelöscht wird, und andererseits genauso für ihre Persönlichkeitsrechte, im Falle des zu sparsamen Löschens.

160 Vgl. Singhal et al. (2022). *SoK*, S. 7–9.

161 Ebd., S. 11.

162 Klonick (2018). *The New Governors*, S. 1642.

Das bedeutet, dass *Content Moderation* – gerade auch vor dem Hintergrund einer sich ausweitenden Grundrechtsbindung gegenüber den Plattformen – anspruchsvoll ist und Bemühungen bzgl. der Qualifikation von Moderator:innen erfordert. Daraus erwachsen weitere Herausforderungen, denn *Content Moderation* wird im Zuge der globalen Arbeitsteilung oft an Subunternehmen in Staaten mit sehr niedrigen Lohnkosten ausgelagert, wie z.B. Indien oder den Philippinen.

Solches ist aus einer ökonomischen Logik heraus insofern nachvollziehbar, da die Masse an Inhalten sowie der Bedarf an Moderation z.B. durch Gesetzgebung wie das NetzDG oder den DSA gestiegen ist. Neben den enormen psychischen Belastungen für die Moderator:innen, die der Flut an z.T. furchtbaren Inhalten und zudem prekären Arbeitsbedingungen ausgesetzt sind,[163] entsteht durch die globale Verteilung der Moderationsaufgaben das Problem eines kulturellen und wertebasierten Bias‹, der Auswirkungen auf Moderationsentscheidungen haben kann.

An *Facebook* lässt sich zeigen, dass es ein mehrstufiges Verfahren zur *Content Moderation* gibt, das die Moderator:innen in drei Stufen einteilt.[164] Das *Gros* der Clickworker:innen sind Stufe 3-Moderator:innen, die das Alltagsgeschäft bewältigen und von Stufe 2-Moderator:innen beaufsichtigt und kontrolliert werden. Die Stufe 2-Moderator:innen sind zudem für priorisiert zu bearbeitende gemeldete Inhalte zuständig. Stufe 1-Moderator:innen arbeiten oft am Firmensitz der Plattformen in den USA und sind Anwält:innen oder anderweitig hoch qualifizierte Entscheider:innen.[165] Die zumeist nicht besonders gut bezahlte Arbeit der Stufe 2- und Stufe 3-Moderator:innen ist global verteilt, wobei Länder wie die Philippinen, Indien, Mexiko, Türkei, Irland, Portugal, Spanien, Deutschland, Lettland oder Kenia eine bekannte Rolle spielen.[166] So wird die meiste

163 Menschliche Moderator:innen, v.a. in Süd- und Südostasien, bearbeiten einen Großteil des als problematisch oder rechtswidrig gemeldeten oder von automatischen Moderationssystemen identifizierten *Contents*. Die Moderationsalgorithmen sind noch nicht in der Lage, dieselben Moderationsleistungen wie Menschen zu erbringen. Dadurch sind die Moderator:innen enormen Mengen invektiver Inhalte ausgesetzt, die von *Hate Speech* über Bilder extremer Gewalt, allen Arten von Nacktheit und Pornografie bis hin zu politischer oder extremistischer Propaganda reichen. Was dies für die betroffenen *Click-Worker:innen* bedeutet, nämlich massive soziale und psychische Folgen bis zum Suizid, lässt sich eindrucksvoll und emotional einnehmend im Dokumentarfilm »The Cleaners« oder im zugrundeliegenden Recherchebuch »Digitale Drecksarbeit« nachvollziehen. Vgl. etwa Block, Hans & Rieseweck, Moritz (2018). The Cleaners: Im Schatten der Netzwelt, 88 Minuten, *Bundeszentrale für politische Bildung, farbfilm Verleih*, abgerufen am 13.12.2022, von: https://www.bpb.de/mediathek/video/273199/the-cleaners/; Rieseweck, Moritz (2017). *Digitale Drecksarbeit: Wie uns Facebook & Co. von dem Bösen erlösen*, München: dtv. Weitere wissenschaftliche Arbeiten, die das Thema aufgreifen, siehe nur: Steiger, Miriah; Bharucha, Timer J.; Venkatagiri, Sukrit; Riedl, Martin J. & Lease, Matthew (2021). *The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support*, in: *CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan May 2021, ACM, https://doi.org/10.1145/3411764.3445092; Barrett (2020). *Who Moderates the Social Media Giants.*

164 Es ist wahrscheinlich, dass andere sehr große Plattformen ähnlich vorgehen.

165 Vgl. Klonick (2018). *The New Governors*, S. 1639–1640.

166 Vgl. Barrett (2020). *Who Moderates the Social Media Giants*, S. 3; Klonick (2018). *The New Governors*, S. 1640.

Moderationsarbeit der Stufe 3-Moderator:innen in Callcentern an Niedrigstlohnstandorten wie den Philippinen erledigt und beaufsichtigt von Stufe 2-Moderator:innen vor Ort oder an Niedriglohnstandorten in den USA oder Deutschland.[167]

Während die Grundlagen der Inhaltsmoderation in den letzten Jahren transparenter geworden sind, ist es noch immer schwer nachzuvollziehen, wer die Moderationsarbeit erledigt. Schätzungen aus dem Jahr 2014 gehen von insgesamt etwa 100.000 bezahlten Moderator:innen weltweit aus.[168] Etwas neuere Zahlen aus dem Jahr 2020 gibt es für *Facebook* (ca. 15.000 Moderator:innen), *Alphabet* und dort v.a. für *YouTube* (ca. 10.000 Moderator:innen) sowie für *Twitter* (ca. 1.500 Moderator:innen).[169]

Die wenigsten Moderator:innen verfügen über eine umfassende juristische Ausbildung, vielmehr setzen die Plattformen auf Training, insbesondere wenn es darum geht, Moderator:innen dazu zu schulen, kulturelle Vorurteile und emotionale Reaktionen bei der Anwendung der Moderationsregeln auszublenden. Dies ist der Kern des Trainings. Durch beständiges Wiederholen von Trainingsentscheidungen sollen emotional und kulturell geprägte Reaktionen und Entscheidungen durch die Rationalität der Moderationsregeln überschrieben werden.[170] Dieses Vorgehen ist dann sinnvoll, wenn es darum geht, global einheitliche Regeln auf einer Plattform zu schaffen. Sie erzeugt jedoch neue Probleme, wenn es darum geht, den unterschiedlichen Regimen der Äußerungsfreiheiten in den verschiedenen Staaten und politischen Systemen gerecht zu werden – und dabei ist von unterschiedlichen gesellschaftlichen Wertvorstellungen noch nicht einmal die Rede.

Klonick zeigt, dass die Moderationspraxis *Facebooks*, und das ist auch für die anderen großen US-amerikanischen Plattformen anzunehmen, von der US-amerikanischen Rechtstraditionen und dem US-amerikanischen Recht geprägt ist.[171] Daraus erwachsen neue Schwierigkeiten für die *Content Moderation*, denn die Gewichtung von Äußerungsfreiheiten im Verhältnis zu Persönlichkeitsrechten unterscheidet sich in vielen Teilen der Welt erheblich vom Verständnis der Rechtstraditionen in den USA.

Content Moderation ist zuvorderst von globaler Arbeitsteilung und prekären Arbeitsbedingungen geprägt. Die Bewertungen und die Einordnung bedürfen menschlicher Entscheidungen. Die Menschen, die diese Moderationsarbeit leisten, erfüllen schwierige und wichtige Aufgaben, die mitunter schwerwiegende Folgen für ihre psychische Gesundheit mit sich bringen. Die Auseinandersetzung mit invektivem Material kann krank machen.[172]

Allerdings geht manuelle *Content Moderation* mit automatisierten bzw. algorithmisierten Mechanismen Hand in Hand und wird zunehmend von diesen übernommen. Aus der Nutzung automatisierter Moderationsmechanismen erwachsen Chancen und Risiken für die Wirkung von *Content Moderation*.

---

167 Vgl. Klonick (2018). *The New Governors*, S. 1640–1641.

168 Vgl. Steiger et al. (2021). *The Psychological Well-Being of Content Moderators*, S. 1.

169 Vgl. Barrett (2020). *Who Moderates the Social Media Giants*, S. 2.

170 Vgl. Klonick (2018). *The New Governors*, S. 1643.

171 Vgl. Ebd., S. 1644–1648.

172 Siehe nur: Steiger et al. (2021). *The Psychological Well-Being of Content Moderators*; Kreutzer, Lili & Köllner, Volker (2020). *Posttraumatische Belastungsstörung in der digitalen Arbeitswelt: Ein Fallbeispiel*, in: *Rehabilitation* 59 (5), S. 298–302.

