ISD | Institute for Strategic Dialogue

Gefördert durch:

Auswärtiges Amt

aufgrund eines Beschlusses
des Deutschen Bundestages

Neue Plattformen und Technologien

Ein Überblick zur aktuellen Bedrohungslage und ihrer politischen Implikationen

Mauritius Dorn, Sara Bundtzen,
Christian Schwieter & Milan Gandhi

Über das Digital Policy Lab

Als zwischenstaatliche Arbeitsgruppe engagiert sich das Digital Policy Lab (DPL) dafür, politische Lösungen zur Verhinderung und Bekämpfung der Verbreitung von Desinformation, Hassrede sowie extremistischen und terroristischen Inhalten im Internet aufzuzeigen. Die Arbeitsgruppe besteht aus Vertreter:innen der zuständigen Ministerien und Aufsichtsbehörden ausgewählter liberal-demokratischer Länder. Die Arbeit des DPL zielt darauf ab, den regierungsübergreifenden Dialog zu fördern, politischen Entscheidungsträger:innen und Aufsichtsbehörden Zugang zu einschlägigem Fachwissen und Forschungsergebnissen zu verschaffen sowie eine internationale Arbeitsgemeinschaft zur Bewältigung der wichtigsten digitalpolitischen Herausforderungen aufzubauen. Wir danken dem Auswärtigen Amt für die Unterstützung des Projekts.

Über diesen Bericht

Im Rahmen des DPL organisierte das Institute for Strategic Dialogue (ISD) zwischen **Mai und Juni 2023 zwei Arbeitsgruppentreffen zum Thema Neue Plattformen und Technologien**. Die Arbeitsgruppe bestand aus Vertreter:innen nationaler Ministerien und Aufsichtsbehörden, unter anderem aus Australien, Deutschland, Neuseeland, Großbritannien, der Slowakei, den USA und der Europäischen Union. Zu den Teilnehmer:innen gehörten zudem Vertreter:innen der Zivilgesellschaft und der Wissenschaft. Auch wenn die Teilnehmer:innen an diesen Treffen teilgenommen und/oder an diesem Bericht mitgewirkt haben, spiegeln die darin geäußerten Ansichten nicht unbedingt die Ansichten aller Teilnehmer:innen oder der an diesem Projekt beteiligten Regierungen wider.

Über die Autor:innen

Mauritius Dorn ist Senior Digital Policy & Education Manager beim ISD Germany. Er leitet das Projekt AHEAD – eine Dialogreihe, die Politikschaffenden ein integriertes Verständnis hybrider Bedrohungen mit Schwerpunkt auf Desinformation vermitteln will. Er unterstützt das DPL als Experte für internationale Digitalpolitik.

Sara Bundtzen ist Analystin beim ISD Germany. Sie erforscht die Verbreitung von Informationsmanipulation in mehrsprachigen Online-Umgebungen. Im Rahmen des DPL analysiert Sara politische Wege zur Bekämpfung von Desinformation, Hass und Extremismus.

Christian Schwieter ist Fellow beim ISD und Doktorand an der Abteilung für Medienwissenschaften der Universität Stockholm. Bis 2023 war er Project Manager beim ISD Germany und leitete das Forschungsprojekt »Radikalisierung in rechtsextremen Online-Subkulturen entgegentreten«.

Milan Gandhi ist Research Fellow (KI und Politik) am ISD und Master-Student an der Universität Oxford. Er ist auch der Gründer von Legal Forecast, einer gemeinnützigen Organisation, die sich mit der Schnittstelle zwischen Recht und neuen Technologien beschäftigt.

Herausgeberische Verantwortung:
Huberta von Voss, Executive Director, ISD Germany & Henry Tuck, Head of Digital Policy, ISD

Danksagungen

Wir bedanken uns bei allen Teilnehmer:innen der Arbeitsgruppe, einschließlich der Expert:innen aus Wissenschaft und Zivilgesellschaft, für ihre Beiträge. Besonderer Dank gilt den Referent:innen sowie allen Mitwirkenden an diesem Bericht für ihre wertvollen Einblicke und ihr Feedback: Diederik Don (Europol), Dr. Elena Gubenko (Bundesamt für Justiz, Deutschland), Adam Hadley (Tech Against Terrorism), Dominik Hammer (ISD Germany), Dr. Oliver Marsh (CASM Technology), und Jenna Omassi (Ofcom).

ISD | Institute for Strategic Dialogue

www.isdgermany.org

Inhaltsverzeichnis

Glossar

ActivityPub ist ein offenes und dezentrales Netzwerkprotokoll. Als offenes Protokoll gehört es keiner bestimmten Firma und ist nicht auf bestimmte Produkte limitiert. Es bietet Client-to-Server und Server-to-Server APIs. ActivityPub bildet einen Standard fürs Fediversum.

Application Programming Interfaces (APIs) sind Softwareschnittstellen, die eine Kommunikation zwischen zwei Anwendungen ermöglichen. APIs haben eine riesige Vielzahl an Verwendungszwecken, aber im Zusammenhang mit diesem Bericht ermöglichen sie Forscher:innen den Zugriff auf bestimmte Daten von Online-Plattformen über Datenanfragen. Als zwischengeschaltete Instanz stellen APIs eine zusätzliche Sicherheitsebene bereit, indem sie keinen direkten Zugriff auf Daten zulassen und das Volumen und die Häufigkeit der Anfragen protokollieren, verwalten und überwachen.

