WAS SIND
UPLOADFILTER?

Foto mit freundlicher Genehmigung von Marcel Felde

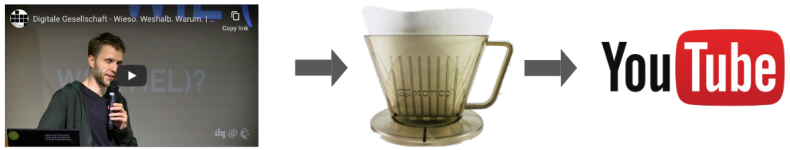
DIGITALE GESELLSCHAFT

Was sind Uploadfilter?

Im Frühling 2019 fanden breite Straßenproteste gegen die drohende Einführung von Uploadfiltern statt. Dennoch wurden mit der europäischen Urheberrechtsreform viele Onlineplattformen verpflichtet, Uploadfilter zu installieren. Ist die politische Debatte damit beendet? Wie funktionieren Uploadfilter? Haben sie etwas mit künstlicher Intelligenz zu tun? Können sie Memes erkennen? Werden sie ohnehin schon längst verwendet? Droht eine zentrale Kontrollinfrastruktur für das World Wide Web (WWW)? Was bedeuten sie für die Freiheit des Internets? Und was für die Zukunft von Sharing-Plattformen?

Inhaltsverzeichnis

Wofür sollen Plattformen verantwortlich sein?



Die Bezeichnung „Uploadfilter“ meint technische Systeme, die auf Plattformen installiert sind, auf denen Nutzende Inhalte hochladen und veröffentlichen können. Sie untersuchen hochgeladene Inhalte auf bestimmte Eigenschaften und lehnen die Veröffentlichung gegebenenfalls ab.
Das Ziel ist, automatisiert Inhalte zu erkennen, die illegal oder unerwünscht sind, zum Beispiel weil sie Urheberrechte verletzen, Nacktheit zeigen oder für Organisationen werben, die als terroristisch eingestuft werden. Zur Inhaltserkennung kommen dabei verschiedene Technologien zum Einsatz: Fingerprinting, Hashing und auch maschinelles Lernen.
In den letzten Jahren hat in der Diskussion um die Verantwortlichkeit von Online-Plattformen ein Paradigmenwandel stattgefunden. In der europäischen e-Commerce-Richtlinie von 2000 wurde eine Haftungsfreistellung für Plattformen (Providerprivileg) festgeschrieben, die dem Grundsatz nach noch heute gilt. Danach haften Plattformen grundsätzlich nicht für die Inhalte, die Nutzende auf ihren Seiten publizieren – solange sie nicht über rechtswidrige Inhalte informiert werden. Damit war der rechtliche Grundstein der Sharing-Plattform gelegt: Die Plattformbetreibenden waren nicht gezwungen, alles zu überprüfen, was auf ihren Seiten veröffentlicht wurde, weil sie keine Haftung fürchten mussten. Den EU-Staaten war verboten, ihnen allgemeine Überwachungspflichten aufzuerlegen. Ab Anfang der 2000er Jahre standen Fragen um die Verbreitung von urheberrechtlich geschützten Inhalten im Internet immer stärker im Zentrum der Auseinandersetzungen. Der Musik und Filmindustrie reichte es nicht mehr, die Plattformen nachträglich zum Löschen von Verstößen auffordern zu können. Diese sollten Urheberrechtsverletzungen bereits beim Upload sperren. Auch seitens der Sicherheitspolitik wurden Stimmen laut, die nach Verantwortlichkeit der Plattformen riefen: Insbesondere nach den terroristischen Anschlägen in Frankreich 2015 sahen europäische Regulierer in der Propaganda des sogenannten Islamischen Staates (IS) eine weitere Inhalte-Kategorie, die aus dem Netz verschwinden sollte.
In der Folge kam es zunächst nicht zu gesetzlichen Regulierungen, aber der Druck auf Plattformen, Urheber an den Werbeeinnahmen aus der Nutzung ihrer Werke zu beteiligen, wuchs. Die Verwertungsgesellschaften und die großen Plattformen handelten zum Teil Vereinbarungen über Vergütungen aus. So einigten sich die für Musikrechte zuständige GEMA und Youtube 2016 nach achtjährigem Streit zu nicht veröffentlichten Bedingungen. Die Fehlermeldung, dass Videos, da Musikrechte von der GEMA nicht eingeräumt wurden, „in Deutschland leider nicht verfügbar“ seien, verschwand. Im Hinblick auf als terroristisch klassifizierte Inhalte etablierten Dienste in Kooperation mit der EU-Kommission Selbstregulierungsmechanismen. Dabei ging es darum, Inhalte zu entfernen. Gegenwärtig tendiert die EU immer stärker zu gesetzlichen Pflichten, die Veröffentlichung von bestimmten Inhalten zu verhindern. Diese finden sich nicht nur im umstrittenen Artikel 17 der 2019 beschlossenen Urheberrechtsreform.[1] Die EU-Kommission möchte sie auch in einer Verordnung gegen terroristische Inhalte,[2] über die die europäischen Institutionen derzeit noch verhandeln, festschreiben. Mit dem Digital Services Act, der für die nächsten Jahre auf der Agenda steht, könnte eine solche Filterpflicht schließlich bereichsübergreifend eingeführt werden. Die politische Debatte über Uploadfilter steht deshalb erst am Anfang.