Automatisierte Content Moderation

Im Zusammenhang mit der Rachepornografie wurde bereits von automatisierten *Uploadfiltern*[173] berichtet, die verhindern, dass einmal gesperrte Bild-Inhalte wieder auf eine Plattform geladen werden können.[174] Diese *ex post* in Kraft gesetzte und für die Zukunft *ex ante* wirksame Moderationsmethode ist von prädiktiven KI-Moderationssystemen zu unterscheiden, die aufgrund ihrer Parameter ausschließlich *ex ante* wirken und damit potenziell eine Bedrohung für die Rede- und Meinungsäußerungsfreiheit darstellen.[175]

Die Rede von »*algorithmic commercial content moderation* [kursiv i. Orig.; Anm. P.B.]« beschreibt Systeme, die nutzer:innengenerierte Inhalte auf der Grundlage von Übereinstimmungen (*Matching*) oder Vorhersagen (*Prediction*) klassifizieren, was zu Einzelentscheidungen auf der Basis von Governance-Richtlinien führt. Im Ergebnis werden bspw. einzelne Inhalte entfernt, für bestimmte Orte gesperrt (*Geoblocking*) oder Accounts komplett gelöscht.[176] Dies ist nur ein Teil dessen, was als (algorithmisierte) *Content Moderation* verstanden werden kann, und zwar jener Teil, der als »hard moderation« bezeichnet wird. Der andere Teil, die »soft moderation«, umfasst Dinge wie Empfehlungssysteme, Rankings, Regeln, Architekturen sowie Designentscheidungen und ist von der harten Moderation abzugrenzen.[177] Beide Bereiche haben jedoch enormen Einfluss auf Äußerungen und Äußerungsmöglichkeiten sowie auf die Formierung digitaler Öffentlichkeiten.

Zusammenfassend lässt sich der *Status quo* der automatisierten *Content Moderation* in zwei Bereiche einteilen. Zum einen in *ex ante* wirkende *Matching-*, *Geoblocking-* und Vorhersagemodelle und zum anderen in die wiederum enorm durch automatisierte Abläufe unterstützte menschliche *Ex-post-*Moderation.[178]

Matching-Systeme sind mit der Frage befasst, ob ein Inhalt derselbe ist wie ein bereits bekannter und gegebenenfalls zu löschender Inhalt. Vorhersage- oder Klassifikationssysteme sollen – und tun das auch bereits mit Erfolg – Inhalte analysieren und in Kategorien einteilen. Sie beantworten also Fragen wie »Was für ein Inhalt ist das?«, »Ist die Äußerung *Hate Speech*?«, »Ist auf dem Bild Nacktheit zu erkennen?« usw.[179]

---

173 Von gemeldeten Bildern wird ein sog. *Hash-Wert*, in etwa ein digitaler Fingerabdruck, errechnet, der ein unverwechselbares Erkennungsmerkmal darstellt. Wenn ein Bild mit gesperrtem *Hash-Wert* hochgeladen werden soll, kann dies automatisch verhindert werden. Vgl. z.B. Klonick, Kate (2020). *The Facebook Oversight Board: Creating an Independent Institution to Adjucate Online Free Expression*, in: *Yale Law Journal* 129, S. 2418–2498, S. 2429–2431, insb. Fn. 31.

174 Siehe Kapitel 4.2.2.

175 Vgl. Gorwa, Robert; Binns, Reuben & Katzenbach, Christian (2020). *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, in: *Big Data & Society*, S. 1–15, hier: S. 3.

176 Vgl. ebd.

177 Vgl. Gorwa; Binns & Katzenbach (2020). *Algorithmic content moderation*, S. 3.

178 Vgl. Bloch-Wehba, Hannah (2020). *Automation in Moderation*, in: *Cornell International Law Journal* 53, S. 41–96, hier: S. 55; Gorwa; Binns & Katzenbach (2020). *Algorithmic content moderation*, S. 3; Klonick (2020). *The Facebook Oversight Board*, S. 2430.

179 Vgl. Gorwa; Binns & Katzenbach (2020). *Algorithmic content moderation*, S. 3. Zum *Matching* siehe insb. S. 4, zur Klassifikation insb. S. 4–5.

Solche Systeme sind insbesondere dann sehr hilfreich, wenn es darum geht, einmal als zum Löschen identifizierte Inhalte für die Zukunft von Plattformen fernzuhalten, etwa dann, wenn es um Bilder im Bereich der Rachepornografie geht. Oftmals liegen ihnen Datenbanken mit von menschlichen Moderator:innen *gelabelten*, also als zu löschen identifizierten Inhalten zugrunde. Ein Problem ist jedoch die Möglichkeit der technischen Umgehung von *Matching*-Systemen durch minimale Veränderungen an den beanstandeten Inhalten. So kann der *Hash-Wert*, also der digitale Fingerabdruck eines Inhalts verändert werden. In der Folge kann der *Matching-Algorithmus* den Inhalt nicht mehr als identisch identifizieren. Eine Lösung für dieses Problem kann zumindest im Bildbereich *Software* zur automatischen Gesichtserkennung sein.[180]

Vorhersage- und Klassifikationssysteme sind technisch wesentlich anspruchsvoller, denn sie müssen anhand von Trainingsdaten, also von *zu löschen* bzw. als *nicht zu beanstanden* gelabelten Inhalten lernen, welche Art von Inhalten sie markieren sollen.[181] Dieses Verfahren ist sehr hilfreich, wenn Algorithmen durch Markierung von Inhalten menschlichen Moderator:innen zuarbeiten, die dann letztendlich die Entscheidung über den Umgang mit dem jeweiligen Inhalt treffen.

Der Vorteil automatisierter Systeme gegenüber menschlicher Moderation ist die enorme Geschwindigkeit und Skalierbarkeit, mit der sie moderieren und demnach die virale Verbreitung toxischer, invektiver oder gefährlicher Inhalte unterbinden können.

Algorithmen und auf ihnen basierende KI-Systeme, insbesondere *Machine-Learning*-Technologien, sind gerade und zunehmend bei den großen *UGC-Plattformen* im Einsatz. Dafür gibt es verschiedene Gründe.

Die Menge an Inhalten ist durch ausschließlich menschliche Moderator:innen nicht zu bewältigen, technische Lösungen beschleunigen und vergünstigen Moderationsprozesse. Hinzu addiert sich der Umstand, dass viele Unternehmen – etwa *Facebook*, *Twitter* und *YouTube* – seit Beginn der Covid-Pandemie beschleunigt auf automatisierte Verfahren umgestiegen sind, da Moderator:innen während der Lockdowns nicht an ihre Arbeitsplätze kommen konnten.[182]

Je nach Bereich werden schon heute wesentlich mehr Inhalte durch algorithmische Moderation markiert bzw. gemeldet als durch menschliche Moderator:innen oder durch Nutzer:innen. So werden nach eigenen Angaben bei *Facebook* 99 Prozent der gelöschten Inhalte mit Terrorbezug von KI-Systemen markiert und bei *YouTube* sind es 83 Prozent, was dazu führt, dass dreiviertel der Videos mit terroristischen Inhalten gelöscht werden können, bevor sie auch nur einmal angesehen wurden.[183] Auch *Twitter* nutzt schon länger *Tools* zur Erkennung von *Spam-Accounts*, um Profile, die zur Verbreitung von Terrorpropaganda genutzt werden, zu identifizieren und zu sperren. Nach eigenen Angaben identifizieren diese *Tools* über 90 Prozent der zu löschenden *Accounts*.[184]

---

180 Vgl. Gorwa; Binns & Katzenbach (2020). *Algorithmic content moderation*, S. 5.

181 Vgl. ebd.

182 Vgl. Nahmias & Perel (2021). *The Oversight of Content Moderation by AI*, S. 171–173; Elkin-Koren, Niva (2020). *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence*, in: *Big Data & Society*, S. 1-13, hier: S. 1 & 3; Gillespie, Tarleton (2020). *Content moderation, AI, and the question of scale*, in: *Big Data & Society*, S. 1–5, hier: S. 1–2.

