Transkript

Podcast Inside Instagram

S01: [00:00:02] Jede Sekunde werden auf Social Media tausende neue Kommentare geschrieben. Die meisten sind freundlich, viele sind sogar lustig, aber einige davon verletzen, greifen an oder verbreiten sogar Hass. Und genau darüber müssen wir sprechen, denn Hate Speech ist längst kein Randphänomen mehr, sondern mitten in unserem Alltag angeht. Und damit herzlich willkommen zu unserem Podcast Inside Instagram, wie wir Hate Speech auf Social Media begegnen und vor allem, was wir dagegen tun können. Mein Name ist Johann Wortmann und hier sprechen wir die nächsten Minuten über eine Plattform, die wir wahrscheinlich alle jeden Tag nutzen. Eine Plattform, die weltweit mehr als zwei Milliarden Menschen täglich nutzen und eine Plattform, die zu einem der größten Tech-Giganten der Welt gehört. Es geht natürlich, wie im Titel schon gesagt, um Instagram. (....) Es wird jetzt aber kein epischer Story-Podcast zu Instagram-Story, dafür gibt es eine ganze Menge gute Bücher, sondern ein Podcast zu einem Thema, das uns alle auf Instagram etwas angeht, nämlich Hasskommentare. Egal, ob als Hate Speech, Hass im Netz oder als kommunikative Gewalt bezeichnet, es gibt viele, viele Begriffe für digitale Anfeindungen auf Plattformen wie Instagram. [00:01:06] Und wir haben uns in den letzten Monaten an der Uni Münster sehr intensiv mit diesem Thema, nämlich kommunikativer Gewalt im Netz, beschäftigt. Wir haben Interviews geführt mit Menschen aus Politik, Sport, Justiz und Gesellschaft. Wir haben eigene Forschungsberichte geschrieben, Interviews ausgewertet, Daten analysiert und jetzt zum Abschluss produzieren wir diesen Podcast. Und damit starten wir direkt rein und tauchen ein in die Welt der Hasskommentare und Algorithmen. Das hier ist Inside Instagram. (14 seconds pause) Okay, ich habe ja im Intro vorhin schon geteasert, dass wir heute hinter die Fassade von Instagram blicken wollen. Genauer gesagt aber hinter den mysteriösen Algorithmus. Das ist natürlich die Frage, wie funktioniert der eigentlich und vor allem, wie entscheidet Instagram, was wir sehen und wann wir das überhaupt sehen und warum. Und vor allem, warum löst man hier oft den Begriff

Blackbox in die Debatte ein? Darüber sprechen wir jetzt mit Isaac. Er hat sich intensiv mit dem Instagram-Algorithmus beschäftigt und er ist jetzt bei mir hier im Studio.

[00:02:10] Sag mal hallo. Hi. Hi, grüß dich. So, es gibt ja, fangen wir mal an, ganz viele Mythen über den Algorithmus. Meta hält sich zu dem Thema ja allgemein recht bedeckt. Deshalb fangen wir vielleicht doch erstmal mit den Basics an. Was ist denn der Instagram-Algorithmus überhaupt und warum gibt es den?

S02: [00:02:26] Naja, die Forschung zeigt ganz klar, Instagram hat nicht einen Algorithmus, sondern mehrere. Also, das ist ein System, sortiert den Feed, ein anderes empfiehlt Inhalte auf der Explorer-Seite und wieder andere kümmern sich um Reels oder um die Moderation von Inhalten. All diese Systeme haben die Aufgabe, Ordnung in die riesige Menge an Post zu bringen. Damit das funktioniert, nutzen sie bestimmte Signale. Das sind zum Beispiel Likes, Kommentare, Speicherungen oder geteilte Inhalte. Auch die Beziehung zu einem Account spielt eine entscheidende Rolle. Wenn ich mit jemandem häufiger interagiere, sehe ich diese Inhalte eher. (.) Also, das Ergebnis ist ein ganz persönlicher Feed. Jeder bekommt eine eigene Version von Instagram angezeigt. Selbst zwei Menschen, die denselben Account folgen, sehen nicht dasselbe.