Die Technologien hinter den Filtern

Uploadfilter lassen sich in drei Komponenten unterteilen:

1. (Wieder-)Erkennung bestimmter Inhalte: Diese Komponente überprüft automatisiert, ob ein Inhalt mit einem in einer Datenbank hinterlegten Inhalt übereinstimmt oder bestimmte Merkmale aufweist.

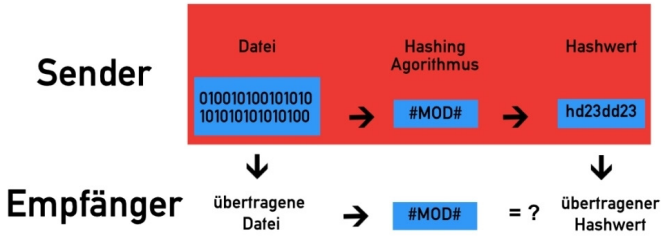
2. Entscheidungsregeln: Sie legen fest, was mit einem erkannten Inhalt geschehen soll – z.B. eine Information an den Nutzer versenden, zur menschlichen Überprüfung vorlegen oder die Veröffentlichung verhindern.

3. Systeme, in denen Informationen über die Rechte an Inhalten oder darüber, welche Inhalte nicht erscheinen dürfen, hinterlegt sind, z.B. Hash-Datenbanken.

AUTOMATISCHE INHALTSERKENNUNG (AUTOMATED CONTENT RECOGNITION, ACR)

Bereits seit einiger Zeit ermöglichen Dienste wie Shazam die Erkennung von Audioinhalten. Dabei werden meist sogenannte digitale Fingerabdrücke verwendet, um ein bestimmtes Werk zu identifizieren. Neben Dienstleistern, die das für die Bereiche Musik und Video ermöglichen, gibt es auch eine Vielzahl von Unternehmen, die Software anbieten, um Bilder wiederzuerkennen. Im Bereich Text wird automatische Inhaltserkennung unter anderem zur Plagiatserkennung in der Wissenschaft eingesetzt. 2007 führte YouTube das System Content ID ein, das ursprünglich von Audible Magic entwickelt wurde. Sein Zweck besteht darin, Videos wiederzuerkennen, um Urheberinnen und Urhebern die Verfolgung ihrer Rechtsansprüche zu ermöglichen. In einer Datenbank werden Fingerabdrücke von Inhalten gespeichert, die Rechteinhaber dort als ihnen zugehörig einreichen können. Wenn eine Übereinstimmung zwischen einem gespeicherten und einem neu hochgeladenen Inhalt erkannt wird, kann der Rechteinhaber entscheiden, ob er ein Video sperren, es monetarisieren oder Daten darüber erhalten möchte, wo und wann das Video wie oft angesehen wurde.[3] Ähnlich ermöglicht die Funktion „Rightsmanager“ auf Facebook seit 2016, Inhalte als geschützt zu kennzeichnen und die Verwendung durch andere Nutzende auf Facebook zu begrenzen.[4]
Zur Erkennung unerwünschter Inhalte setzen einige Plattformen außerdem maschinelles Lernen ein. So sollen zuvor unbekannte Inhalte auf bestimmte Merkmale untersucht werden. Zum Beispiel hat Tumbler mit „NSFW“ (not suitable for work) ein System entwickelt, dass Nacktheit auf Fotografien erkennt.[5] YouTube verwendet Systeme, die auf maschinellem Lernen beruhen, um damit Inhalte zu markieren, die unter die Kategorie „gewalttätiger Extremismus“ fallen können. Im nächsten Schritt werden die markierten Inhalte einer menschlichen Kontrolle unterzogen.[6]

Diese Dienste arbeiten mit verschiedenen Technologien, die das Wiedererkennen eines bestimmten Inhalts, der einmal in eine Datenbank eingespeist wurde, oder das Erkennen bestimmter Merkmale ermöglichen. Was sie nicht ermöglichen, ist eine automatische Einordnung des Kontextes, in dem ein hochgeladener Inhalt steht, d.h. ob es sich um erlaubte Zitate, Satiren, kritische Auseinandersetzungen mit Themen oder eine journalistische Berichterstattung handelt.