183 Vgl. Nahmias & Perel (2021). *The Oversight of Content Moderation by AI*, S. 172.

184 Vgl. Gorwa; Binns & Katzenbach (2020). *Algorithmic content moderation*, S. 2

Automatisierte Moderationssysteme funktionieren nicht ohne Probleme. Eines ist, dass ihnen immer noch viele zu löschende Inhalte entgehen, oder dass sie mitunter zu löschenden Inhalt nicht erkennen und als unproblematisch einstufen.[185] Zudem sind KI-basierte Moderationslösungen regelmäßig an nur auf ein Problem ausgerichtet, wie z.B. der Wahrung von Urheberrechten oder dem Verhindern des *Uploads* von Pornografie. Das führt zur Vernachlässigung anderer Normen und Faktoren, wie sie in der Abwägung zwischen Meinungsäußerungsfreiheit und Persönlichkeitsrechten oder in Bezug auf die Diskussion gesellschaftlicher Werte notwendig sind.[186]

Ein anderes Problem ist, dass es automatisierten Systemen nicht gelingt, Äußerungen in ihrem Kontext zu interpretieren. Sie können keine Ironie oder keinen Sarkasmus verstehen und auch subtile Äußerungen oder subkulturelle Einbettungen nicht einordnen. In Fällen, in denen in journalistischen, politisch bildnerischen oder wissenschaftlichen Beiträgen Symbole terroristischer Organisationen verwendet werden, kann der Algorithmus dies oft nicht von echter Terrorpropaganda unterscheiden.[187] Dies gilt auch für Fälle des Zeigens von nackten Körpern, die viele Community-Standards untersagen.

Berühmt wurde der Fall der ikonischen Vietnamkriegsfotografie »The Terror of War« auch bekannt als »Napalm Girl«, das ein nacktes weinendes Mädchen auf der Flucht nach einem Luftangriff zeigt. Die norwegische Zeitung *Aftenposten* nutzte das Bild, um in einem Artikel über die Löschung desselben Bildes auf dem *Facebook-Profil* des Autors Tom Egelands zu berichten. Genau wie Egeland wurde *Aftenposten* zum Löschen des Bildes aufgefordert. Egeland beschwerte sich bei *Facebook*, woraufhin sein Profil zeitlich begrenzt gesperrt und der *Aftenposten*-Artikel samt Bild durch *Facebook* entfernt wurde. Der ganze Vorgang entwickelte sich zum medialen und politischen Ereignis, in dessen Folge *Facebook* das Zeigen des Bildes gestattete.[188]

Der Vorgang illustriert, dass sowohl die Moderationsvorgaben zu verbotener Nacktheit als auch hinsichtlich Kinderpornografie zwischen Bildnissen der Zeitgeschichte und Kunst sowie verbotenen Inhalten unterscheiden müssen. Solches ist insbesondere im Rahmen automatisierter Moderation herausfordernd, da sie auf die Erkennung von Mustern ausgelegt ist und Schwierigkeiten mit der Kontextualisierung von Inhalten hat. Weiterhin liegt bei automatisierter *Content Moderation* via *Machine-Learning-Technologie* das Problem von fortgesetzten Bias vor. *Machine-Learning* beruht auf Trainingsdaten, sodass die Algorithmen nur aus diesen Daten auf die Zukunft schließen können und wiederkehrende Muster erkennen. Neuartige invektive Muster fallen so durch das Raster und Regeländerungen müssen in die Trainingsdaten eingepflegt werden.[189]

Eine weitere Kritik der algorithmisierten *Content Moderation*, die jedoch für jede Art unregulierter Inhaltsmoderation gilt, sind die Tendenzen zum *Overblocking* und zu über-

185 Vgl. Nahmias & Perel (2021). *The Oversight of Content Moderation by AI*, S. 173; Bloch-Wehba (2020). *Automation in Moderation*, S. 45.

186 Vgl. Elkin-Koren (2020). *Contesting algorithms*, S. 2.

187 Vgl. Gillespie (2020). *Content moderation, AI, and the question of scale*, S. 3.

188 Vgl. Klonick (2020). *The Facebook Oversight Board*, S. 2439–2441; Reuter, Markus (09.09.2016). Napalm-Mädchen: Facebook zensiert ikonografisches Kriegsbild, *Netzpolitik.org*, abgerufen am 14.12.2022, von: https://netzpolitik.org/2016/napalm-maedchen-facebook-zensiert-ikonografisches-kriegsbild/#netzpolitik-pw.

189 Vgl. Gillespie (2020). *Content moderation, AI, and the question of scale*, S. 3.

mäßiger Zensur. Wenn übermäßige Moderationsmaßnahmen für die Plattformbetreiber:innen ohne Folgen bleiben, wird lieber mehr als weniger entfernt oder gesperrt.[190]

Während Gründe und Methoden der *Content Moderation* global gültig sind, haben sich durch unterschiedliche Regulierungsansätze zwei Hauptlinien der *Content Moderation* entwickelt. Zum einen die unregulierte Selbstregulierung in den Vereinigten Staaten und zum anderen das europäische Modell der regulierten Selbstregulierung. Beide werden im Folgenden umrissen, denn es sind jeweils eigene Antworten auf den Umgang mit invektiven Herausforderungen für die Meinungsäußerungsfreiheit.

6.2.3 Content Moderation in den USA – unregulierte Selbstregulierung

Obgleich die Regulierung von Äußerungen aufgrund der verfassungsrechtlichen Grenzen ein wenig realistisches Szenario darstellt, ist die Debatte um Äußerungsregulierung auf digitalen Plattformen in vollem Gange. Dies zeigt schon der Blick in aktuelle Fachliteratur.[191] Die Diskussion wird insbesondere durch den Handlungsdruck anhaltender invektiver Kommunikation und invektiven Konstellationen im Netz befeuert.

Section 230 CDA entbindet die Plattformbetreiber:innen weitgehend von einer Haftung für von Dritten produzierten Inhalten. Der Zweck der weitreichenden Plattformimmunität ist es, die Plattformen zu ermutigen, als »Good Samaritans« tätig zu werden und anstößige Inhalte zu entfernen und zu moderieren. Zugleich geht die US-Gesetzgebung somit der Gefahr aus dem Weg, das Äußerungsregulierung auf digitalen Plattformen von den Gerichten als nicht mit dem *First Amendment* vereinbar angesehen wird.[192] Den Plattformen sind in den USA weitreichende Möglichkeiten gegeben, selbst Regeln zu setzen.

Neben der weichen Moderation, also den Design-Entscheidungen der Plattformen, sind die Plattformen auch ohne gesetzlichen Zwang bereit, harte Moderationsentscheidungen zu treffen. Nahezu alle großen Plattformen verfügen über Moderationssysteme, die auf ihren Nutzungs- und Geschäftsbedingungen beruhen und im Kern lässt sich die Hauptaufgabe der Inhaltsmoderation bzw. dessen, was die US-amerikanischen Plattformunternehmen darunter verstehen, wie folgt auf den Punkt bringen: »Content moderation is the industry term for a platform's review of user-generated content posted on its site and the corresponding decision to keep it up or take it down.«[193]

---

190 Vgl. Bloch-Wehba (2020). *Automation in Moderation*, S. 45.

191 Siehe nur: Bastian (2022). *Content Moderation Issues Online*; Elkin-Koren, Niva; De Gregorio, Giovanni & Perel, Maayan (2022). *Social Media as Contractual Networks: A Bottom up Check on Content Moderation*, in: *Iowa Law Review* 107 (3), S. 987–1050; Wright (2022). *Automated Platform Governance Through Visibility and Scale*; Bone, Tanner (2021). *How Content Moderation May Expose Social Media Companies to Greater Defamation Liability*, in: *Washington University Law Review* 98 (3), S. 937–964; Lee, Edward (2021). *Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance*, in: *American University Law Review* 70 (3), S. 913–1060; Mikaelyan (2021). *Reimagining Content Moderation: Section 230 and the Path to Industry-Government Cooperation*; Nahmias & Perel (2021). *The Oversight of Content Moderation by AI*; Schiff (2021). *Informationsintermediäre*, S. 30–32; Bloch-Wehba (2020). *Automation in Moderation.*

192 Vgl. Klonick (2018). *The New Governors*, S. 1602.

193 Klonick (2020). *The Facebook Oversight Board*, S. 2427.

Jedoch bewegt sich die *Content Moderation* in den USA nicht im luftleeren Raum, sondern unterliegt gesellschaftlichen, politischen und wirtschaftlichen Zwängen. Anhaltende Problematisierungen von und Debatten um Moderationspraxen sowie weitreichende Kritik an den Regeln einzelner Plattformen führen zumindest zu punktueller Weiterentwicklung der unregulierten Selbstregulierung, wie *Metas Oversight Board* zeigt.