BitTorrent ist ein Peer-to-Peer-basiertes Filesharing-Protokoll, das für das Verteilen großer Datenmengen verwendet wird, weil es die Serverbelastung schont. Beim Herunterladen einer Datei via BitTorrent-Technologie wird die Datei nicht als Ganzes, sondern in Datenstücken, von allen mit dem Netzwerk verbundenen Geräten übertragen.

Blockchain-Technologie, die insbesondere für alternative Währungen entwickelt wurde, zeichnet sich durch ihre besondere Datenstruktur aus, die aufgrund ihrer Transparenz und ihres dezentralisierten Aufbaus als besonders fälschungssicher gilt. Daten werden dort an vielen verschiedenen Orten gespeichert und regelmäßig verglichen. Blockchain-Technologie ermöglicht (pseudo-)anonyme Transaktionen und Kommunikation – ein Umstand, der die Technologie auch für Kriminelle und Extremist:innen attraktiv macht.

Verschwörungsideologien sind Erklärungsversuche für bestimmte Phänomene, indem ein finsteres, von mächtigen Akteuren inszeniertes Komplott unterstellt wird. Verschwörungen werden als geheim oder esoterisch dargestellt, wobei sich die Anhänger:innen eines Erklärungskonstrukts als die wenigen Eingeweihten sehen, die Zugang zu verborgenem Wissen haben. Die Anhänger:innen von Verschwörungsideologien sehen sich in der Regel in direkter Opposition zu den Mächten, die das Komplott inszenieren, bei denen es sich in der Regel um Regierungen oder Autoritätspersonen handelt.

Deplatforming bezeichnet die Sperrung von Accounts und Gruppen auf den sozialen Medien. Es führt regelmäßig dazu, dass Reichweite für Agitation verloren geht und Einnahmequellen wegbrechen. Zugleich haben Deplatforming und die Angst vor der Sperrung oder Löschung von Konten und Webseiten dazu beigetragen, dass alternative Online-Plattformen entstanden sind.

Desinformation bezeichnet die absichtliche Verbreitung von falschen oder irreführenden Inhalten, um zu täuschen, wirtschaftliche und/oder politische Vorteile zu erzielen, und einen Schaden für die Öffentlichkeit zu verursachen. Wenn wir uns auf solche Informationen beziehen, die unabsichtlich verbreitet werden, verwenden wir den Begriff **Fehlinformationen**.

Extremismus gilt als die Befürwortung einer Weltanschauung, die die Überlegenheit und Dominanz einer identitätsbasierten »Eigengruppe« (*in-group*) über alle »Fremdgruppen« (*out-groups*) propagiert. Extremismus beinhaltet somit eine Dehumanisierung der »Anderen« (*othering*), die mit Pluralismus und universellen Menschenrechten unvereinbar ist. Extremistische Gruppen verfolgen und befürworten einen systemischen politischen und gesellschaftlichen Wandel, der ihre Weltanschauung widerspiegelt. Sie können dies sowohl mit gewaltlosen und subtileren Mitteln als auch mit gewalttätigen oder expliziten Mitteln tun. Extremismus kann sowohl von staatlichen als auch von nichtstaatlichen Akteuren befürwortet werden.

Das **Fediversum** ist ein Versuch, eine dezentralisierte Alternative zu großen sozialen Netzwerken zu erstellen. Zum Fediversum gehören Mikroblogging-, Video- und Image-Sharing-Dienste. Sofern die Dienste das gleiche Netzprotokoll verwenden, können die verschiedenen Server innerhalb des Fediversums miteinander kommunizieren.

Basismodelle sind eine neuere Entwicklung, bei der KI-Modelle auf der Grundlage von Algorithmen entwickelt werden, die im Hinblick auf Allgemeinheit und Vielseitigkeit der Ergebnisse optimiert wurden. Diese Modelle werden häufig auf der Grundlage eines breiten Spektrums von Datenquellen und großer Datenmengen trainiert, um eine Fülle nachgelagerter Aufgaben zu erfüllen, darunter auch solche, für die sie nicht speziell

entwickelt und trainiert wurden. Das Basismodell kann unimodal oder multimodal sein und durch verschiedene Methoden wie überwachtes Lernen oder bestärkendes Lernen trainiert werden.

Schädliche Inhalte und Verhaltensweisen sind ein breites Spektrum von Online-Aktivitäten, die negative Auswirkungen auf die Menschenrechte, die Gesellschaft und/oder die Demokratie haben können. Diese können von der gezielten Belästigung von Einzelpersonen über die Anstiftung zu Gewalt gegen eine bestimmte Gruppe bis hin zur Verbreitung von Desinformationen und schädlichen Verschwörungsideologien reichen. In einigen Fällen ist das Schadensrisiko bereits mit dem Inhalt selbst verbunden, wobei die Gefahr durch dessen Verbreitung noch verstärkt wird. In anderen Fällen wird die Schadensgefahr eher durch aggregierte Verhaltensmuster als durch die Art des Inhaltes selbst verursacht. Je nach geografischem und rechtlichem Kontext können verschiedene Formen schädlicher Inhalte oder Verhaltensweisen illegal sein oder nicht. Und je nach Plattform können schädliche Inhalte oder Verhaltensweisen auch durch die »Community-Richtlinien«, Standards oder Regeln eines Unternehmens abgedeckt sein oder nicht.

Unter **Hass** versteht das ISD Überzeugungen oder Praktiken, die eine ganze Gruppe von Menschen aufgrund geschützter Merkmale wie ethnische Zugehörigkeit, Religion, Geschlecht, sexuelle Orientierung oder Behinderung angreifen, verleumden, delegitimieren oder ausgrenzen. Hassakteure sind Einzelpersonen, Gruppen oder Gemeinschaften, die sich aktiv und offen an den oben genannten Aktivitäten beteiligen, ebenso wie diejenigen, die implizit Menschengruppen angreifen, beispielsweise durch die Verwendung von Verschwörungsideologien und Desinformation. Hasserfüllte Aktivitäten laufen dem Pluralismus und der universellen Anwendung der Menschenrechte zuwider.