S01: [00:03:19] Okay, spannend. Das heißt, es ist eine sehr individuelle Experience, wenn man es so sagen kann. Jetzt habe ich es ja gerade schon gesagt, viele Leute beschreiben den Algorithmus so als Blackbox, also als mysteriöse schwarze Box, wo man irgendwie nicht eingucken kann. Erklär uns doch kurz mal, was es damit auf sich hat. (.)

S02: [00:03:35] Ja, damit ist gemeint, dass wir nicht genau sehen können, wie die Entscheidungen im Detail entstehen. Also, Studien zeigen oder nennen dafür drei Gründe. Erstens, die Plattform legt nur einen kleinen Teil ihrer Verfahren offen. Und zweitens, die Systeme sind technisch so komplex, dass sie selbst mit Einblick schwer nachzuvollziehen wären. Und drittens, die Algorithmen verändern sich ständig, weil sie aus Daten und Verhalten lernen. (.) Und das führt zu Unsicherheit. Manchmal läuft

ein Account gut und plötzlich geht die Reichweite runter ohne erkennbare Erklärung. Die Forschung nennt das algorithmische Unsicherheit. Und sie betrifft eigentlich alle Kreatorinnen und Kreatoren, Organisationen und Unternehmen.

S01: [00:04:25] Okay, das ist jetzt noch spannender für mich. Das heißt, wenn ich das mal quasi runterbreche, es geht darum, dass quasi der Algorithmus ja schon ziemlich mächtig ist wahrscheinlich. Jetzt natürlich irgendwie so die Millionenfrage. Ich weiß, aber was mag denn der Algorithmus und vielleicht was er nicht? Und vielleicht noch, gibt es da irgendwie einen Zusammenhang zum Thema kommunikative Gewalt? Darüber sprechen wir heute.

S02: [00:04:47] Naja, der Algorithmus belohnt Inhalte, die viele Reaktionen auslösen. Also Likes, Kommentare, das Speichern und Teilen von Beiträgen. Studien zeigen, dass emotionale oder provozierende Inhalte oft besonders viel Interaktion erzeugen und dadurch automatisch weiter oben im Feed landen. Das kann zum Problem werden, denn provokante oder polarisierende Aussagen lösen häufig starke Reaktionen aus. Auch negative. Dadurch können solche Inhalte mehr Sichtbarkeit bekommen. Selbst wenn sie eigentlich unerwünscht oder verletzend sind. (.) Gleichzeitig gibt es Maßnahmen, die Reichweite bremsen. Zum Beispiel eingeschränkte Sichtbarkeit über einen bestimmten Zeitraum oder vorübergehende Blockierungen bestimmter Funktionen. Diese Mechanismen sollen Missbrauch verhindern, erzeugen aber oft neue Unsicherheiten bei Nutzer und Nutzerinnen. Insgesamt zeigt die Forschung, der Algorithmus bewertet Inhalte nicht moralisch.

[00:05:47] Er sortiert nach Aktivität und genau dadurch können auch aggressive Inhalte sichtbar bleiben, wenn sie viele Reaktionen erzeugen.

S01: [00:05:56] Okay, ähm, I see, dann sehr spannende Erkenntnis und auch, ja, vielleicht spannend für euch, äh, ein bisschen drüber nachzudenken, wie Instagram dann eigentlich funktioniert. Gerade Thema, äh, Engagement. Ja, vielen Dank dir, Isaac, für diese Einschätzung. Und, äh, wir machen hier gleich direkt weiter. Gleich geht es aber um die etwas andere Seite. Äh, das hört ihr dann direkt nach der Pause. (.....) Bevor ihr jetzt das Gefühl bekommt, hey, die Plattform, die sind irgendwie so

mächtig, da kann ich als Userin oder User ja eh nichts tun, schauen wir uns doch direkt mal auch die andere Seite an. Nämlich genau die Perspektive, die wir wahrscheinlich alle tagtäglich einnehmen, die Sicht der Nutzerinnen und Nutzer. Was ihr als Instagram-Userinnen und User tun könnt und welche Tipps es da womöglich gibt, darüber sprechen wir jetzt mit Hannah, die ist jetzt bei mir im Studio. Grüß dich.

S00: [00:06:43] Hallo.