6.2.4 Die Weiterentwicklung unregulierter Selbstregulierung (Beispiel Metas Oversight Board)

Das ambitionierteste Moderationsprojekt US-amerikanischer Plattformen ist *Metas Oversight Board*, welches eine Kontrollebene oberhalb der plattformimmanenten Moderationsverfahren der *Meta-Plattformen Facebook* und *Instagram* einführt. Das *Board* wird im Rahmen einer von *Meta* gegründeten und finanzierten, jedoch unabhängigen Treuhandgesellschaft betrieben. Im Kern beschäftigt sich das Expert:innengremium mit strittigen Fällen, die Nutzer:innen nach Abschluss des regulären Moderationsverfahrens der Plattformen als Beschwerde vor das *Board* bringen. Diese Weiterentwicklung der staatlich unregulierten Plattformmoderation in ein System unregulierter Selbstregulierung unter Einbindung von und Unterwerfung unter die Entscheidungen Dritter nicht staatlicher Akteure (»regulatory intermediation mechanisms«), nennt Rotem Medzini »enhanced self-regulation.«[194] Er bietet eine präzise Definition: »Enhanced self-regulation is a framework in which rule-makers of self-regulatory regimes design the regime with *regulatory intermediation mechanisms* to constrain rule-takers‹ behavior and to improve their self-regulation and policy implementation [Herv. P.B.].«[195]

Medzini macht mit seiner Studie eine wichtige Beobachtung für das Verständnis der Rolle von *Content Moderation* im Rahmen der Regulierung neuer Schule. Durch die Einführung unabhängiger Moderationsinstanzen im Rahmen der Selbstregulierung verändert sich das Machtgefüge zwischen Nutzer:innen und Plattformen zumindest etwas. Von Moderationsentscheidungen betroffene Nutzer:innen können nun neben dem Staat eine weitere Institution anrufen.

Momentan ist noch nicht vollständig abzusehen, was dieser Umstand für den Umgang digitaler Plattformen mit den Herausforderungen durch invektive Online-Konstellationen bedeutet. Schon allein deshalb, weil bisher lediglich der *Meta-Konzern* für *Facebook* und *Instagram* einen dritten Mechanismus eingeführt hat. Jedoch ist *Facebook* immer noch das meistgenutzte soziale Medium und *Instagram* eine der relevantesten *UGC-Plattformen* weltweit, zumal in Deutschland und den Vereinigten Staaten.

Nicht nur Gesetzgebung und Gerichte sorgen für Anpassungen in Bezug auf das System der *Content Moderation*. Druck von Nutzer:innen, Öffentlichkeit, Zivilgesellschaft und aus der Politik führte dazu, dass *Meta-Chef* Mark Zuckerberg im November 2018 verkündete, dass es zukünftig einen fortschrittlichen Mechanismus für Beschwerden

194 Medzini, Rotem (2021). *Enhanced self-regulation: The case of Facebook's content governance*, in: *New Media & Society*, S. 1–25, insb. S. 3.

195 Ebd.

gegen Moderationsentscheidungen der Plattform *Facebook* geben soll.[196] Dies ist der vorläufige Höhepunkt der Entwicklung der Selbstregulierung der meistgenutzen *Social-Media*-Plattform weltweit.[197]

Tatsächlich wurden 2020 die ersten Mitglieder des sog. *Oversight Boards* benannt.[198] Es handelt sich dabei um bekannte und renommierte Expert:innen aus verschiedenen professionellen Bereichen sowie mit unterschiedlichen kulturellen und nationalen Hintergründen, die alle einen Bezug zu »digitalen Inhalten« und »Digital Governance« haben.[199] Das Gremium wird von Zuckerberg, aber auch in der Presse verschiedentlich als eine Art *Supreme Court von Facebook* bezeichnet.[200] Diese Metapher verdeutlicht die Bedeutung, welche die Plattformbetreiber:innen dem Gremium – mindestens in der Kommunikation nach außen – zubilligen.

Ziel des *Boards* ist es, für die *Content Moderation Facebooks* und *Instagrams* höhere Legitimität *ergo* Akzeptanz zu generieren und das Verfahren transparenter zu gestalten.[201] Entscheidungsgrundlage des *Boards* sind die jeweiligen *Community Standards* bzw. *Content-Richtlinien*, die Werte des Unternehmens, frühere Moderationsentscheidungen, Grund- und Menschenrechtserwägungen und insbesondere Erwägungen zur Meinungsäußerungsfreiheit. Am Ende des Prozesses steht die verbindliche Entscheidung, Inhalte zu entfernen oder eben nicht.[202]

Das *Board* richtet sich an die Nutzer:innen der Plattformen, aber nicht an Personen, die kein Profil bei *Facebook* oder *Instagram* haben. Nutzer:innen können sich *nach Erschöpfung ihrer Widerspruch- und Beschwerdemöglichkeiten* im regulären Moderationsverfahren an das *Board* wenden. Zudem können auch *Facebook* oder *Instagram* selbst dem *Board* sowohl relevante und komplizierte Fälle vorlegen, als auch eine vom konkreten Fall losgelöste Prüfung der Richtlinien für die Moderation verlangen. Allerdings sind die »Richtlinienempfehlungen« des *Boards* für die Plattformen nicht verbindlich. Entscheidungen

196 Vgl. Brosch, Marlene (2021). *Alles neu macht das … Facebook Oversight Board? Kritische Untersuchung der geplanten Rechtsschutzmöglichkeiten für Plattformnutzer gegen Moderationsentscheidungen*, in: *Multimedia und Recht (MMR)*, S. 26-29, hier: S. 26.

197 Siehe zur Entwicklung von *Facebooks* Content Moderation: Medzini (2021). *Enhanced self-regulation*, S. 7–14; Klonick (2020). *The Facebook Oversight Board*, S. 2427–2448.

198 Vgl. Brosch (2021). *Alles neu macht das … Facebook Oversight Board?*, S. 26.

199 Vgl. *Oversight Board* (0J). Das Gremium: Kompetenz aus aller Welt, abgerufen am 21.12.2022, von: https://www.oversightboard.com/meet-the-board/.

200 So eine Analogie Zuckerbergs, die immer wieder in der Berichterstattung aufgegriffen wird und die wohl auf ein Memo des Verfassungsrechtlers Noah Feldman zurückgeht. Vgl. Klonick (2020). *The Facebook Oversight Board*, S. 2425 & 2449; Doueck, Evelyn (2019). *Facebook's »Oversight Board:« Move Fast with Stable Infrastructure and Humility*, in: *North Carolina Journal of Law & Technology* 21 (1), S. 1–77, hier: S. 3. Eine kritische Auseinandersetzung mit Verfassungsmetaphern, wie der Rede vom »Supreme Court of Facebook«, in Bezug auf Plattformgovernance findet sich bei: Cowls, Josh; Darius, Philipp; Santistevan, Dominiquo & Schramm, Moritz (2022). *Constitutional Metaphors: Facebook's »supreme court« and the legitimation of platform governance*, in: *New Media & Society*, S. 1–25.

201 Vgl. Brosch (2021). *Alles neu macht das … Facebook Oversight Board?*, S. 26.

202 Vgl. Cowls et al. (2022). *Constitutional Metaphors*, S. 2; Brosch (2021). *Alles neu macht das … Facebook Oversight Board?*, S. 27–28.

in Bezug auf den Umgang mit Inhalten (z.B. löschen oder stehen lassen) sind dagegen verbindlich, es sei denn, die jeweilige Entscheidung verstößt gegen geltendes Recht.[203]

Das sog. »Einspruchsverfahren« erinnert an die Praxis der Verfassungsbeschwerde. Nachdem Moderator:innen von *Facebook* oder *Instagram* eine finale Entscheidung über den Umgang mit einem Inhalt getroffen haben, können betroffene Nutzer:innen Einspruch beim *Oversight Board* einlegen. Das *Board* prüft die Einsprüche und sucht sich Fälle heraus, »die dazu beitragen, die Richtlinien von Facebook und Instagram zu optimieren.« Wenn der Fall zur Entscheidung ausgewählt wird, trifft das *Board* eine verbindliche Entscheidung.[204]

Die an den Rechtsweg erinnernde Form des Verfahrens und die semantische Aufladung des *Boards* als Oberstes Gericht *Facebooks*, dürfen nicht darüber hinwegtäuschen, dass es sich um ein privates *Compliance-Verfahren*, also um Selbstregulierung handelt. Das gesamte Beschwerdeverfahren besteht neben dem *echten* Rechtsweg.