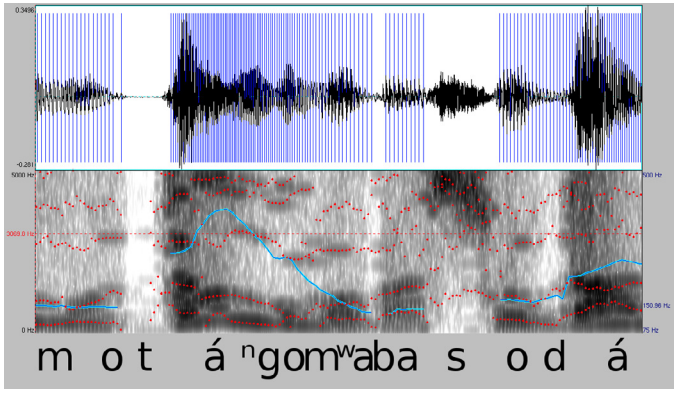


HASHING

Eine Grundlage vieler Inhaltserkennungstechnologien ist Hashing. Ein Hashing-Algorithmus berechnet aus den Daten einer Audio- oder Videodatei einen Wert. Dabei soll ein Hashwert nur genau einem Eingabedatum zugeordnet werden können. Hashfunktionen werden z.B. zur Bildung von Prüfsummen angewendet, mit denen Veränderungen bei der Datenübertragung aufgespürt werden sollen.
In Datenbanken können Datensätze mittels Hashwerten sehr effizient gesucht werden. Viele Inhaltserkennungstechnologien arbeiten so, dass sie mittels eines Hashing-Algorithmus eine eindeutige Kompression eines Inhalts erzeugen und diesen mit einer Datenbank von als zulässig oder unzulässig eingeordneten Inhalten abgleichen. So können Inhalte, die Bit für Bit identisch sind, wiedererkannt werden.

FINGERPRINTING

Fingerprinting ist ein Verfahren, das auf Hashing aufbaut. Dabei wird ein Inhalt aber nicht bitweise gehasht, sondern auf einer semantischen Ebene erfasst und dann gehasht. Die Technologie verwendet psycho-sensorische Merkmale als Ausgangspunkt. Das führt dazu, dass die Erkennung unempfindlich gegenüber Veränderungen auf Bit-Ebene, wie der Auflösung, und auch gegenüber geringfügigen Veränderungen auf semantischer Ebene, etwa einem Rauschen in einer Audiodatei, einer Farbverschiebung in einem Video oder Ähnlichem, wird.[7] Diese Technologie orientiert sich also stärker an der menschlichen Wahrnehmung.
Z.B. sollen akustische Fingerabdrücke Audioinhalte maschinell wiedererkennbar machen. Als Ausgangspunkt wird dabei ein Spektrogramm eines Audiostücks, also eine Abbildung der



Spektrogramm einer weiblichen Stimme, die Worte in der Sprache Lingala spricht. Gemeinfrei via Wikimedia Commons.

Frequenz auf die Zeit, verwendet. Sodann werden die Frequenzspitzen markiert, denn diese machen die für das menschliche Gehör markanten Merkmale eines Klangs aus. Aus deren Positionen zueinander werden dann Hashwerte berechnet, mit denen der Audioabschnitt wiedererkannt werden kann, wenn man sie gegen eine ausreichend große Datenbank solcher Hashwerte abgleicht.[8] Bereits ein Sample von wenigen Sekunden kann genügen, um ein Lied wiederzuerkennen.

INHALTSERKENNUNG DURCH MASCHINELLES LERNEN

Durch maschinelles Lernen können Algorithmen in die Lage versetzt werden, aus einer großen Zahl von Beispielen (Trainingsdaten) ein Modell eines bestimmten Inhalts zu bilden und dieses Modell anschließend auf neue Daten anzuwenden. So können Algorithmen „lernen“, bestimmte Merkmale zu erkennen. Regeln zur Erkennung von Merkmalen oder Mustern werden dabei nicht vorgegeben.[9] So kann ein Algorithmus beispielsweise darauf trainiert werden, Fotos von Hunden von solchen von Katzen zu unterscheiden, eine Abbildung einer IS-Flagge oder einer Brustwarze zu erkennen.

Wozu werden Uploadfilter verwendet?

Seit der Einführung von automatischen Inhaltserkennungssystemen zur Verfolgung von urheberrechtlichen Ansprüchen hat sich der Anwendungsbereich stark ausgeweitet. Neben Urheberrechtsverletzungen sollen auch kinderpornografische und terroristische Inhalte maschinell (wieder-)erkannt werden. Auch in Sachen Hasskriminalität gelten solche Systeme zunehmend als Mittel der Wahl. Es mehren sich Stimmen, die die Anwendung von Filtern für alle möglichen unerwünschten Inhalte fordern.

URHEBERRECHT

Seit den 2000er Jahren ist die Verletzung von Urheberrechten auf Sharing-Plattformen ein Politikum. Die Plattformen verwenden verstärkt Inhaltserkennungssysteme, um den Rechteinhabern eine Verfolgung ihrer Ansprüche zu ermöglichen. Mit der europäischen Urheberrechtsreform von 2019 werden die Plattformen verpflichtet, Uploadfilter zu installieren, die die Veröffentlichung geschützter Inhalte im Vorhinein verhindern.

PROVIDERPRIVILEG

Die EU-Urheberrechtsrichtlinie von 2019 rührt an einem rechtlichen Prinzip, das 2000 in der europäischen E-Commerce-Richtlinie (2000/31/EG) verankert wurde und seither die Entwicklung von Sharing-Plattformen stark geprägt hat: dem Providerprivileg. Demnach gilt in der europäischen Union der Grundsatz, dass Plattformbetreibende nicht für Rechtsverletzungen durch Nutzende haften, es sei denn, sie wurden darauf aufmerksam gemacht, dass ein Inhalt rechtswidrigerweise hochgeladen wurde. Sie müssen nicht selbst nach Urheberrechtsverletzungen fahnden, sondern ihnen mit dem Notice-and-Take-down-Verfahren begegnen. Das bedeutet: Die Rechteinhaberin oder der Rechteinhaber wendet sich an die Plattform und teilt mit, dass ein Inhalt ohne Erlaubnis hochgeladen wurde. Daraufhin sperrt die Plattform den Inhalt (Take-down).