S01: [00:06:44] So, hallo, hallo. Fangen wir doch erstmal an. Entweder ich poste und mache mich damit angreifbar oder ich poste eben nicht und verliere dadurch Sichtbarkeit. Jetzt frage ich mich direkt natürlich, gilt es nur für die Politik quasi als Sonderfall oder kann ich quasi auch als Privatnutzerin, Privatnutzer direkt zur Zielscheibe werden? Zum Beispiel, sagen wir mal, wenn ich mich in eine Debatte oder einen Skandal einmische. (.)

S00: [00:07:05] Ja, natürlich kannst du auch als privater Nutzer oder private Nutzerin von Social Media oder generell Online-Räumen solche Attacken erfahren. Oftmals sind die Ausprägungen von solchen Angriffen dann allerdings ein bisschen anders. Zum Beispiel haben wir das bei Cyberbullying, dass wir das sehen, dass AngreiferInnen zum Beispiel den Betroffenen auch bekannt sind, was natürlich die Reaktionsmöglichkeiten auch verändert im Vergleich zu Personen, die dann anonym angegriffen werden in anderen Kontexten. Und trotzdem muss man aber sagen, dass die Konsequenzen auch im privaten Raum ziemlich weitreichend sind und sowohl psychische als auch professionelle, soziale oder ökonomische Ausprägungen einnehmen können. Was in der privaten Nutzung aber natürlich anders ist, ist, dass wir nicht zwangsläufig davon abhängig sind, eine öffentliche Präsenz online haben zu müssen. Das heißt, wenn wir zum Beispiel von einer fremden Person online angegangen werden, können wir über Privatsphäre-Einstellungen oder FollowerInnen-Beschränkungen diese Angriffe ein bisschen eindämmen. Außerdem gibt es eben bestimmte Themen, bei denen Kommentarspalten sowieso eher ins Negative kippen.

[00:08:08] Und da kann man sich dann auch selber fragen, ob man gerade die Kapazität bei sich sieht, um in so eine Debatte reinzugehen. Oft ist das zum Beispiel

der Fall bei Beiträgen zu Personen mit Einwanderungsgeschichte oder bei feministischen Themen oder bei anderweitig polarisierenden Themen wie Abtreibung. Oder wenn Personen in ihrer Rolle, die sie online einnehmen, als normverletzend wahrgenommen werden. Also zum Beispiel, wenn weiblich gelesene Journalistinnen über Technik schreiben. Ich selber habe zum Beispiel auch mal einen Beitrag von der Tagesschau kommentiert, bei einem Thema, das mir persönlich wichtig war. Und auch ich habe da Reaktionen bekommen. Und ich habe mir vorher auch Gedanken gemacht, habe ich gerade die Kapazität, um mit solchen Reaktionen umgehen zu können? Habe ich vielleicht auch im Netzwerk, um gegebenenfalls mit Leuten darüber sprechen zu können? (.) Und habe dann für mich in dem Moment entschieden, ja, das Thema ist mir so wichtig, ich möchte das machen. Aber genauso ist es eben auch legitim, wenn man gerade sagt, ich habe diese Kapazität nicht oder ich kann das gerade nicht leisten. (.) Man kann also sein Verhalten online auch im Privaten ein bisschen reflektieren. Aber egal, wie man sich entscheidet, ist es auch hier wichtig zu betonen, dass die Betroffenen nicht schuld sind an solchen Anfeindungen. [00:09:09] Also egal, ob sie sich öffentlich äußern oder nicht, ob sie in so einer professionellen KommunikatorInnenrolle sind, wie PolitikerInnen oder JournalistInnen oder ob sie eben Privatpersonen sind, das ist dabei egal. (.) Generell sehen wir bei diesen Anfeindungen nämlich sehr oft, dass es sich eben um eine Wiederholung von gesellschaftlichen Machtstrukturen handelt. (.) Und die Tatsache, dass wir in Online-Räumen erreichbar sind, rechtfertigt diese Angriffe eben auch nicht.