Die Genese des *Boards* erinnerte an eine Art Verfassungsgebungsprozess. Es kam zu umfangreichen Vorbereitungen mit globalen Konsultationen verschiedener Stakeholder:innen und mit Workshops an verschiedenen Orten rund um den Globus mit insgesamt 650 Beteiligten aus 88 Ländern.[205] Im Ergebnis lag dann das dieses *Board* begründende Dokument vor:

> »The Oversight Board Charter was imagined as, and ultimately became, a constitution-like document that laid out the structural relationship between Facebook, the Oversight Board, and the Trust that would sit between them. […] The nine-page foundational document was split into seven articles: Members, Authority to Review, Procedures for Review, Implementation, Governance, Amendments and Bylaws, and Compliance with Law.«[206]

Die Ambitionen und Herausforderungen des *Boards* werden noch einmal klarer, wenn man sich vor Augen führt, dass das *Board* sich an alle Nutzer:innen der *Meta-Plattformen* richtet. In dieser Arbeit wurden die unterschiedlichen Regime der Meinungsäußerungsfreiheit in Deutschland und den Vereinigten Staaten untersucht. Schon dabei werden erhebliche rechtliche und kulturelle Unterschiede bei der Gewichtung unterschiedlicher Werte und Rechte deutlich. Es fällt nicht schwer, sich vorzustellen, dass sich die Herausforderung potenziert, wenn es um eine beinahe globale Nutzer:innenschaft eines Unternehmens geht und dieses Unternehmen selbst nicht neutral, sondern wertegeleitet

203 Vgl. Brosch (2021). *Alles neu macht das … Facebook Oversight Board?*, S. 26 & 29; *Oversight Board* (o.J.). Abgerufen am 21.12.2022, von: https://www.oversightboard.com/.

204 Vgl. *Oversight Board* (o.J.). Einspruch beim Oversight Board einlegen, abgerufen am 21.12.2022, von: https://www.oversightboard.com/login/?redirect_url=https%3A%2F%2Fwww.oversightboard.com%2Fsubmit%2F.

205 Vgl. Klonick (2020). *The Facebook Oversight Board*, S. 2451–2457.

206 Ebd., S. 2457–2458. Siehe auch *Facebook* (2019). Oversight Board Charter, abgerufen am 22.12.2022, von: https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf; *Oversight Board* (o.J.). Bylaws, abgerufen am 22.12.2022, von: https://www.oversightboard.com/governance/.

und kapitalistischen Logiken unterworfen ist. In ihrer umfangreichen Untersuchung des *Oversight Boards* bringt Klonick *Facebooks* Positionierung auf den Punkt:

> »Facebook's projected rules and values are not neutral. For example, Facebook has a very vocal predisposition to and commitment to freedom of expression. This is also reflected in its Values, which prioritize voice over concerns of safety, privacy, dignity, and authenticity. Facebook's removal of hate speech, harassment, and bullying, all of which would be permissive in a First Amendment framework, mark concessions to limiting unfettered expression. Moreover, Facebook's categorization of content cannot draw on universal categories because no such consensus exists. Hate speech to a Spaniard might be political expression to a New Yorker. Obscenity to a Canadian might be art to a Korean. Fake news to an Australian might be satire to a Brazilian.«[207]

Bis Juli 2024 hat das *Board* 102 Fallentscheidungen zu Beschwerden hinsichtlich Moderationsentscheidungen auf *Facebook* und *Instagram* veröffentlicht und darüber hinaus vier »policy advisory opinions« abgegeben. In der Auswahl der veröffentlichten Fälle spiegelt sich der globale Anspruch des Gremiums wider. Bisher wurde drei Fälle mit Deutschlandbezug veröffentlicht. 13 Fälle haben Bezug zu einem EU-Land und 24 zu den Vereinigten Staaten. Die anderen Fälle verteilen sich rund um den Globus.[208] Die Entscheidungen sind leicht im Internet abrufbar und in Deutschland sowohl auf Deutsch als auch in der Sprache des betroffenen Landes verfasst. Anders als Gerichtsurteile sind die Entscheidungen grafisch ansprechend aufbereitet. Sie beginnen mit dem stichpunktartigen Verweis auf die betroffenen »Policies and topics«, »Region and countries« und die »Plattform«. Danach folgt eine Fallzusammenfassung mit Fallübersicht, dem Herausstellen wichtiger Erkenntnisse und der Entscheidung des *Boards* inklusive Empfehlungen an die betroffene Plattform. Nachfolgend wird die Fallentscheidung ausführlich ausgeführt. In ihrer Form erinnert sie durchaus an Urteile von Verfassungsgerichten. Zunächst werden Sachverhalt und Verfahrensgang geschildert und die Zuständigkeit des *Boards* geprüft. Ähnlich eines Maßstabteils werden sodann die relevanten Entscheidungs-dokumente dargestellt sowie auch die Stellungnahmen der Verfahrensbeteiligten und Dritter. Hiernach führt das zuständige Panel des *Oversight Boards* seine Entscheidungsgründe auf, um mit einer Entscheidung und Richtlinien-Empfehlungen für die plattforminterne *Content Moderation* zu schließen.

Insgesamt sind *Metas* Bemühungen der erweiterten *Content Moderation* und die Schaffung einer dritten Ebene, neben der plattforminternen Moderation und dem ordentlichen Rechtsweg, eine Ergänzung des Rechtswegs und der regulären Moderation von Inhalten. Das *Board* kann von jenen angesprochen werden, die sich im regulären Moderationsprozess unfair behandelt fühlen und die sich nicht an die Justiz wenden können oder wollen. Jedoch gilt es zu beachten, dass nur eine kleine Anzahl von Fällen entschieden wird. Zwar sind dies stets Fälle mit Symbolwirkung, aber die überwiegende Moderationsarbeit verbleibt bei den Plattformmoderator:innen. Zudem ist völlig offen,

207 Klonick (2020). *The Facebook Oversight Board*, S. 2475.

208 Vgl. *Oversight Board* (o]). Case decisions and policy advisory opinions, abgerufen am 21.07.2024, von: https://www.oversightboard.com/decision/.

ob sich das *Oversight Board* bewährt und ob die Weiterentwicklung der privaten *Content Moderation* aus den Plattformunternehmen heraus Schule macht.

Viele der Ausführungen zu *Content Moderation* sowie der (erweiterten) unregulierten *Content Moderation* gelten auch für Deutschland und die EU. Dennoch verlassen sich die europäische und deutsche Gesetzgebung nicht ausschließlich auf das Handeln der Plattformen.

6.2.5 Content Moderation in Deutschland und der EU – regulierte Selbstregulierung

In Deutschland und der EU wird das System der *Content Moderation* zunehmend durch Regulierung und Rechtsprechung beeinflusst, sodass von einem System regulierter Selbstregulierung gesprochen werden muss. Grundrechtlich lässt sich für die Moderation von Inhalten in den Artikeln 12 (Berufsfreiheit), 14 (Eigentumsrechte) und 2 Abs. 1 (Allgemeine Handlungsfreiheit) des GG eine verfassungsrechtliche Grundlage finden.[209] Diese bilden die positive Freiheit der Plattformbetreiber:innen ab, eigene Regeln für ihre Angebote aufzustellen und durchzusetzen.

Aufgrund der Herausforderungen, die im Spannungsfeld von Meinungsäußerungsfreiheit und Persönlichkeitsrechten auf digitalen Plattformen bestehen, geht die Gesetzgebung in Deutschland und in der EU im Zuge der Regulierung neuer Schule dazu über, Regeln neuer Schule für die Inhaltsmoderation aufzustellen. Durch dieses Vorgehen können die Grundrechte aller Beteiligten, insbesondere die der Nutzer:innen digitaler Plattformen, effektiv gewährleistet werden. Die Gesetzgeber:innen schaffen gewissermaßen durch ihr Handeln zumindest einen teilweisen Ausgleich der Machtasymmetrie zwischen Plattformen und Nutzer:innen. Die Plattformen ihrerseits profitieren durch einen sicheren Rechtsrahmen.