ARTIKEL 17

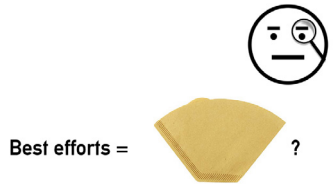
Dies ändert sich gerade: Die EU-Urheberrechtsrichtlinie, die im April 2019 beschlossen wurde und bis 2021 umgesetzt werden muss, enthält den umstrittenen Artikel 17.
Artikel 17 (vormals 13) Absatz 1 macht die Plattformen selbst zu den für die auf ihnen publizierten Inhalte Verantwortlichen. Er schafft die neue Verpflichtung für Plattformen, vor der Veröffentlichung von den Inhabern der Rechte an Musik, Videos, Fotos usw., die ihre Nutzer hochladen, eine Erlaubnis einzuholen, meist eine vergütungspflichtige Lizenz.
Wird ein Inhalt hochgeladen, für den keine Autorisierung der Berechtigten vorliegt, so haftet die Plattform, das heißt, sie muss z.B. Schadenersatz zahlen. Artikel 17 Absatz 4 sieht jedoch eine letzte Möglichkeit vor, der Haftung zu entgehen.

Dazu müssen drei Voraussetzungen gleichzeitig erfüllt sein:

die Plattform muss mit allen Anstrengungen versucht haben, eine Erlaubnis der Berechtigten einzuholen und

die Plattform muss alle **Anstrengungen** unternommen haben, um sicherzustellen, dass bestimmte Werke, die zuvor mit Erkennungsinformationen von den Berechtigten gemeldet wurden, nicht verfügbar sind und

die Plattform muss nach Erhalt eines Hinweises der Berechtigten den Inhalt sofort entfernt haben und alle Anstrengungen unternommen haben, um das erneute Hochladen zu verhindern.



→ Sobald also ein einziger Rechteinhaber oder eine einzige Rechteinhaberin sich dazu entschließt, Erkennungsinformationen seiner oder ihrer Werke, z.B. einen Fingerprint oder den Inhalt selbst, an eine Plattform

Foto mit freundlicher Genehmigung von Marcel Felde

zu senden, damit diese das Werk unverfügbar macht, ist die Plattform gezwungen, mit allen zur Verfügung stehenden Mitteln die Veröffentlichung dieses Werks zu verhindern. Dazu muss sie sämtliche Uploads überwachen. Aus technischer Sicht kann unter „allen Anstrengungen", mit denen sichergestellt wird, dass bestimmte Inhalte nicht verfügbar sind, nichts anderes zu verstehen sein als die Einrichtung automatischer Filtersysteme, die urheberrechtlich geschützte Inhalte erkennen und blockieren sollen. Darin sind sich Juristinnen und Juristen einig.[10] Da die Nichtverfügbarkeit sichergestellt werden muss, muss jeder Upload vor der Publikation geprüft werden.

KÖNNEN FILTER DAS?

Wie gezeigt können Inhaltserkennungssysteme einen Inhalt, den man einmal in eine Datenbank eingepflegt hat, wiedererkennen. Artikel 17 verlangt Plattformen viel mehr ab: Sie sollen vollautomatisiert entscheiden, ob die Nutzung eines Inhalts in einem konkreten Fall urheberrechtlich zulässig ist oder nicht. Dazu reicht es nicht, eine Übereinstimmung eines Inhalts mit einem, der in einer Datenbank hochgeladen wurde, festzustellen. Man muss überprüfen, ob es sich um ein erlaubtes Zitat, eine Parodie oder eine kritische Auseinandersetzung mit einem Inhalt handelt. Automatisierte algorithmische Systeme können jedoch den Kontext von Inhalten nicht erfassen und die rechtlich gebotenen Güterabwägungen nicht durchführen. Deshalb können sie diese Entscheidung nicht treffen.

„ZENSURHEBERRECHT"?

Urheberrecht schützt die berechtigten wirtschaftlichen Interessen von Kunstschaffenden. Es kann jedoch von den Berechtigten auch instrumentalisiert werden, um die Verbreitung von unerwünschten Veröffentlichungen zu verhindern. Das könnte mit automatischen Filtern noch viel effizienter bewerkstelligt werden.

> Nachdem der Medienblog Übermedien ein Video über Spekulationen der Zeitung Bild zum Attentat in Hanau im Februar 2020 auf YouTube veröffentlicht hatte, ließ der Axel Springer Verlag das Video kurzer Hand sperren – es waren Abbildungen von Überschriften aus der Bild enthalten, die hier als zulässige Zitate verwendet wurden.