S01: [00:09:34] Okay, ja, alles wichtige Punkte. Jetzt stelle ich mir natürlich die Frage, Instagram ist eine gigantische Plattform. Also wir haben es vorhin im Intro angeteasert, knapp zwei Milliarden Menschen sind hier tagtäglich unterwegs, knapp. Und da kann man jetzt, geht zumindest mir, so leicht ein Gefühl der Ohnmacht bekommen. Jetzt die Frage, habe ich denn quasi auch als einzelner Nutzer oder NutzerIn Möglichkeiten mit Hate Speech und auch kommunikativer Gewalt quasi umzugehen? Und vielleicht das sogar irgendwie zu bekämpfen? Irgendwie kann ich dagegen steuern?

S00: [00:10:01] Ja, das stimmt. Und so ein Gefühl von Ohnmacht kann sehr, sehr schnell entstehen, gerade weil es sich halt um wirklich Massenphänomene ja auch

handeln kann. Aber letzten Endes setzt sich diese Masse ja eben auch aus den Einzelpersonen zusammen. Und dementsprechend haben wir schon auch Möglichkeiten, Angriffen etwas entgegenzustellen, wenn wir eben, wie gesagt, gerade eine Kapazität dafür bei uns sehen. Und diese Möglichkeiten sind auch relativ unterschiedlich. Also generell gilt es zum Beispiel, dass man mit problematischen Inhalten halt ein bisschen weniger interagiert, damit diese Inhalte dann nicht in Folge von den Interaktionen durch den Algorithmus verstärkt ausgespielt werden. (.) Außerdem kann man die Teilenden oder die problematischen Beiträge dann auch direkt auf der Plattform melden. Oder wenn Inhalte triggern sind, kann man diese Personen auch blockieren oder den entfolgen oder auch die Nicht-Interessiert-Funktion zum Beispiel auf Instagram nutzen. Aber abgesehen davon, nicht auf diese negativ geprägten Inhalte einzugehen oder halt gegebenenfalls diese Inhalte zu melden, (.) gilt es eben auch, wichtige, ist es eben auch wichtig, positive oder solidarische Inhalte zu unterstützen [00:11:05] und gegebenenfalls auch zu publizieren, wie gesagt, wenn man da bei sich gerade die Kapazität sieht. Das kann so aussehen, dass man zum Beispiel bei Falschnachrichten Faktenchecks von anderen Personen liked oder kommentiert. Und oftmals macht es auch Sinn, sich für solche Aktionen zu vernetzen. Das sehen wir zum Beispiel auch bei Ich bin hier. Das ist eine Initiative, die sich für Gegenrede online vernetzt und organisiert. Und genauso kann man sich im privaten Bereich eben auch mit FreundInnen absprechen, dass man gegenseitig die eigenen Kommentare liked, darauf antwortet und generell mit den geposteten Inhalten interagiert, um so A mit Betroffenen Solidarität zu zeigen und eben auch ein bisschen mit zu beeinflussen, welche Inhalte überhaupt sichtbar werden, eben auch in dem Moment dann für positive Inhalte. (..) Und wenn man halt gerade aber sagt, man hat nicht die Kapazität, um in dieses Öffentliche reinzugehen, kann man zum Beispiel Betroffenen auch Solidarität zeigen, indem man dem mal eine DM schickt, also eine Direktnachricht, ohne eben sich öffentlich positionieren zu müssen. [00:12:09] Generell geht es aber dabei eben darum, dass wir über diese organisierte und vernetzte Herangehensweise und die Veröffentlichung von zum Beispiel so positiven Beiträgen effektiv versuchen, auch aktiv das direkte Klima zum Beispiel in den Kommentarspalten zu verändern.

S01: [00:12:25] Okay, also spannend, Stichwort auch aktiv werden. Jetzt ist so ein bisschen der Punkt, wenn man sich mehr mit dem Thema Hate Speech und auch kommunikative Gewalt beschäftigt, dann taucht ja oft dieser Begriff auf, Counter-Speech, den hast du jetzt gerade auch schon eingebracht. Vielleicht einfach mal für unsere Zuhörerinnen und Zuhörer, erklär uns doch mal kurz, worum geht es da genau und vor allem, wie funktioniert das im Einzelnen, also wie funktioniert das konkret?