Auf nationaler Ebene ist im Bereich der Äußerungsregulierung zuvorderst das NetzDG zu nennen, welches für die wichtigsten *UGC-Plattformen* Regeln für die Inhaltsmoderation vorgibt. Das NetzDG sieht ein wirksames und transparentes Verfahren für Beschwerden vor, legt Fristen für die Beschwerdebearbeitung fest und regelt die Sanktionsmöglichkeiten gegenüber rechtswidrigen Inhalten und ihren Urheber:innen bzw. Verbreiter:innen. Ferner gibt es für Betroffene – sowohl für die sich Äußernden als auch diejenigen, über die sich geäußert wurde und über die sich beschwert wurde – die gesetzlich garantierte Möglichkeit, sich im Rahmen eines Gegenvorstellungsverfahrens zu den im Moderationsverfahren getroffenen Entscheidungen der Plattformmoderator:innen zu äußern.

Bei den Ausführungen im vorangegangenen Absatz darf nicht vergessen werden, dass sich das NetzDG nur auf *offensichtlich rechtswidrige* und *rechtswidrige* Inhalte bezieht. Sprich es soll die verfassungskonformen Grenzen der freien Meinungsäußerung besser durchsetzen und so invektiven Auswüchsen und Exzessen begegnen. *Content Moderation* darf jedoch auch engere Grenzen als die Gesetzgebung setzen. Es sieht so aus, dass sich – neben den deutschen Vorschriften zur besseren Durchsetzung v.a. des Strafrechts – ein eigenes wiederum durch Gesetzgebung und Rechtsprechung gestaltetes System der

209 Vgl. König (2019). *Vertragliche Gestaltung der Meinungsfreiheit in sozialen Netzwerken*, S. 631.

Inhaltsmoderation etabliert (hat), in welchem die Plattformen entsprechend ihren Wertvorstellungen Regeln setzen, die sie in AGB, *Community Standards* usw. festhalten. Der Gerichtsbarkeit kommt v.a. die Rolle der Kontrolle der selbstgesetzten Standards der Plattformen zu, insbesondere unter Beachtung gleichheitsrechtlicher Anforderungen, der Gewährleistung eines fairen Verfahrens sowie der Vermeidung von Willkür.

Zunehmend schalten sich Gerichte in die Ausgestaltung der Anforderungen an *Content Moderation* in Deutschland ein. Drei Entscheidungen sind dafür exemplarisch und schaffen Präzedenzfälle. Es handelt sich um die bereits thematisierte Entscheidung des BGH vom 29. Juli 2021[210] sowie um zwei Entscheidungen des *Landgerichts Frankfurt a.M. (LG Frankfurt)* aus dem Jahr 2022.[211]

Der BGH musste entscheiden, ob *Facebook* Äußerungen einer Nutzerin sanktionieren darf, wenn diese zwar »Hassrede« im Sinne der Gemeinschaftsstandards *Facebooks* sind, jedoch nicht gegen Gesetze verstoßen. Im Zuge seiner Entscheidung konkretisiert er die Anforderungen an die *Content Moderation* im Rahmen des deutschen Rechts. Grundsätzlich dürfen Plattformen:

> »den Nutzern ihres Netzwerks in Allgemeinen Geschäftsbedingungen die Einhaltung bestimmter Kommunikationsstandards [vorgeben], die über die strafrechtlichen Vorgaben hinausgehen […] [und] bei Verstoß gegen die Kommunikationsstandards Maßnahmen […] ergreifen, die eine Entfernung einzelner Beiträge und die (vorübergehende) Sperrung des Netzwerkzugangs einschließen.«[212]

Im konkreten Fall bedeutet das, dass *Facebook* gegen »Hassrede« gemäß der *Facebook-AGB* vorgehen darf, auch wenn diese nicht gegen Strafrecht verstößt.[213] Diese Regeln und Rechte der Plattformen gelten nicht unbeschränkt, sondern verlangen einen Ausgleich der Grundrechtspositionen der Plattformen (Handlungsfreiheit, Berufsfreiheit und Eigentumsfreiheit – in der BGH-Entscheidung wird nur die Berufsfreiheit angeführt) mit den Grundrechten der Nutzer:innen (Meinungsäußerungsfreiheit und Gleichbehandlungsgebot).[214] Konkret bedeutet das laut BGH-Urteil:

- dass es einen »sachlichen Grund« für eine Löschentscheidung geben muss und willkürliche Löschungen nicht zulässig sind;[215]
- dass solche Entscheidungen nachvollziehbar sind und »an objektive, überprüfbare Tatbestände anknüpfen«;[216]

210 BGH, Urteil v. 29.07.2021, Az. III ZR 179/20.

211 LG Frankfurt a.M., Urteil v. 14.12.2022, Az. 2–03 O 325/22 und Urteil v. 08.04.2022, Az. 2–03 O 188/21, openJur.

212 BGH, Urteil v. 29.07.2021, Az. III ZR 179/20, S. 42 Rn. 78.

213 Vgl. ebd., S. 49–50 Rn. 92.

214 Vgl. ebd., S. 42–43 Rn. 78 & 80.

215 Vgl. ebd., S. 43–44 Rn. 81.

216 Vgl. ebd., S. 44 Rn. 82.

- dass »zumutbare […] Anstrengungen zur Aufklärung des Sachverhaltes« unternommen werden, etwa eine Anhörung der Betroffenen;[217]
- dass Plattformen Informationspflichten gegenüber Betroffenen von Moderationsentscheidungen haben, denn sie sind »regelmäßig gehalten […], den Sachverhalt zu ermitteln und zu diesem Zweck die Beanstandung zunächst an den für den monierten Inhalt verantwortlichen Nutzer zur Stellungnahme weiterzuleiten«;[218]
- dass Kontosperrungen oder Teilsperrungen nur restriktiv von den Plattformen verhängt werden dürfen und in der Regel nur nach Anhörung der betroffenen Nutzer:in statthaft sind, während es bei der Löschung von einzelnen Beiträgen eine zügige Entfernung durch die Plattform ohne vorherige Anhörung zulässig ist, wenn den Nutzer:innen für diesen Fall in den AGB »ein Recht auf unverzügliche nachträgliche Benachrichtigung, Begründung und Gegendarstellung mit anschließender Neubescheidung« eingeräumt wird[219]
- und dass Betroffene von Moderationsentscheidungen ein Recht auf Erwiderung haben.[220]

Zusammengefasst entschied der BGH, dass die Plattformen verpflichtet sind, »den Nutzern in ihren Geschäftsbedingungen das Recht auf Benachrichtigung, Begründung und Gegendarstellung mit anschließender Neubescheidung einzuräumen.«[221] Damit legt der BGH schlüsselhaft die Richtlinien für die Ausgestaltung der *Content Moderation* fest. Somit sind aber nicht alle Fragen rund um die Verpflichtungen der Plattformen im Rahmen der regulierten Selbstregulierung geklärt. Das zeigen die beiden im Folgenden dargestellten Beschlüsse des LG Frankfurt.

Das LG entschied am 14. Dezember 2022 im Eilverfahren, dass *Twitter* »falsche oder ehrverletzende Tweets« löschen muss. So weit ist die Entscheidung inhaltlich nicht neu, jedoch ist bemerkenswert, dass *Twitter* »[a]uch sinngemäße Kommentare mit identischem Äußerungskern« entfernen muss, wenn es »von der konkreten Persönlichkeitsrechtsverletzung Kenntnis erlangt« hat.[222] Das heißt allerdings nicht, dass es für *Twitter* eine Pflicht gibt, Inhalte proaktiv auf identische Äußerungskerne zu untersuchen, sondern nur, dass es eine Pflicht hat zu reagieren, wenn Inhalte mit solcherlei identischen Äußerungskernen gemeldet werden.

In dem Verfahren ging es um ehrverletzende Äußerungen über den Baden-Württembergischen Antisemitismusbeauftragten Michael Blume. Dieser hatte die verleumderischen Tweets an *Twitter* gemeldet, was jedoch nicht zur gewünschten Löschung führte.