Informationsmanipulation beschreibt ein meist nicht illegales Verhaltensmuster, das Werte, Verfahren und politische Prozesse bedroht oder das Potenzial hat, diese negativ zu beeinflussen. Derartige Aktivitäten haben einen manipulativen Charakter und werden absichtlich und koordiniert durchgeführt.

Instanz bezeichnet in diesem Kontext eine Online-Plattform, die durch PeerTube oder eine andere Fediversum-Software aufgesetzt wurde. Auf Instanzen können meist – wie bei herkömmlichen sozialen Netzwerken – Konten erstellt und Inhalte hochgeladen werden. Jede Instanz wird unabhängig verwaltet, kann aber durch optionale Vernetzungen mit anderen Instanzen kommunizieren.

Non-fungible tokens (NFTs) sind Dateneinheiten, die in einem digitalen Register gespeichert sind, das Aufzeichnungen über den Kauf führt und Fälschungen verhindert. Sie sind einzigartige Vermögenswerte in einer digitalen Welt, was bedeutet, dass sie wie materielle Güter verkauft und gekauft werden können und als virtueller Eigentumsnachweis angesehen werden können.

Radioaktive Daten beziehen sich auf Markierungen (Datenisotope), die durch den Lernprozess hindurch bestehen bleiben und mit hoher Sicherheit in einem neuronalen Netz erkannt werden können.

Social Bots arbeiten von Accounts auf Online-Plattformen aus. Sie sind Computerprogramme, die nach ihrer Aktivierung ohne menschliches Zutun automatisiert im Internet agieren und beispielsweise dazu eingesetzt werden, um Postings zu teilen, zu liken oder zu kommentieren.

Virtuelle Realität (VR) ist eine Technologie, die nahezu reale und/oder glaubwürdige Erfahrungen auf synthetische oder virtuelle Weise bietet, während **Augmented Reality (AR)** die reale Welt durch Überlagerung von computergenerierten Informationen erweitert. **Mixed Reality (MR)** bietet eine Erfahrung, bei der die reale Umgebung des Benutzers und digital erstellte Inhalte nahtlos ineinander übergehen, wobei beide Umgebungen nebeneinander bestehen und miteinander interagieren können. **Extended Reality (XR)** ist ein Sammelbegriff, der Technologien wie VR, AR und MR umfasst.

Executive Summary

Dieser Bericht gibt einen Überblick über die einschlägigen Erkenntnisse zu den Schadensrisiken neuer Plattformen und Technologien und zeigt eine Reihe politischer Implikationen auf. Im Folgenden werden »dezentralisierte«, »generative« und »immersive« Technologien hinsichtlich ihrer Auswirkungen auf Desinformation, Hass und Extremismus analysiert. Es ist jedoch anzumerken, dass die beschriebenen Trends voneinander abhängig sein können und dass die Konvergenz neuer Plattformen und Technologien rasch voranschreitet. In der Folge können auch die zugehörigen Risiken konvergieren. Dies wirft wichtige Fragen über die Manifestierung künftiger Schäden und das Potenzial für schwerwiegendere und extremere Auswirkungen auf. Daher müssen die politischen Entscheidungsträger:innen und die Regulierungsbehörden die spezifischen Schadensrisiken erkennen können und gezielte Initiativen unterstützen, um die Risiken entsprechend ihrer Beschaffenheit zu mindern. Darüber hinaus müssen sie bei der Entwicklung und Durchsetzung neuer und bereits bestehender Regelwerke konvergierende Risiken berücksichtigen. Hierfür wird eine länder- und funktionsübergreifende Zusammenarbeit zwischen politischen Entscheidungsträger:innen und Regulierungsbehörden entscheidend sein.

Allgemeine Bedrohungslage

- **Das dezentralisierte Social Web ist für rechtsextreme und konspirative Milieus zu einer förderlichen Umgebung geworden, um schädliche Inhalte zu verbreiten und schädliche Verhaltensweisen auszuüben.** PeerTube-Instanzen beispielsweise ermöglichen die vollständige individuelle Kontrolle über die Inhaltemoderation und die Verbreitung schädlicher Inhalte über verschiedene miteinander verbundene Dienste. Gleichzeitig werden Nutzer:innen mit abweichenden Ideologien aus dem Informationsraum verdrängt. Ebenso können böswillige Akteure auf Odysee ihre schädlichen Inhalte und Verhaltensweisen durch neue Optionen monetarisieren.

- **Große Sprachmodelle (*Large Language Models, LLMs*) können eine Vielzahl von unvorhersehbaren Ergebnissen erzeugen und bieten derzeit ein erhebliches Potenzial für die Ausnutzung durch böswillige Akteure.** Forscher:innen haben festgestellt, dass Trainingsdatensätze Vorurteile oder Stereotypen enthalten können und dass selbst einfache Eingabeaufforderungen schädliche Inhalte, einschließlich Fehlinformationen, erzeugen können. Technisch fortgeschrittene Akteure können auch die Code-Generierungsfunktion von LLMs zur Informationsmanipulation ausnutzen. Ein unmittelbarer Zugang zu Endanwendungen kann ebenfalls zur Täuschung der Endnutzer:innen beitragen.