S00: [00:12:47] Ja, du hast es gerade schon angesprochen, Counter-Speech oder Gegenrede auf Deutsch, (.) schließt sehr stark an das an, was ich eben auch schon gesagt habe. Also Gegenrede generell soll eben online oder kann online dazu dienen, dass alternative Inhalte zu diesen Attacken oder zu kommunikativer Gewalt verbreitet werden. Also es geht um solche Sachen wie Faktenchecks oder auch Hinweise auf Konsequenzen der Angriffe, um zum Beispiel Empathie zwischen den verschiedenen Parteien zu fördern oder es geht auch um solidarische Äußerungen gegenüber Betroffenen. Und so wird eben aktiv versucht, irgendwie das Kommunikationsklima online auch positiv zu verändern. (.) Und noch eine Alternative ist zum Beispiel auch, dass aktiv nach Moderation oder nach der Wunsch nach Moderation in den Kommentaren geäußert wird. (..) Generell ist Gegenrede aber eben so ein aktiver Einsatz gegen Angriffe im Netz. Und deswegen, das meinte ich ja eben auch schon, geht es halt dabei darum, dass man auch guckt, welche Kapazität man gerade hat, sowas auch zu leisten. Und auch hier kommt die Rolle von Netzwerken eben wieder hinzu, [00:13:47] damit man eben nicht das Gefühl bekommt, wie du auch vorhin meintest, alleine irgendwie gegen viele zu stehen.

S01: [00:13:52] Ich finde gerade auch den Punkt spannend, quasi zu sagen, naja, Thema Kapazität, ne? Weil wir haben gerade drüber geredet, das ist ja auch emotional unfassbar belastend und es frisst ja auch einfach Zeit. Also das muss man auch dazu sagen. //S00: Total.// Deswegen, vielleicht zum Abschluss, nehmen wir mal an, ich scrolle nachher irgendwie auf der Couch, auf dem Bett durch meinen Instagram-Feed und sehe einen Beitrag, der halt wirklich offen Hate-Speech betreibt. Also können wir uns jetzt irgendein Beispiel nehmen, wo ich wirklich sage, das geht überhaupt nicht und ich möchte was tun, ich möchte eben nicht das Gefühl der

Ohnmacht haben. Wie und wo kann ich den melden? Hast du da irgendwie eine Empfehlung oder einen Tipp für mich?

S00: [00:14:23] Ja, also die Möglichkeiten zu melden sind tatsächlich auf den verschiedenen Plattformen auch recht unterschiedlich. Auf Instagram kannst du anstößige Beiträge direkt melden oder flaggen, in Anführungszeichen. Und Instagram prüft solche Beiträge dann. Und wenn der gesperrt werden sollte, dann bekommst du auch eine Rückmeldung. Allerdings ist da zu reflektieren, dass diese Prozesse oft recht langwierig und auch nicht immer erfolgreich zwangsläufig sind. (..) Bei Instagram findest du diese Möglichkeiten zum Beispiel über die drei Punkte, die du neben dem Post siehst. Aber generell gilt es da, sich zu informieren, was auf welchen Plattformen halt möglich ist. Und wenn du diesen Hass zum Beispiel selber erfährst und etwas tun könntest, du kannst zum Beispiel auf Instagram auch ein Sperrformular ausfüllen und damit die Löschung des Beitrags oder des Kommentars fordern. Da ist allerdings der Disclaimer zu nennen, dass Instagram in diesem Sperrformular auch fordert, dass man sich selber vorher rechtlich beraten lässt. Da sollte es durch diese, durch das Sperrformular zu einem Gerichtsprozess kommen, [00:15:26] man da eben auch wieder involviert werden würde. (.) Infos dazu, wo man aber was wie melden kann, gibt es aber bei verschiedenen Unterstützungsorganisationen, wie eben zum Beispiel eben schon angesprochen, ich bin hier oder bei HateAid oder in dem Verbund ToneShift. und diese Organisationen, die helfen auch, sich generell zu dem Thema zu vernetzen und bieten Beratungen an, welche Schritte denn eingeleitet werden können, auch in Richtung Strafverfolgung. Die geben zum Beispiel auch ein bisschen Input dazu, wie man denn eigentlich so eine Beiträge online sichern kann, also wie man wirklich eine Spurensicherung online auch produktiv leisten kann. (.) Denn generell ist es immer wichtig, sich bewusst zu machen, dass Angriffe, die menschenverachtend sind oder Hass verbreiten, auch im Online-Raum strafbar sind und eben auch offline nachverfolgbar sind. Das ist zum Beispiel der Fall, wenn so ein Kommentar offen zu Gewalt aufruft. (..) Und wie gesagt, das kann dann eben auch ein Fall für die Strafverfolgung werden. Da helfen dann auch Polizei oder auch die Staatsanwaltschaft weiter.