---

217 Vgl. BGH, Urteil v. 29.07.2021, Az. III ZR 179/20, S. 44–45 Rn. 83. Zur Anhörung siehe auch S. 47–48 Rn. 86 & 88 des Urteils.

218 Vgl. ebd., S. 45–46 Rn. 84, Zitat: S. 46 Rn. 84.

219 Vgl. ebd., S. 47–49 Rn. 87–88, Zitat S. 48 Rn. 88.

220 Vgl. ebd., S. 46 Rn. 84.

221 Ebd., S. 48–49 Rn. 89, siehe auch S. 46 Rn. 85 für weitere Begründungen des Gerichts.

222 LG Frankfurt a.M. (14.12.2022). Pressemitteilung: Persönlichkeitsrecht: Ehrverletzung durch herabwürdigenden *Tweet*, abgerufen am 15.12.2022, von: https://ordentliche-gerichtsbarkeit.hessen.de/presse/ehrverletzung-durch-herabwuerdigenden-tweet; LG Frankfurt a.M., Urteil v. 14.12.2022, Az. 2–03 O 325/22.

Zudem verletzte *Twitter* die Pflicht aus dem NetzDG, Blume eine Beschwerdemöglichkeit einzuräumen (Gegenvorstellungsverfahren) und ihn auf den Rechtsweg zu verweisen.[223]

Bereits am achten April 2022 hatte das LG entschieden, dass *Facebook* auch »Varianten mit kerngleichem Inhalt« eines *Memes* (vom Gericht als Word-Bild-Kombination definiert) entfernen muss, auch ohne durch erneute Hinweise auf die jeweilige URL verwiesen worden zu sein.[224] Es ging um ein *Meme*, das Abbildung und Namen Renate Künasts mit der nie von ihr getätigten Aussage »Integration fängt damit an, dass Sie als Deutscher mal türkisch lernen!« in Zusammenhang brachte. Das *Meme* wurde in unterschiedlichen Varianten – etwa durch Pixelveränderungen oder mit verschiedenen Kommentierungen im Bild, sog. *Captions* – verbreitet, sodass eine automatisierte Moderation *Facebooks* mittels Matchingverfahren wegen der unterschiedlichen *Hash-Werte* der *Memevarianten* nicht erfolgversprechend war. Dennoch kam das Gericht zu der Ansicht, dass es für *Facebook* zumutbar ist, auch inhaltsgleiche *Memes* zu entfernen, ohne dass diese von der Betroffenen gemeldet werden:

> »Das deutsche Recht mutet jedem Verpflichteten eines Unterlassungsgebots grundsätzlich zu, selbst festzustellen, ob in einer ihm bekannten Abwandlung das Charakteristische der konkreten Verletzungsform zum Ausdruck kommt und damit kerngleich ist. Dies gilt auch in diesem Fall. Die Beklagte [*Facebook*; Anm. P.B.] befindet sich damit in keiner anderen Situation, als wenn die Klägerin oder ein anderer Nutzer ihr eine Rechtsverletzung meldet. Auch in diesem Fall muss sie mittels *›menschlicher Moderationsentscheidung‹* feststellen, ob die gemeldete Rechtsverletzung zu einer Sperrung führt oder nicht. Vor der gleichen Situation steht sie in Anbetracht der gefilterten Kandidaten. Der einzige Unterschied ist, dass die Kandidaten nicht aufgrund einer Meldung der URL durch die Klägerin identifiziert werden, sondern durch die beschriebenen Methoden künstlicher Intelligenz. Soweit es der Technik nicht möglich ist, diese Kandidaten als Verbreiter von Falschzitaten auszuschließen, bedarf es in gleicher Form wie bei der Meldung einer konkreten URL einer von der Beklagten zu leistenden *›menschlichen Moderationsentscheidung‹*. Diese menschliche Moderationsentscheidung muss sie ohnehin treffen. Der einzige Unterschied ist, dass es nicht mehr der Klägerin aufgebürdet ist, diese Kandidaten der *›menschlichen Moderationsentscheidung‹* durch die Beklagte zuzuführen, sondern die Beklagte diese mittels der genannten Bilderkennungsverfahren selbst filtert [Herv.i.O.; Anm. P.B.].«

Plattformen wie *Facebook* müssen also ihre technischen Möglichkeiten einsetzen, um die Persönlichkeitsrechte Betroffener zu schützen. Dabei ist entscheidend, dass es keine Ex-ante-Pflicht zur Inhaltsüberwachung gibt, sondern dass die Plattformen –

223 Vgl. *Legal Tribune Online* (14.12.2022). Antisemitismusbeauftragter vor LG Frankfurt erfolgreich: Twitter muss ehrverletzende Tweets löschen, abgerufen am 16.12.2022, von: https://www.lto.de/recht/nachrichten/n/lgfrankfurt-2-03032522-twitter-antisemitismusbeauftragte-bawue-persoenlichkeitsverletzung-unterlassungsanspruch/.

224 Vgl. *Landgericht Frankfurt a.M.* (08.04.2022) Pressemitteilung: Facebook: Ehrverletzung durch Falschzitat in sozialem Netzwerk, abgerufen am 16.12.2022, von: https://ordentliche-gerichtsbarkeit.hessen.de/presse/ehrverletzung-durch-falschzitat-in-sozialem-netzwerk; LG Frankfurt a.M., Urteil v. 08.04.2022, Az. 2–03 O 188/21, openJur.

wie im vorliegenden Fall bei Facebook – ihre automatisierten Moderationsinstrumente einsetzen müssen, um inhaltsgleiche *Memes* menschlichen Moderationsentscheidungen zuzuführen, ohne dass es dafür Meldungen von Nutzer:innen oder Dritten bedarf.

Nicht nur nationale Gerichte beschäftigten sich in der Vergangenheit mit den Pflichten von Plattformen im Rahmen von *Content Moderation*. Im Oktober 2019 entschied der EuGH den Fall *Glawschnig-Piesczek vs. Facebook Ireland*. Er entschied als Vorlage des österreichischen Obersten Gerichtshofs den Fall der österreichischen Grünen-Politikerin Eva Glawschnig-Piesczek. Diese wurde nach dem Wechsel aus der Politik in die Wirtschaft auf Facebook als »miese Volksverräterin«, »korrupter Trampel« und als Mitglied einer »Faschistenpartei« geschmäht.

Gegen diese Äußerungen konnte Glawschnig-Piesczek eine Unterlassungserklärung erwirken, die *Facebook* zum Löschen der entsprechenden Postings für eine Sichtbarkeit in Österreich verpflichtete. Zugleich ging es um die Frage, ob *Facebook* auch zum Löschen wort- oder sinngleicher Behauptungen verpflichtet werden dürfe.

Die erste Instanz, das Handelsgericht Wien, bejahte beides; die zweite Instanz, das *Oberlandesgericht Wien (OLG Wien)*, verpflichtete dagegen *Facebook* lediglich zum Entfernen wortgleicher Äußerungen. Für sinngleiche Äußerungen gilt dies laut OLG nur, wenn die entsprechende Plattform »positive Kenntnis« von einer solchen Äußerung erlangt.

Der Fall kam schließlich vor den Obersten Gerichtshof Österreichs, welcher dem EuGH wesentliche Fragen zur Vorabentscheidung vorlegte. Dieses Verfahren kann der Oberste Gerichtshof wählen, wenn er europäisches Recht als maßgeblich zur Entscheidung einer Sache ansieht. Konkret ging es darum, ob das EU-Recht (wesentlich waren damals Art. 14 Absatz 1 und 15 Absatz 1 der E-Commerce Richtlinie/RL 2000/31) zulässt, dass *Social-Media*-Plattformen von einem nationalen Gericht dazu verpflichtet werden können, auch Äußerungen zu entfernen, die »wort- und/oder sinngleich« zu einer einmal gemeldeten und als rechtswidrig klassifiziert und damit als zu entfernende Äußerung eingestuft sind. Zudem ging es um die Reichweite und Geltung der Entscheidungen nationaler Gerichte in Bezug auf Moderationsentscheidungen, sprich um die Frage, ob solche Entscheidungen nur für Österreich, für die EU oder weltweit gelten.