TERRORISTISCHE INHALTE

Im Zuge der Ausbreitung des sogenannten Islamischen Staates, der Social-Media-Plattformen rasch für seine Propagandazwecke einzusetzen wusste, nahmen Sicherheitspolitikerinnen und Sicherheitspolitiker Inhalte von Terrororganisationen, die z.B. auf Videoplattformen verfügbar waren, immer stärker als Problem wahr. Bereits 2015 startete die EU-Kommission das EU Internet Forum, einen Kooperationsrahmen zwischen den nationalen Innenmi-

nistern, Europol und Vertretern der Internetwirtschaft. Mittlerweile sind die meisten großen Internetdienstleister beteiligt. Sie arbeiten seit 2016 an einer Datenbank, die von den Unternehmen gemeinschaftlich mit Fingerprints als terroristisch klassifizierter Inhalte gefüttert wird. Sie wird mittlerweile vom Global Internet Forum to Counter Terrorism (GIFCT) betrieben, einem Zusammenschluss, der ursprünglich von Facebook, Microsoft, Google und Twitter ins Leben gerufen wurde. Was Inhalt dieser Datenbank ist, wie ein Inhalt dort hineinkommt und ob ein fälschlich als Terror-Propaganda eingeordnetes Video jemals wieder heraus gelangt, ist für die Öffentlichkeit nicht nachvollziehbar.

Im April 2018 forderten der Bundesinnenminister Horst Seehofer und der damalige französische Innenminister Gérard Collomb in einem Schreiben an die EU-Kommission ein schärferes Vorgehen gegen Terror-Propaganda im Netz: Terroristische Inhalte sollten binnen einer Stunde nach Veröffentlichung entfernt werden, dies müsse durch Sanktionsmechanismen sichergestellt werden. Dabei sollten terroristische Inhalte nur den Anfang machen, zukünftig wünschen sich die Innenminister, dass auch gegen sonstige rechtswidrige Inhalte mit diesen Mitteln vorgegangen wird.

Im September 2018 veröffentlichte die Kommission einen Verordnungsentwurf, der das umsetzen soll. Dabei sollen auch Uploadfilter verpflichtend eingesetzt werden. Die Verordnung wird derzeit im Trilog zwischen Parlament, Kommission und Rat verhandelt. Das Parlament steht der Filterpflicht eher kritisch gegenüber.

Nach dem Livestream des Attentats im neuseeländischen Christchurch im März 2019, dessen Aufzeichnung noch einige Tage auf sozialen Medien verteilt wurde, forderte eine Vielzahl von Regierungsvertretenden weltweit, dass Inhalte dieser Art so schnell wie möglich völlig aus dem Internet verschwinden sollen.[11]

Das Löschen von als terroristisch eingestuften Inhalten, vor allem automatisiert, ist aus mehreren Gründen hochproblematisch:

> **Was Terrorismus ist, liegt nicht in der Natur der Sache, sondern ist Gegenstand politischer Debatten. Zum Beispiel stand Nelson Mandela bis 2008 auf US-Terrorlisten. Autoritäre Regierungen nennen Oppositionelle oftmals Terroristen. Andererseits standen terroristische Taten von Rechts, etwa die des NSU, lange Zeit völlig außerhalb des Blickfelds von Terrorismusbekämpfung.**
>
> **Oftmals wird die Darstellung von gewalttätigen Anschlägen unterbunden. Die Darstellung von Anschlägen kann affirmativ erfolgen, sie kann aber auch aus schützenswerten Gründen geschehen: als Berichterstattung oder zur Dokumentation von (Kriegs-)Verbrechen. Eine solche Einordnung kann unmöglich maschinell durchgeführt werden.**

HASSKRIMINALITÄT

Beleidigungen, Bedrohungen, Holocaust-Leugnung, die Billigung von Straftaten, das Verbreiten verfassungsfeindlicher Symbole: Politisch motivierte Straftaten in sozialen Netzwerken, bei denen die Betroffenen aufgrund ihrer Gruppenzugehörigkeit zum Ziel werden, sind derzeit in vielen europäischen Ländern Gegenstand von Regulierungsbemühungen. Gleichzeitig sind die Begriffe Hasskriminalität und Hassrede unscharf und spiegeln keinen fixen Bestand an Straftatbeständen wider. Die Strafverfolgung verläuft häufig schleppend oder findet gar nicht statt. In Deutschland verpflichtet das Netzwerkdurchsetzungsgesetz (NetzDG) soziale Netzwerke, „offensichtlich“ rechtswidrige Inhalte zu löschen, wenn sie ihnen gemeldet werden und unter einen der im Gesetz genannten Straftatbestände fallen. Eine Pflicht zur automatisierten Suche nach solchen Inhalten besteht bisher nicht. Forderungen, Hasspostings automatisch zu erkennen, werden aber lauter. Besondere Bedeutung erlangte der Fall der österreichischen Politikerin Eva Glawischnig-Piesczek.[12] Sie hatte nach heftigen Beleidigungen auf Facebook von dem Unternehmen nicht

nur die Löschung eines einzelnen Posts gerichtlich verlangt, sondern auch sämtlicher wortgleicher und sinngleicher weiterer Postings. Der Europäische Gerichtshof entschied, dass es nicht gegen das Verbot allgemeiner Überwachungspflichten in der e-Commerce-Richtlinie verstößt, wenn das österreichische Recht Facebook dazu verpflichtet. Das Gericht hat dabei sogar ausdrücklich auf „EDV-gestützte Hilfsmittel" hingewiesen, die der Plattform zur Verfügung stünden.
Inhaltsgleiche Postings können leicht automatisiert gefunden werden. Wie aber kann ein Algorithmus damit umgehen, dass jemand eine Äußerung zitiert, um über einen Fall zu berichten? Dieselben Worte können in einem Kontext beleidigend sein und in einem anderen nicht.[13] Noch komplexer wird die Erkennung sinngleicher Äußerungen. Die Trennlinie zwischen einer strafbaren Beleidigung und einer Äußerung, die gerade noch von der Meinungsfreiheit gedeckt ist, ist oftmals dünn, gerade wenn es um Aussagen gegenüber Politikerinnen und Politikern geht. Diese Abgrenzung erfordert ein Verständnis des Zusammenhangs des Gesagten und eine wertende Einordnung der betroffenen Rechtsgüter. Auch birgt die automatische Einstufung von Inhalten als Hasskommentare die Gefahr, bereits marginalisierte Gruppen zu diskriminieren, wie einige Studien belegen.[14] Automatische Spracherkennungssysteme arbeiten in thematisch eng begrenzten Gebieten am besten. Ein einziges Werkzeug auf die Fülle von Inhalten, die in sozialen Netzwerken besprochen werden, loszulassen, ist daher wenig aussichtsreich.