S01: [00:16:28] Okay, das heißt, wir haben verschiedene Eskalationsstufen, wenn man das vielleicht so nennen kann. Also fassen wir kurz zusammen, auch als Einzelperson kann man eine Menge gegen Hass im Netz tun. Ihr könnt vor allem Solidarität zeigen, zum Beispiel, indem ihr Beiträge meldet, indem ihr Gegenräte, beziehungsweise wie wir es gerade hatten, Counterspeech betreibt und eben auch solche Kommentare quasi kontextualisiert. Und vor allem ganz wichtig, vernetzen. Vernetzt euch, tauscht euch aus und seid gemeinsam stark dagegen. Und in diesem Sinne, danke an Hannah für das spannende Gespräch.

S00: [00:16:56] Danke dir.

S01: [00:16:57] Wir machen jetzt einen kleinen Sprung weg von der individuellen Perspektive, zurück zum großen Ganzen, wie man so schon sagt. Es geht nämlich um das Thema Moderation. Das ist ja bei uns gerade auch schon angeklangen, aber es ist natürlich die Frage, warum geht es da eigentlich genau und warum ist es vielleicht auch tückisch? (.....) Für die einen ist es aktive Zensur. Für die anderen hingegen ist es der Inbegriff eines demokratischen Internets. Es geht natürlich um das Thema Content-Moderation oder Content-Moderation. Genauer gesagt aber um die Frage, wie Plattformen wie Instagram eigentlich ihre Inhalte absegnen oder eben auch nicht. Wie das genau funktioniert und warum Content-Moderation ein so komplexes Thema ist. Darüber sprechen wir jetzt hier mit einem Gast und natürlich habe ich hier auch wieder einen Experten dabei, nämlich Tristan. Hi. Grüß dich. Fangen wir doch mal ganz vorne an. Nehmen wir anderes an. Ich lade jetzt hier aus dem Studio ein Selfie auf Instagram hoch. Ja, kann ich ja kurz machen. Setze eine coole Caption drunter. Prüft Instagram dann direkt meinen Post vor dem Upload? Oder wie funktioniert das? Erklär uns doch mal kurz, wie das abläuft. (..)

S03: [00:17:58] Ja, also grundsätzlich ist das richtig, aber nicht so redaktionsmäßig, wie man sich das vielleicht vorstellt. Also, es sitzt natürlich niemand bei Instagram irgendwie im stillen Kämmerlein und prüft jeden Post, bevor er online geht. Aber mithilfe von künstlicher Intelligenz wird zumindest so eine erste automatisierte Vorabprüfung vorgenommen, wo dann oberflächlich schon mal irgendwie auch Verstöße gegen Community-Richtlinien geprüft wird. Also beim Text, aber auch

natürlich auch bei Bildinhalten. (.) Konkret bedeutet das so eine Art Fahndung nach bestimmten Mustern, wie Nacktheit, Beleidigungen oder Hasssymbolen. Ja, und wenn der KI-Algorithmus etwas Verdächtiges findet, kann er den Inhalt mit Warnhinweisen versehen oder herabstufen oder im Ernstfall sogar blockieren. (.) Aber solche Entscheidungen sind natürlich nicht immer schwarz oder weiß. Und auch Sprache hat ja von Ironie bis Humor einfach so viele verschiedene Facetten. Und ja, deshalb braucht es weiterhin die menschliche Komponente sozusagen, die bestimmte verdächtige Inhalte nachträglich prüft und bewertet. Also der Post würde direkt nach der automatisierten KI-Prüfung [00:19:02] erstmal online gehen, sofern du nicht gegen die Community-Richtlinien verstößt.