- Extended Reality (XR) kann schwerwiegendere Varianten bestehender schädlicher Inhalte und Verhaltensweisen ermöglichen. So könnte XR ein noch nie zuvor gesehenes Maß an emotionaler Manipulation und überzeugenderer Propaganda (beispielsweise durch Avatare) ermöglichen und auch das Risiko von körperlichen Schäden erhöhen. Darüber hinaus kann sich XR zu einem Rückzugsort für böswillige Akteure zur Rekrutierung, Finanzierung und Planung von Aktivitäten entwickeln. XR bietet auch eine Reihe von Möglichkeiten für die Integration des dezentralisierten Social Web und von LLMs, einschließlich deren spezifischer Schadensrisiken.

Empfehlungen

- **Die politischen Entscheidungsträger:innen und Regulierungsbehörden müssen klären, welche bereits bestehenden regulatorischen Vorgaben auf dezentralisierte Social-Web-Dienste anwendbar sind und welche Ansätze zur Regeldurchsetzung in Frage kommen.** Die Verpflichtung der Diensteanbieter, gesetzliche Vertreter:innen im Inland zu benennen, kann als ein wichtiges politisches Element zur Erreichung einer Rechenschaftspflicht angesehen werden. Die Regeldurchsetzung muss sich auf eine bessere internationale Koordinierung und öffentlichen Druck stützen. Zudem müssen Regulierungsbehörden die Regeleinhaltung durch die Diensteanbieter proaktiv unterstützen (z. B. durch die Entwicklung von Compliance-Plugins).

- **Die Schadensrisiken von LLMs können bei einer Vielzahl von Endanwendungen auftreten. Um dem entgegenzuwirken, müssen politische Entscheidungsträger:innen und Regulierungsbehörden neue Regeln für den Zugriff, die Rechenschaftspflicht, die Haftung, die Sicherheit und die Erkennung von LLMs festlegen.** Selbstregulierung kann eine Zwischenlösung sein, bis neue Regeln in Kraft treten. Gleichzeitig müssen sich die Regulierungsbehörden der Taktiken und Techniken bewusst sein, die von böswilligen Akteuren eingesetzt werden, um LLMs für ihre Strategien auszunutzen (z. B. zur Informationsmanipulation).

- Die politischen Entscheidungsträger:innen und **Regulierungsbehörden müssen definieren, was ein Schadensrisiko in XR-Umgebungen darstellt, und gewährleisten, dass ein anwendbarer (Co-) Regulierungsrahmen besteht.** Zu diesem Zweck müssen sie bereits bestehende Regeln für Plattformen und Technologien (wie z. B. der europäische Digital Services Act (DSA) und AI Act, der australische Online Safety Act 2021) und nationale Strafgesetze überprüfen und unter Umständen eine XR-spezifische Taxonomie im Zusammenhang mit Schäden und Straftaten entwickeln. Darüber hinaus müssen in einem Multi-Stakeholder-Dialog Standards für die Sammlung von Beweisen, die Transparenzberichten und die Inhaltemoderation entwickelt werden.

Einleitung

Am 11. April 2022 veröffentlichte das Oberste Volksgericht der Volksrepublik China einen Fall, in dem der Anbieter eines KI-Systems das Bild einer Person ohne Erlaubnis nutzte und sich dadurch haftbar machte. Der Anbieter schuf eine virtuelle Figur als »KI-Begleiter«, die den Namen und das Portrait der Person benutzte, und ermöglichte die Erstellung von interaktiven Inhalten.[1] Der Fall veranschaulicht den anhaltenden Trend der zunehmenden Konvergenz digitaler Technologien mit Auswirkungen auf die Persönlichkeitsrechte. Selbst in China, wo der digitale Überwachungsstaat seit Jahren auf dem Vormarsch ist, wird zumindest die wirtschaftliche Nutzung persönlicher Daten zunehmend eingeschränkt. Der Informationsraum ist seit jeher durch technologische Innovation und Wechselbeziehungen zwischen Medien und Gesellschaft gekennzeichnet gewesen,[2] was neue Vorteile, aber auch Schadensrisiken mit sich gebracht hat. Gemeinsam mit der sich stetig verändernden Wahrnehmung und Wirkung der Medien haben sich gleichzeitig jedoch auch die Governance-Ansätze weiterentwickelt. Infolgedessen wurde auch der Rechtsrahmen in demokratischen Staaten in vielen Zusammenhängen verschärft, wie z. B. im Harmful Digital Communications Act (2015) in Neuseeland und durch das Netzwerkdurchsetzungsgesetz (2017) in Deutschland. Es ist noch nicht vollends geklärt, inwieweit sich diese neuen rechtlichen Rahmenbedingungen auf die Grundrechte ausgewirkt haben. Gleichsam werden neue Plattformen und Technologien wie das dezentralisierte Social Web, große Sprachmodelle (LLMs) und Extended Reality (XR) zunehmend von böswilligen Akteuren, einschließlich der extremistischen und propagandistischen Online-Ökosysteme, ausgenutzt, um schädliche Inhalte zu verbreiten und/oder schädliche Verhaltensweisen auszuüben. Dies wird auch von der Global Coalition for Digital Safety erkannt – diese bestätigt, dass die zuvor aufgeführten Technologien und Plattformen »neue Typen von Schäden hervorrufen oder bestehende Schadenstypen verschärfen könnten«.[3] Angesichts dieser sich wandelnden Bedrohungslage müssen politische Entscheidungsträger:innen und Regulierungsbehörden auf der ganzen Welt analysieren, inwieweit ihr Vorgehen gegen Schadensrisiken zukunftssicher ist. Aufbauend auf den Diskussionen einer Arbeitsgruppe des Digital Policy Lab (DPL)[4] zum Thema »Neue Plattformen und Technologien« zwischen Mai und Juni 2023 erörtert der vorliegende Bericht die Schadensrisiken und die politischen Implikationen neuer »dezentralisierter«, »generativer« und »immersiver« Plattformen und Technologien. Die vorliegende Analyse soll einen Überblick über mögliche negative Auswirkungen sowie Impulse für die Entwicklung der Regeln auf der Grundlage aktueller Forschungsergebnisse und Beispiele geben. Die Analyse beschränkt sich in erster Linie auf ausgewählte Schadensrisiken im Zusammenhang mit Desinformation, Hass und Extremismus. In drei Abschnitten werden die drei Typen von neuen Plattformen und Technologien untersucht: Abschnitt eins befasst sich mit dem dezentralisierten Social Web, Abschnitt zwei mit LLMs und Abschnitt drei mit XR. Jeder Abschnitt enthält Infoboxen, die zusätzliche Erklärungen zu den wichtigsten Begriffen liefern.