KINDERPORNOGRAFIE

Um die Verbreitung kinderpornografischer Darstellungen zu verhindern, entwickelte Microsoft zusammen mit dem Darthmouth College das System PhotoDNA. Es basiert auf Fingerprinting. Eine große Hash-Datenbank wird vom National Center for Missing & Exploited Children geführt. Das System kann von Plattformen als Plug-In implementiert werden.[15]

ILLEGALE INHALTE?

Bereits 2017 erwog die EU-Kommission, die Verantwortlichkeit von Plattformen für die auf ihnen veröffentlichten Inhalte auszudehnen und dabei „illegale Inhalte" im Allgemeinen zu adressieren.[16] Nach der Regulierung der Plattformhaftung für Urheberrechtsverletzungen und terroristische Inhalte wird in den nächsten Jahren ein Rundumschlag erwartet: In einem Kommissionspapier, das netzpolitik.org im Sommer 2019 leakte, wurden erste Pläne zu einem Rechtsakt über digitale Dienste (Digital Services Act) bekannt, der unter anderem die Regulierung von Hasskriminalität und Onlinewerbung vereinheitlichen soll. Darin ist von „illegalen" und „schädlichen" Inhalten die Rede. Laut diesem Papier soll das Verbot allgemeiner Überwachungspflichten aus der e-Commerce-Richtlinie zwar bewahrt werden. Dennoch werden „proaktive Maßnahmen" gefordert, ein Ausdruck, unter dem in der Regel automatisierte Filter zu verstehen sind, in jedem Fall aber Maßnahmen, die über ein Reagieren auf Sperranforderungen Dritter hinausgehen. Automatisierte Systeme werden insofern explizit erwähnt, als dass Regeln für ihre Transparenz und Zuverlässigkeit aufgestellt werden sollen – „wo diese eingesetzt werden".[27]
Denkbar wäre also schlimmstenfalls, dass Sharing-Plattformen für sämtliche Rechtsverstöße ihrer Nutzenden haften würden und Uploadfilter für Rechtsverstöße aller Art bereithalten müssten.

Was ist das Problem?

FAILS – KOLLATERALSCHÄDEN DURCH FEHLFILTERUNGEN

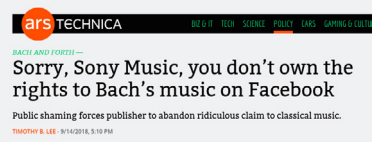
Es gibt verschiedene Projekte, in denen unter anderem falsche Sperrungen durch automatische Tools gesammelt werden – etwa die Lumen Database[18] oder die „Takedown Hall of Shame" der Electronic Frontier Foundation[19]. Einige Konstellationen tauchen dabei immer wieder auf:

→ Ein Werk wird innerhalb einer Fernsehsendung wiedergegeben (z.B. ein Musikvideo im Frühstücksfernsehen). Der Fernsehsender meldet die Sendung routinemäßig bei den Plattformen als geschützt. Damit wird auch das enthaltene Musikvideo, dessen Rechte möglicherweise bei anderen Personen liegen, dem Rechteinhaber der Fernsehsendung zugerechnet und herausgefiltert, wenn es von anderen hochgeladen wird.

> RTL hat ein freilizenziertes Video des Kollektivs Pinkstinks gesendet und die gesamte Sendung bei YouTube als urheberrechtlich geschützt gemeldet. Daraufhin wurde das Video auf dem Kanal von Pinkstinks gesperrt.

→ Eine Aufführung eines gemeinfreien Werks (z.B. eines Musikstücks, dessen Komponist schon länger als 70 Jahre verstorben ist) wird bei einem Filtersystem als urheberrechtlich geschützt angemeldet. Alle anderen Aufnahmen desselben Werkes, deren Rechte anderen Interpreten zustehen, werden als Urheberrechtsverletzung gesperrt.

> Sony Music Entertainment hatte 2018 gegenüber Facebook eine Interpretation eines Musikstücks von Bach als urheberrechtlich geschützt angemeldet. Facebook sperrte daraufhin den Ton eines Videos des Musiker James Rhodes, das ihn zeigte, wie er selbst dieses Musikstück spielte. Sogar eine erste Beschwerde des Musikers wurde von Sony zurückgewiesen.