S01: [00:19:07] Es ist aber natürlich der Punkt, aktive Moderation, also auch sowas, was du gerade erklärt hast, ist ja durchaus umstritten, sagen wir es mal so. Vor allem die Frage, wer hier eigentlich bestimmt, was denn sagbar ist, also was in dem Fall quasi online gehen würde, oder was eben nicht. Wer hat denn aktuell diese Macht? (.)

S03: [00:19:24] Ja, das ist natürlich so ein bisschen die Gretchenfrage und das ist etwas zu komplex, um es in einem Satz zu beantworten. Also im Moment liegt diese Macht hauptsächlich bei den Plattformen selbst, würde ich sagen. Also im Falle von Instagram bei Meta. Denn sie legen ihre Community-Standards selbst fest. Das heißt, die Grundlage dafür, welche Inhalte erlaubt sind und welche nicht. Also da geht es zum einen um juristische Belange und Expertise, die dann dafür zur Rate gezogen wird. Aber natürlich auch um wirtschaftliche, beziehungsweise um unternehmerische Interessen für Meta. (.) Ja, dann gibt es die Moderatoren, die von der Plattform eingesetzt werden und täglich meist sehr schnell unzählige Entscheidungen treffen. Sie haben also eine sehr konkrete, praktische Macht darüber, was sichtbar bleibt und was nicht. Aber sie sind natürlich in ihren Entscheidungen letztlich an die Community-Standards der Plattform gebunden. (.) Ja, und beim Thema Macht und Durchsetzung müssen wir natürlich auch über Politik sprechen. Also in Deutschland gibt es ja das Netzwerk Durchsetzungsgesetz jetzt schon seit einigen Jahren. Das ist auf EU-Ebene jetzt weitestgehend [00:20:25] eigentlich in den DSA, also den Digital Services Act, übergegangen. Ja, diese Gesetze verpflichten Plattformen wie Instagram natürlich auch dazu, unter anderem Hass und strafbare Inhalte schnell zu entfernen, klare

Meldewege anzubieten, etc. (.) Also Plattformen haben kein Machtmonopol, aber liegen natürlich trotzdem immer noch im Machtzentrum bei Fragen rund um das Thema Moderation. Moderation. Und genau das macht Moderation ja so umstritten, weil gigantische Unternehmen wie Meta ja damit faktisch über Teile unseres öffentlichen Diskurses entscheiden.

S01: [00:20:58] Warum ist es denn so schwer, dann quasi Inhalte auf Instagram zu moderieren, ohne dass die Leute quasi Angst vor dem einen oder dem anderen haben? (.)

S03: [00:21:06] Jetzt bezogen auf Instagram, also wir alle wissen, wie viele Posts jede Stunde, jede Minute weltweit veröffentlicht werden. Ja, und gleichzeitig fordern wir dann von Moderationen normativ trotzdem ein, dass sie präzise und fair richtige Entscheidungen trifft und natürlich am besten so schnell wie möglich. Und das ist technisch und menschlich schlichtweg nicht möglich und nicht realisierbar. Zum Beispiel Sprache zu bewerten, ist extrem kontextabhängig. Also die KI versteht diesen Unterschied oft nicht. und Menschen hingegen meist schon, weil sie den Kontext kennen, verstehen und so ein bisschen diese sozialen Regeln der Kommunikation ja auch besser kennen. (.) Und da Menschen aber auch langsam sind, haben sie halt nur wenige Sekunden pro Fall für ihre Bewertung, stehen dabei unter Stress und das erschwert natürlich Moderation gewaltig. (.) Und dann gibt es diese Fälle, bei denen selbst Juristen uneins sind. Also wenn es um Ironie, um Satire oder Kritik an Politikern geht und die Frage im Raum steht, ist es noch Free Speech oder schon Strafe? (.) Ja, die Grenzen der Meinungsfreiheit sind manchmal weder endgültig geklärt [00:22:07] noch gut bewacht. Und wer soll das als Moderator fair und objektiv bewerten, wenn selbst Juristen vortrefflich darüber streiten können? Und natürlich befindet sich Instagram auch selbst in einer Dilemmasituation. Also die Plattform lebt ja von Aufmerksamkeit, von Traffic, generiert dadurch ihre Einnahmen und provokante Inhalte erzeugen häufig genau diese Interaktion. Also Engagement und aktiver Austausch sind gut fürs Geschäft, aber halt schlecht, wenn dadurch Hass sichtbarer wird. Also Moderation ist also schwer, weil es keine einfachen Ja- oder Nein-Entscheidungen gibt. Und ja, weil die Plattformlogik natürlich so ein bisschen auch gegen die gesellschaftliche Verantwortung arbeiten kann.