Dezentralisiert, generativ und immersiv: das fortschreitende Ökosystem der Online-Extremisten

Abschnitt 1: Dezentralisiertes Social Web

Die soziale Medienlandschaft erlebte in den letzten Jahren rasante Veränderungen. Während die marktbeherrschenden Plattformen wie Facebook, X (ehemals Twitter) und YouTube aufgrund ihrer Rolle bei der Verstärkung von Desinformation und Hass zunehmend in den Blickpunkt der Öffentlichkeit gerieten, entwickelten sich alternative digitale Dienste. Einige dieser neuen Dienste wie die Videoplattform Odysee haben das Ziel, einen sicheren Zufluchtsort für verschwörungsideologische, rechtsextremistische und andere böswillige Akteure zu schaffen, die das Gefühl haben, dass ihre Inhalte auf großen Online-Plattformen zu Unrecht zensiert werden. Andere wurden etabliert, um eine Alternative zum Modell des Überwachungskapitalismus von »Big Tech« zu bieten, und streben danach, ein soziales Medienuniversum frei von Unternehmensinteressen zu schaffen – das sogenannte Fediversum. Sowohl Web3-Dienste wie Odysee als auch das Fediversum werden oft als »dezentralisierte« Alternativen zu den dominierenden großen Online-Plattformen beschrieben. Beide unterscheiden sich jedoch in Bezug auf die Technologie und die weiter gefasste Vision, wobei Schnittstellen zwischen ihren Informationsräumen entstehen.[5] Im folgenden Abschnitt wird die potenzielle Ausnutzung des dezentralisierten Social Web im Zusammenhang mit Desinformation, Hass und Extremismus untersucht, bevor die entsprechenden politischen Implikationen bewertet werden.

Infobox: Web3

Odysee ist Teil der Web3-Bewegung, die der Journalist Gilead Edelman als »ein dezentralisiertes Online-Ökosystem auf Grundlage der Blockchain« beschreibt. Er fährt fort, dass »Plattformen und Anwendungen, die auf Web3 aufgebaut sind, nicht im Besitz eines zentralen Informationsregulators sein werden, sondern im Besitz der Nutzer:innen. Sie verdienen ihren Anteil, indem sie an der Entwicklung und Pflege dieser Dienste mitwirken«.[6] Das Herzstück von Web3 ist die Blockchain-Technologie, mit der die Nutzer:innen sich und ihre Inhalte authentifizieren können, sowie die Kryptowährung, mit der die Web3-Teilnehmer:innen (Infrastrukturanbieter:innen und regelmäßige Nutzer:innen) finanzielle Belohnungen erhalten können. Der besondere Stellenwert von Finanztransaktionen, die frei von staatlichen Eingriffen sind, hat das Web3 bei libertären Gruppen beliebt gemacht. Odysee ist eine Video-Hosting-Plattform, die von ihrem Entwickler LBRY als YouTube-Alternative vermarktet wird. Sie ermöglicht Streaming und Dateidownloads und basiert auf dem LBRY-Protokoll, einem dezentralisierten Filesharing-Netzwerk, das Blockchain- und BitTorrent-Technologien integriert und LBRY Credit (LBC) als Währung verwendet. Infolge eines Gerichtsurteils im US-Bundesstaat New Hampshire musste LBRY im Juli 2023 seine Schließung ankündigen, wodurch die Zukunft von Odysee und vergleichbaren Web3-Diensten in Frage gestellt wurde.[7]

Infobox: Das Fediversum

Das Fediversum verfolgt das gleiche Ziel eines dezentralisierten Social Web auf der Grundlage von Peer-to-Peer-Netzwerken, wenn auch ohne den Schwerpunkt auf finanziellen Belohnungen oder der Blockchain-Technologie. Stattdessen entwickelte sich das Fediversum aus der sogenannten »Bewegung für freie Software« heraus. Wie die Befürworter von Open-Source-Software erlaubt freie Software jedem, Software frei zu nutzen, zu verändern und zu verbreiten. Die Befürworter freier Software betonen jedoch auch ethische Grundsätze wie den Verzicht auf herstellergebundene Software, die die Zusammenarbeit behindert. Im Fediversum ermöglichen Protokolle wie ActivityPub, dass verschiedene Plattformen (auch bekannt als Instanzen) miteinander sprechen können, wodurch die Interoperabilität über den Wettbewerb gestellt wird. Das bedeutet, dass die Nutzer:innen sowohl untereinander als auch mit ihren Inhalten interagieren können, selbst wenn sie sich nicht auf der gleichen Instanz befinden. Auf diese Weise können Nutzer:innen von Mastodon (einem X-ähnlichen Microblogging-Dienst im Fediversum) ein Video auf PeerTube kommentieren (hierbei handelt es sich um eine Fediversum-Software, die die Erstellung von YouTube-ähnlichen Videodiensten ermöglicht), ohne ein neues Konto erstellen zu müssen. Das Fediversum wird bereits von den offiziellen Stellen der Europäischen Union[8] und der deutschen Bundesregierung[9] verwendet und von den jeweiligen Datenschutzbehörden unterstützt.