Screenshot via arstechnica.com

→ Erlaubte Zitate von Werken werden regelmäßig als Urheberrechtsverletzungen eingeordnet. Häufig sind davon ironischerweise Urheberrechtsvorlesungen betroffen – sie arbeiten natürlich mit Zitaten, wenn sie Beispiele für Urheberrechtsverletzungen zeigen.

→ Legitime Darstellungen von Gewalt oder Nacktheit. Viele Plattformen blockieren Gewaltdarstellungen, terroristische Inhalte oder Nacktheit. Häufig werden Darstellungen dieser Art aber berechtigt verwendet.

> Das vielleicht bekannteste Beispiel ist das Foto des Napalm-Mädchens aus dem Vietnam-Krieg von 1972, ein wichtiges zeitgeschichtliches Dokument, das Facebook 2016 gesperrt hat, weil die Plattform die Darstellung nackter Kinder untersagt.

> Die syrische NGO The Syrian Archive sammelt unter anderem Videos zur Dokumentation von Kriegsverbrechen. YouTube sperrte diese Videos gleich zu Tausenden, weil sie als terroristische Inhalte klassifiziert wurden.

→ Filter, die sich gegen ihre Verwender wenden

> Warner Bros. hatte 2016 Google notifiziert, einige seiner eigenen URLs wegen Urheberrechtsverletzungen aus der Suche auszuschließen.[20]

OVERBLOCKING

Im Falle einer Urheberrechtsverletzung schwebt das Risiko einer Klage über den Plattformen. Mit der geplanten Verordnung über terroristische Inhalte drohen ihnen hohe Bußgelder, wenn sie einen Inhalt rechtswidrig nicht entfernen. Für den Fall, dass Inhalte fälschlich gelöscht werden, haben die Plattformen hingegen kaum Konse-

quenzen zu befürchten. Derzeit setzen die Regulierungen also Anreize, im Zweifel eher zu sperren. Deshalb ist zu erwarten, dass es zu Overblocking kommt. Das meint, dass die Plattformen aus Vorsicht deutlich mehr Inhalte blockieren als eigentlich nötig. Gerade kleinere Anbieter fürchten Rechtsstreitigkeiten und können es sich kaum leisten, es „darauf ankommen" zu lassen.

ZENSURINFRASTRUKTUR

Filter, die vor oder nach der Veröffentlichung bestimmte Inhalte erkennen und blockieren, sind auf den großen Plattformen schon recht verbreitet. Mit der Tendenz zur umfassenden gesetzlichen Filterpflicht besteht aber die Gefahr, dass sich eine Filterinfrastruktur etabliert, die das ganze Internet umfasst und um die herumzukommen immer schwieriger wird. Je mehr Anbieter dabei auf dieselben technischen Systeme oder dieselben Hash-Datenbanken zurückgreifen, desto größer werden die Auswirkungen einer Fehleinordnung. Eine solche Infrastruktur kann – wenn die Betreibenden unter ausreichenden Druck geraten – systematisch dazu verwenden werden, politisch unliebsame Inhalte zu verdrängen. Auch hierbei wäre der Effekt umso schwerwiegender, je mehr Plattformen dieselben Systeme verwenden.

WHITELIST STATT BLACKLIST

Bisherige Filtersysteme suchen nach den Inhalten, die nicht auf Plattformen erscheinen sollen – sie nehmen eine Blacklist-Filterung vor. Dieses Prinzip könnte leicht in sein Gegenteil verkehrt werden: Eine Whitelist-Filterung, also ein Internet, in dem jeder veröffentlichte Inhalt seine „Identität", etwa eine Lizenz, vorweisen müsste. Dies würde die Vielfalt und Freiheit des Internets auf ein Minimum beschränken.

DATENSCHUTZ

Je nach Gestaltung können Uploadfilter erhebliche datenschutzrechtliche Implikationen haben. Sollten kleinere Plattformen auf Angebote größerer Anbieter zurückgreifen, wäre zu befürchten, dass über letztere ein nicht unerheblicher Teil des Internetverkehrs laufen wird, gab der Bundesdatenschutzbeauftragte Ulrich Kelber im Februar 2019 zu bedenken.[21] Damit würden schlimmstenfalls Anbieter wie YouTube und Facebook auch Daten über Uploads auf einer Vielzahl anderer Plattformen erhalten. Entscheidungen, die auf ausschließlich automatischer Verarbeitung beruhen und Personen erheblich beeinträchtigen können, erlaubt Artikel 22 der Datenschutz Grundverordnung außerdem nur unter erhöhten Voraussetzungen.

Die Zukunft der Filter

Die Haftungsprinzipien, die Sharing-Plattformen in den letzten Jahren prägten, werden in der EU gegenwärtig neu verhandelt. Nicht nur urheberrechtlich geschütztes Material, als terroristisch klassifizierte Inhalte und Hasskommentare sollen nach Auffassung der EU-Kommission und einiger Mitgliedstaaten aus dem Netz verschwinden. Mit dem Digital Services Act könnte dieses Prinzip auf alle Internetdienste, alle Gefahrenlagen und alle illegalen und möglicherweise auch „schädlichen" Inhalte ausgedehnt werden. Uploadfilter sind nicht die einfache technische Lösung, mit denen dies umgesetzt werden kann. Während automatische Inhaltserkennungstechnologien Videos, Audios, Texte oder andere Inhalte identifizieren können, gibt es keine Technologie, die juristische Abwägungsentscheidungen automatisiert vornehmen kann – und damit die Entscheidung zwischen legal und illegal treffen könnte. Mit der Pflicht zu Uploadfiltern wird eine neue Infrastrukturschicht in all jene Bereiche des Internets eingezogen, die es allen Menschen ohne Aufwand ermöglichen, eigene Werke zu veröffentlichen.