Der EuGH entschied, dass sowohl identische als auch sinngleiche Äußerungen durch die Plattformbetreiber:innen entfernt werden müssen. Dies bezieht sich allerdings nur auf Äußerungen, die bereits im Moderationsverfahren oder auf dem Rechtsweg als zu löschen eingestuft wurden und bei denen es keine Abwägung in Bezug auf die Sinngleichheit durch die Plattformmoderator:innen geben muss. Zudem muss es der jeweiligen Plattform möglich sein, die wort- bzw. sinngleichen Äußerungen automatisiert erkennen zu können, denn sie dürfen nicht zu vorauseilenden Schritten verpflichtet werden, um proaktiv gegen rechtswidrige Inhalte vorzugehen. Der EuGH sieht ferner keine Gründe in der *E-Commerce-Richtlinie*, die einer weltweiten Geltung der Rechtsprechung der österreichischen Gerichte in der Äußerungssache entgegenstehen, womit Facebook im konkreten Fall die beanstandeten Äußerungen in Bezug auf Glawschnig-Piesczek weltweit entfernen muss.

Zwar verlor die alte *E-Commerce-Richtlinie*, also die Rechtsgrundlage der EuGH-Rechtsprechung, im Fall Glawschnig-Piesczek durch ihre Aktualisierung in Form des DSA ihre Bedeutung, jedoch zeigt das Urteil, dass der EuGH bereit ist, bei der Regulierung von Content Moderation durch seine Entscheidungen konkret mitzuwirken.

Das Urteil wurde weithin kritisiert, da etwa Aspekte des Persönlichkeitsrechts- oder Datenschutzes nicht beachtet, die verschiedenen (konkreten) Ausprägungen der unterschiedlichen EU-Rechtsordnungen in Bezug auf das Spannungsfeld von Meinungsäußerungsfreiheit und Persönlichkeitsrechten nicht thematisiert, zu hohe Erwartungen an die aktuell möglichen automatisierten *Moderationstools* formuliert und den Plattformen zu viel Entscheidungsspielräume in Bezug auf die Bewertung von Inhalten eingeräumt wurde. Gerade aus fehlerhaften Entscheidungen der Moderator:innen oder durch übermäßiges Blocken auf der Grundlage automatisierter *Moderationstools* könnte es im Rahmen von kollateraler Zensur zu unverhältnismäßigen Einschränkungen der Meinungsäußerungsfreiheit kommen.

Die Glawschnig-Piesczek-Entscheidung ist nicht das einzige Anzeichen dafür, dass die EU-Ebene erheblichen Einfluss auf die Ausgestaltung von *Content Moderation* im Rahmen der Regulierung von Selbstregulierung nimmt. Zentral ist hier wiederum der DSA, bei dem davon auszugehen ist, dass sowohl die Gesetzgebung und die Gesetzgebungsprozesse der Nationalstaaten, wie bspw. die Erfahrungen mit dem deutschen NetzDG, als auch die prominente Rechtsprechung der Gerichte der Mitgliedsstaaten der EU und des EuGHs in ihn Eingang gefunden haben. Dass beim Ziel der Harmonisierung der Regulierung digitaler Plattformen die Inhaltsmoderation nicht außer Acht gelassen wurde, zeigt schon ein Blick in den Teil des DSA, der seine wesentlichen Begriffe definiert. So wird *Content Moderation* im DSA wie folgt verstanden:

> »›Moderation von Inhalten‹ die – automatisierten oder nicht automatisierten – Tätigkeiten der Anbieter von Vermittlungsdiensten mit denen insbesondere rechtswidrige Inhalte oder Informationen, die von Nutzern bereitgestellt werden und mit den allgemeinen Geschäftsbedingungen des Anbieters unvereinbar sind, erkannt, festgestellt und bekämpft werden sollen, darunter Maßnahmen in Bezug auf Verfügbarkeit, Anzeige und Zugänglichkeit der rechtswidrigen Inhalte oder Informationen, z.B. Herabstufung, Demonetarisierung, Sperrung des Zugangs oder Entfernung oder in Bezug auf die Fähigkeit der Nutzer, solche Informationen bereitzustellen, z.B. Schließung oder Aussetzung des Kontos eines Nutzers.«[225]

Es handelt sich also um die Anerkennung der Moderationspraxen privater Plattformen und die Beschreibung ihrer konkreten Methoden. Diese umfassen algorithmische oder menschliche Kontrolle, weiterhin die Erfassung und Identifizierung von Inhalten, die gegen Gesetze verstoßen oder mit den AGB der Plattformen nicht vereinbar sind und schließlich die Sanktionsmöglichkeiten der Plattformen.

Die *Content-Moderation*-Definition der europäischen Gesetzgebung vereint den Anspruch der Durchsetzung von Recht, wie er etwa durch das NetzDG verfolgt wird, mit dem Anspruch und dem Willen der privaten Plattformen, weitergehende Regeln zu setzen und durchzusetzen.

Ähnlich dem Anforderungskatalog des BGHs in Deutschland stellt auch der DSA einen Katalog an Pflichten für die Inhaltsmoderation auf. Sie wurden bereits bei der

225 Art. 3 t) DSA.

Vorstellung des Gesetzes beschrieben und sollen hier nur noch einmal knapp aufgegriffen werden. Wie auch der BGH in seiner Entscheidung aus dem Juli 2021 darstellte, ist die Grundlage der *Content Moderation*, aber auch ihrer Regulierung, der Ausgleich der Grundrechtspositionen von Nutzer:innen und Plattformen.[226] In Bezug auf strittige Inhalte gibt es ein nutzer:innfreundlich ausgestaltetes Melde- und Abhilfeverfahren mit Begründungs- und Informationspflichten sowie einem Objektivitätsgebot und einem Willkürverbot (Art. 16 und 17). Zudem ist ein internes Beschwerdemanagement vorgeschrieben, welches Nutzer:innen verwenden können, wenn sie mit Moderationsentscheidungen nicht einverstanden sind (Art. 20). Darüber hinaus gibt es ein Verfahren zur Schlichtung (Art. 21), wenn im Rahmen von Moderation und Beschwerdemanagement kein Einvernehmen hergestellt werden kann. Im Rahmen des Moderationsprozesses werden sog. »vertrauenswürdige Hinweisgeber« bei der Bearbeitung ihrer Hinweise privilegiert (Art. 22 Abs. 1) und zuletzt müssen die Plattformen im Verdachtsfall von Straftaten mit »Gefahr für das Leben oder die Sicherheit einer Person oder von Personen« diese an die zuständigen Behörden melden (Art. 18 Abs. 1).

Bei diesem Katalog handelt es sich v.a. um »hard moderation«, also um den konkreten Umgang mit Inhalten.[227] Die Schlüsselfrage der »hard moderation« bleibt dabei, ob ein Inhalt am Ende des Prozesses bestehen bleiben kann oder entfernt bzw. (teil-)gesperrt wird, wobei für kommerzielle Profile noch das Instrument der Demonetarisierung besteht, also des Ausschlusses der Möglichkeit, mit einem Profil Geld verdienen zu können.

Der DSA nimmt jedoch weitergehen Instrumente der »soft moderation«[228] in den Blick. Das ist v.a. in Bezug auf Berichts- und Transparenzpflichten zu Plattformpraktiken der Werbung (Art. 26, 39 und 46) und der automatisierten und personalisierten Empfehlung (Art. 27, 38 und 45) der Fall, aber auch durch z.T. fakultative Verbote. Zu nennen sind etwa das Verbot von Nutzer:innen täuschender oder manipulierender Plattformorganisation und -konzeption (Art. 25 Abs. 1) – auch als *dark patterns* bekannt – und das Verbot der personalisierten Werbung gegenüber Minderjährigen (Art. 28 Abs. 2).

Der Abschnitt zeigt, dass es in Deutschland und der EU mittlerweile eine ganze Reihe der Gesetzgebung und der Rechtsprechung entstammender Anforderungen an die Ausgestaltung der *Content Moderation* und des Moderationsverfahrens gibt.

6.3 Zwischenfazit

Am Anfang dieses Kapitels 6 stand die Frage, *wie Herausforderungen für die Meinungsäußerungsfreiheit auf digitalen Plattformen im Angesicht invektiver Konstellationen bewältigt werden können*. Auf den vorangegangenen Seiten wurden zwei Antworten aufgezeigt und diskutiert. Zum einen die Regulierung durch den Staat und zum anderen die private *Content Moderation* der Plattformen.

226 Das macht der DSA in Art. 4 deutlich.

227 Vgl. Gorwa; Binns & Katzenbach (2020). *Algorithmic content moderation*, S. 3.

228 Vgl. ebd.