Die Bedrohungslage

Während sich die Forschung bislang intensiv mit den Risiken von großen Online-Plattformen und deren Nutzung befasst hat, wurde dem dezentralisierten Social Web weniger Aufmerksamkeit zuteil. Nach Angaben von Tech Against Terrorism ist die Ausnutzung dezentralisierter Dienste durch terroristische und gewalttätige Extremisten nach wie vor in erster Linie experimentell, wobei diese Dienste neben oder als Back-up für konventionelle, zentralisierte Plattformen und Dienste genutzt werden.[10] Mit dem Anstieg der Nutzerzahlen ist jedoch auch das Interesse an den potenziellen Schadensrisiken von Web3-Diensten und dem Fediversum gewachsen. Dies spiegelt sich in neuen Initiativen für den Zugriff auf und die Untersuchung von alternativen Informationsräumen. So kündigten Neuseeland, die USA, X und Microsoft im September 2022 eine Investition in eine technologische Innovationsinitiative in Zusammenarbeit mit OpenMined unter dem Banner des »Christchurch Call for Action« an.[11] Ziel der Initiative ist es, die unabhängige Untersuchung der Auswirkungen von Algorithmen und ihrer Interaktionen mit Nutzer:innen zu unterstützen, auch über mehrere Plattformen und Plattformtypen hinweg.

Plattformmigration und Echokammern: Sowohl Web3-Dienste wie Odysee als auch Fediversum-Dienste wie PeerTube haben in den letzten Jahren eine steigende Anzahl an Nutzer:innen verzeichnet. Dies gilt insbesondere für rechtsextreme und verschwörungsideologische Milieus, für die diese Dienste ein günstiges Umfeld zur Verbreitung schädlicher Inhalte, einschließlich Desinformation und Hass, sowie für schädliche Verhaltensweisen bieten können.[12] So haben Akteure und Bewegungen, die von YouTube ausgeschlossen wurden, wie z. B. die »Querdenken«-Bewegung oder bekannte rechtsextreme Hetzer:innen, im dezentralisierten Social Web schnell eine neue Heimat gefunden. Durch die Nutzung von PeerTube können Rechtsextreme beispielsweise Instanzen kreieren, die nur sie allein kontrollieren können. Die dort verbreiteten Inhalte können dann nur noch entfernt werden, indem die Server vom Netz genommen werden. Um die ideologische Vorherrschaft zu erlangen, werden Nutzer:innen, die sich nicht an ihre Ideale halten, nachweislich schikaniert.[13] Im Zuge dieser sogenannten »Community-Kaperunge« werden diese Informationsräume zunehmend zu ideologisch einheitlichen »Echokammern«. Wie im Folgenden dargelegt wird, erschwert die Trennung dieser alternativen Informationsräume ihre Regulierung und ermöglicht es böswilligen Akteuren, ungehindert und unangefochten Ideen auszutauschen.

Instanzübergreifende Freigabe schädlicher Inhalte: Nutzer:innen von Fediversum-Instanzen können Inhalte mit Nutzer:innen anderer Instanzen teilen. Anders als bei sehr großen Online-Plattformen wie Facebook oder X gibt es im Fediversum jedoch keine zentrale Kontrollinstanz, die den Informationsfluss zwischen den Instanzen moderieren oder unerwünschte Inhalte löschen kann.[14] Dies ist der Hauptvorteil und das Hauptmerkmal des Fediversums, also dessen dezentralisiertes Modell – denn »da es keine zentralisierte Fediversum-Instanz gibt, besteht auch keine Möglichkeit, selbst die schädlichsten Inhalte vollständig aus dem Netzwerk auszuschließen«.[15] Wie in einem Kommentar erklärt wird, ist das zuverlässige Löschen von Inhalten aus einem dezentralisierten Netzwerk »einfach nicht möglich«.[16] Dieses Risiko besteht zwar auch bei sehr großen Online-Plattformen, da andere Nutzer:innen den Beitrag vor dem Löschen bereits kopiert haben könnten. Mastodon legt aber beispielsweise für alle Nutzer:innen, die den Beitrag ansehen, einzelne Kopien an. Auf solchen Fediversum-Plattformen ist das »Recht auf Vergessenwerden« gemäß Artikel 17 der EU-Datenschutz-Grundverordnung (DSGVO) praktisch nicht durchsetzbar.