Fußnoten

1 Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG, Abl L 130/92, 17.5.2019.

2 Kommission: Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Verhinderung der Verbreitung terroristischer Online-Inhalte, COM(2018) 640 final, 12.9.2018.

3 YouTube-Hilfe: So funktioniert Content ID, https://support.google.com/youtube/answer/2797370?hl=de, Abruf 6.4.2020.

4 Facebook: Rightsmanager, https://rightsmanager.fb.com/, Abruf 6.4.2020.

5 Open nsfw model, via Github, https://github.com/yahoo/open_nsfw, Abruf 6.4.2020.

6 Wojcicki: Expanding our work against abuse on our platform, https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html, 4.12.2017.

7 ACRCloud Docs: Audio Fingerprinting, https://docs.acrcloud.com/docs/acrcloud/introduction/audio-fingerprinting/, Abruf 6.4.2020.

8 Chirp: Audio fingerprinting — what is it and why is it useful? https://medium.com/chirp-io/audio-fingerprinting-what-is-it-and-why-is-it-useful-33c6cc6bc302, 15.3.2018.

9 Döbel et. al.: Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung, https://www.bigdata.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf, 2018.

10 Spindler: Gutachten zur Urheberrechtsrichtlinie (DSM-RL), Europarechtliche Vereinbarkeit (Artikel 17), Vorschläge zur nationalen Umsetzung und zur Stärkung der Urheberinnen und Urheber, https://www.gruene-bundestag.de/fileadmin/media/gruenebundestag_de/themen_az/netzpolitik/pdf/Gutachten_Urheberrechtsrichtlinie_01.pdf, 14.12.2019.

11 The Ministry of Foreign Affaiers and Trade New Zealand: Christchurch Call, https://www.christchurchcall.com/index.html, Abruf 6.4.2020.

12 Fanta: EuGH-Urteil zu Facebook: Der freien Meinungsäußerung droht Schiffbruch, https://netzpolitik.org/2019/eugh-urteil-zu-facebook-der-freien-meinungsaeusserung-droht-schiffbruch/, 5.10.2019.

13 Keller: The CJEU's new filtering case, the Terrorist Content Regulation, and the future of filtering mandates in the EU, https://cyberlaw.stanford.edu/blog/2019/12/cjeu%E2%80%99s-new-filtering-case-terrorist-content-regulation-and-future-filtering-mandates-eu, 2.12.2019.

14 Sap et al: The Risk of Racial Bias in Hate Speech Detection, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678 Florence, Italy, 28. 7. - 2.8.2019.

15 Microsoft: Photo DNA, https://www.microsoft.com/en-us/PhotoDNA/CloudService, Abruf 6.4.2020.

16 EU-Kommission: Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms, COM(2017) 555 final, 28.9.2017.

17 EU-Kommission, Leak via netzpolitik.org, https://cdn.netzpolitik.org/wp-upload/2019/07/Digital-Services-Act-note-DG-Connect-June-2019.pdf, Abruf 6.4.2020.

18 https://www.lumendatabase.org/.

19 Electronic Frontier Foundation: Takedown Hall of Shame, https://www.eff.org/takedowns, Abruf 6.4.2020.

20 Torrentfreak: Warner Bros. Flags Its Own Website as a Piracy Portal, https://torrentfreak.com/warner-bros-flags-website-piracy-portal-160904/, 4.9.2016.

21 Kelber: Reform des Urheberrechts birgt auch datenschutzrechtliche Risiken, https://www.bfdi.bund.de/DE/Infothek/Pressemitteilungen/2019/10_Uploadfilter.html, 26.2.2019.

Der Kampf für digitale Grundrechte ist nicht umsonst! Unterstütze die Digitale Gesellschaft e.V.

Die Digitale Gesellschaft (DigiGes) ist ein gemeinnütziger Verein, der sich seit seiner Gründung im Jahr 2010 für Grundrechte und Verbraucherschutz im digitalen Raum einsetzt. Zum Erhalt und zur Fortentwicklung einer offenen digitalen Gesellschaft engagiert sich die netzpolitische Organisation gegen den Rückbau von Freiheitsrechten im Netz und für die Realisierung digitaler Potentiale bei Wissenszugang, Transparenz, Partizipation und kreativer Entfaltung. Die Digitale Gesellschaft ist ein Verein mit zwei hauptamtlichen Mitarbeiterinnen und lebt sonst ausschließlich vom Engagement seiner Mitglieder.

Unterstütze uns mit einer Spende oder werde Fördermitglied!

https://digitalegesellschaft.de/unterstuetzen/
https://digitalegesellschaft.de/foerdermitglied/
IBAN: DE88 4306 0967 1125 0128 00
BIC: GENODEM1GLS

Redaktionsschluss: 5. Mai 2020

Herausgeber:
Digitale Gesellschaft e.V.
Groninger Straße 7
13347 Berlin

Text, Redaktion und ViSdP: Elisabeth Niekrenz

Layout: Katrin Brunk