S01: [00:22:45] Okay, das heißt, wenn ich das richtig raussehe, haben wir einen gewissen Interessenkonflikt. Ja, quasi zwischen einerseits ich will irgendwie Engagement, Traffic und andererseits, naja, was kriegt denn Engagement und Traffic? (.) Thema Hasskommentare. Jetzt ist ja irgendwie der Punkt, gerade irgendwie Richtung Abschluss, dass viele Plattformen oder eigentlich die meisten quasi über User-Feedback moderieren. Also das heißt konkret, ich sehe einen anstößigen Post, melde den und nehmen wir mal einen Idealfall an, Instagram benachrichtigt mich, nimmt den runter und sagt, jo, vielen Dank dafür, dass du Instagram zu einem besseren Ort gemacht hast. Jetzt ist natürlich die Frage, gibt es denn noch andere, also quasi alternative Modelle zu so einer aktiven Moderation? (..)

S03: [00:23:27] Ja, also es gibt tatsächlich mehr als nur dieses klassische User meldet, Plattform löscht. Also Instagram moderiert proaktiv. Wir haben das vorhin schon kurz angerissen. Also die KI-Systeme überprüfen Inhalte schon beim Hochladen und greifen natürlich auch ein, wenn irgendetwas klar gegen die Community-Standards verstößt. Und ja, das passiert ja automatisch und meist schon bevor überhaupt irgendwie jemand etwas meldet. (.) Manchmal bleibt ein Beitrag, aber auch online wird jedoch kaum ausgespielt. Also da sprechen wir vor allem dann über Posts im viel zitierten Graubereich, den wir ja gerade auch schon angerissen haben. Ja, und dann gibt es natürlich die Community-Moderation oder Community-Moderation. Das heißt, Instagram gibt uns ja auch viele Tools, mit denen wir selbst filtern können, von eingeschränkten Konten bis hin zur Kommentarfreigabe. Und damit moderieren wir unsere eigene Umgebung ja auch ein Stück weit selbst. Man könnte sogar sagen, jeder von uns ist auf Instagram ein Stück weit selbst Moderator. Denn worauf wir klicken, was wir melden, was nicht, alles beeinflusst, wie unser persönliches Feed aussieht [00:24:28] und welche Inhalte Sichtbarkeit bekommen und welche nicht. Und genau deshalb ist Moderation nicht nur Sache der Plattform, sondern kann auch von uns Nutzenden aktiv mitbestimmt und gestaltet werden.

S01: [00:24:39] Da hast du ein sehr schönes Fazit für einen Abschluss mitgebracht. und in diesem Sinne vielen Dank dir, Tristan, für das ja wirklich total spannende Gespräch, weil ich finde gerade die Frage, wie Inhalte auf Instagram moderiert

werden, die bleibt ja irgendwie spannend, korrigiere mich da gerne, weil es ja eben irgendwie verschiedene Modelle und Möglichkeiten gibt und irgendwie keiner so richtig zu wissen scheint, was davon denn jetzt eigentlich der Way to go ist. Vielleicht ist es auch eine Kombination. Wir wissen es nicht. In diesem Sinne, es ist ja irgendwie ein laufender Prozess, das ist jetzt mein Eindruck, von allem, wo wir uns darüber unterhalten haben. und deswegen da vielleicht auch der Hinweis, sehr spannendes Thema, was wir jetzt natürlich nur zum Stand dieser Produktion abbilden können. Genau, vielen Dank dir nochmal und damit sind wir dann auch am Ende unseres Podcasts angekommen.