Monetarisierung schädlicher Inhalte: Odysee ermöglicht es Nutzer:innen, ihre Inhalte und die Nutzung der Plattform durch verschiedene Belohnungs-, Bonus- und Boost-Funktionen zu monetarisieren. So können Nutzer:innen beispielsweise LBRY Credits (LBCs) verdienen, indem sie einen Kanal erstellen, Videos hochladen, anderen Profilen folgen oder Follower gewinnen. Der Verdienst hängt vom Erreichen verschiedener Stufen ab (z. B. des sogenannten Status »Master of Views«). Die Möglichkeit zum Aufstieg auf bestimmte Level und die Gestaltung der Plattform weisen Elemente einer Gamifizierungsstrategie auf, wonach die Motivation der Nutzer:innen zur Nutzung des Dienstes durch spielähnliche Mechanismen gesteigert werden kann.[17] Dies ist besonders für jene Akteure attraktiv, die von größeren Online-Plattformen wie YouTube entmonetarisiert wurden. Eine ISD-Studie, in der 53 deutschsprachige Odysee-Nutzer:innen aus dem rechtsextremen und verschwörungsideologischen Milieu herangezogen wurden, ergab, dass die Konten seit der Einrichtung der Wallets insgesamt 1.652.786,96 LBRY-Credits erhalten

hatten (dies entspricht 122.306 USD gemäß dem durchschnittlichen Schlusskurs für LBC zwischen Januar 2022 und Mai 2023).[18] Die Volatilität von Kryptowährungen könnte zwar dazu geführt haben, dass die tatsächlichen Einnahmen seit dem Zeitpunkt der Analyse stark zurückgegangen sind, die Zahlen zeigen jedoch, dass Plattformen mit Anreizen wie Odysee zusätzliche Einnahmen für böswillige Akteure schaffen können. Es wurde außerdem festgestellt, dass profitable rechtsextreme Profile Videos verbreiteten, die Verschwörungsideologien wie «The Great Reset« und »QAnon« propagierten. Die untersuchten Videos enthielten auch antisemitische Aussagen, Geschichtsrevisionismus und die Leugnung des Klimawandels. In der untersuchten Videostichprobe erhielten rechtsextreme Kernthemen wie die Leugnung des Holocaust weniger Unterstützung als Videos, in denen aktuelle politische Themen diskutiert wurden.

Politische Implikationen

Die Regulierung alternativer Informationsräume beschäftigt die politischen Entscheidungsträger:innen und Regulierungsbehörden schon seit einiger Zeit. Dies ist auf das zuvor erwähnte Phänomen der Plattformmigration und der Echokammern zurückzuführen, die für das Verständnis der Entwicklung und der Funktionsweise extremistischer Bewegungen im Internet von großer Bedeutung sind. Das dezentralisierte Social Web kann zwar dazu beitragen, dass bestimmte politische Ziele wie »Open Source First«, die Einhaltung des Datenschutzes oder Interoperabilität erreicht werden. Es stellt jedoch zugleich auch einen weitgehend unkontrollierten Informationsraum dar, der von böswilligen Akteuren ausgenutzt werden kann. Obwohl Regeln zuweilen bereits vorliegen, ist die Durchsetzung nach wie vor mit erheblichen Problemen verbunden, da einige Anbieter:innen von Web3- und Fediversum-Diensten bewusst versuchen, sich der rechtlichen Kontrolle zu entziehen.

Finanzierung von Hass: Während auch einige etablierte große Online-Plattformen die Aktivitäten von Nutzer:innen belohnen (z. B. die Super-Chat-Funktion auf YouTube), stellen gamifizierte Erlebnisse auf Web3-Plattformen wie Odysee ein ganz neues Ausmaß an Schadensrisiken im Zusammenhang mit der Rentabilität von Desinformation und Hass durch spielerische Erfahrungen dar.[19] Die Technologie und die finanzielle Anreizstruktur dieser Plattformtypen müssen bei der Ausarbeitung neuer Maßnahmen oder der Aktualisierung bestehender Regelwerke berücksichtigt werden. Zu letzteren gehören nicht nur die Rahmenregelungen zur Regulierung von Plattformen und Technologien, sondern auch Initiativen zur Bekämpfung von Extremismus. Gleichzeitig können die bestehenden Regeln bereits Möglichkeiten bieten, der Monetarisierung von Inhalten und Verhaltensweisen durch böswillige Akteure in Web3-Diensten entgegenzuwirken. Diese Bestimmungen müssten jedoch wirksam durchgesetzt werden, wie das Verfahren der US-Börsenaufsichtsbehörde (*Securities and Exchange Commission, SEC*) gegen LBRY zeigt.[20]

Technologische Bluffs: Die Bezeichnung »dezentralisiert« wird häufig als Marketinginstrument verwendet und kann als Vorwand dienen, um die Einhaltung von einschlägigen Regelwerken zu umgehen. Die Video-Hosting-Plattform Odysee behauptet zwar, dezentralisiert zu sein, kann aber Inhalte moderieren und macht dies auch.[21] Obwohl Odysee Blockchain-Technologie nutzt, kann die Plattform weiterhin Kanäle aus Wiedergabelisten entfernen oder Inhalte geografisch sperren und damit die Inhalte für die meisten Nutzer:innen praktisch unzugänglich machen. Zwar können technisch versierte Nutzer:innen in diesem Fall weiterhin über die Blockchain auf die jeweiligen Inhalte zugreifen, diese können jedoch nicht mehr auf Odysee angesehen werden. Das bedeutet, dass die Plattform grundsätzlich in der Lage ist, Regeln wie der Pflicht zur Einrichtung von Melde- und Abhilfeverfahren nachzukommen.

Bestimmung der Dienstleistungstypen: Viele PeerTube-Instanzen erlauben das Teilen von nutzergenerierten Inhalten. Einige Instanzen bieten aber auch redaktionelle Formate an. Andere erlauben wiederum nur bestimmten Personen oder Medienorganisationen, Inhalte hochzuladen. Diese unterschiedlichen Nutzungsarten bedeuten, dass einige Instanzen rechtlich als soziale Netzwerke oder Online-Plattformen angesehen werden können, während andere als Herausgeber von redaktionellen Inhalten oder Medienplattformen gelten. Diese verschiedenen Kategorien können in verschiedenen Rechtsordnungen unterschiedliche rechtliche Verpflichtungen tragen. Darüber hinaus schreiben einige rechtliche Rahmenbedingungen vor, dass die Dienste Gewinne erzielen müssen (z. B. das deutsche Netzwerkdurchsetzungsgesetz). In den meisten Fällen ist bei PeerTube jedoch nicht klar ersichtlich, welche Art von Geschäftsbeziehung zwischen den Betreiber:innen, den Urheber:innen und den Nutzer:innen besteht. Außer-

dem ist die Unternehmensstruktur hinter den Instanzen oft unklar. Klarheit gibt es hingegen bei der Finanzierung, die in der Regel über Spenden und häufig auch in Form von Kryptowährungen erfolgt.[22]

Verbesserung der Regeldurchsetzung: Da es sich beim Fediversum um einen dezentralisierten Dienst handelt, gibt es dementsprechend auch keine zentrale Plattformbehörde, an die man sich wenden kann, um Inhalte zu entfernen. Stattdessen moderieren Serveradministrator:innen die Inhalte lokal. Einige Instanzen ermöglichen es Nutzer:innen, potenziell illegale Inhalte von verlinkten Instanzen anzusehen, auch wenn die Instanz selbst keine solchen Inhalte beherbergt. Dies wirft die Frage nach der Durchsetzung von Melde- und Abhilfeverfahren für illegale Inhalte in dezentralisierten Netzwerken auf. Darüber hinaus werden einige Instanzen von Einzelpersonen oder Organisationen in Ländern ohne nennenswerte Regelwerke für den digitalen Informationsraum betrieben, wobei auch Scheinfirmen eingesetzt werden. Folglich sollten die politischen Entscheidungsträger:innen die Einführung von Regeln in Erwägung ziehen, die die Anbieter:innen verpflichten, inländische gesetzliche Vertreter:innen für die Zustellung von rechtlichen Dokumenten einzusetzen. Im Jahr 2021 hatte die Bundesregierung jedoch Probleme, Telegram Mitteilungen zuzustellen – trotz einer entsprechenden Verpflichtung im NetzDG. Nach der erfolglosen Zustellung (Telegram, das offiziell seinen Sitz in Dubai hat, reagierte nicht) konnten die Mitteilungen schließlich durch Veröffentlichung im Bundesanzeiger zugestellt werden. Zusammen mit dem Druck, der über die Medien und die internationale Koordination ausgeübt wurde, reagierte Telegram schließlich und lege Berufung ein. In der Folge ist seit 2023 ein Gerichtsverfahren im Gange, das zum Zeitpunkt der Veröffentlichung dieses Berichts noch nicht abgeschlossen war.[23]

Aufbau auf bestehenden Ansätzen: Framasoft, die französische gemeinnützige Organisation, die PeerTube entwickelt, entfernt aktiv problematische Instanzen aus ihrer Suchfunktion, wenn sie nach französischem Recht illegale Inhalte enthalten.[24] Eine solche Verschärfung der Indizierungsstandards könnte ein angemessener Ansatz sein, da PeerTube-Instanzen nur dann ein breiteres Publikum finden können, wenn sie in durchsuchbare Indizes aufgenommen werden. Durch eine fehlende Indizierung kann der Zugang zu extremistischen Inhalten von problematischen Instanzen eingeschränkt werden. Darüber hinaus geben von der Gemeinschaft geführte Sperrlisten den Server-Hosts eine Möglichkeit, die Verbindung mit problematischen Instanzen zu vermeiden und somit die unbeabsichtigte Verbreitung von extremistischen Inhalten zu verhindern. Auf diese Weise können extremistische Instanzen zwar weiterhin direkt zugänglich sein, ihre Inhalte sind jedoch nicht mehr im weiteren Netzwerk sichtbar.

Förderung des Kapazitätenaufbaus: Politische Entscheidungsträger:innen und Regulierungsbehörden könnten die Entwicklung von Open-Source-Plugins unterstützen, die Betreibende von Fediversum-Instanzen verwenden, um die Meldung von Inhalten und die Moderationsprozesse entsprechend den jeweiligen Regelwerken wie beispielsweise dem DSA oder dem australischen Online Safety Act anzupassen. Dies wird das Problem der Inhaltsmoderation zwar nicht lösen, insbesondere nicht für Betreiber:innen, die über kein zusätzliches Personal verfügen, doch es würde immerhin die Bereitschaft beider Seiten zeigen, Transparenzanforderungen so weit wie möglich nachzukommen und einen Dialog zwischen der Fediversum-Gemeinschaft und den Regulierungsbehörden in die Wege zu leiten. Politische Entscheidungsträger:innen und Regulierungsbehörden können auch die Zivilgesellschaft dazu ermutigen und unterstützen, den Betreiber:innen von Instanzen Schulungen zu bewährten Verfahren der Inhaltsmoderation und zur Entwicklung von Standards anzubieten.

Abschnitt 2: Große Sprachmodelle

In einem offenen Brief, der im März 2023 veröffentlicht wurde, forderte das Future of Life Institute (FLI) KI-Unternehmen auf, das Training von KI-Systemen, die leistungsfähiger als GPT-4.0 sind, wegen »tiefgreifender Risiken für die Gesellschaft und die Menschheit«[25] zu unterbrechen. Während der Brief von einigen Forscher:innen unterstützt wurde, kritisierten andere, dass er imaginären apokalyptischen Szenarien eine größere Wichtigkeit zuweist als den Schadensrisiken, die durch den bereits weit verbreiteten Einsatz von Generativer KI entstehen.[26] So bergen bereits die derzeit zugänglichen KI-Systeme, die auf großen Sprachmodellen (»Large Language Models«, LLMs) beruhen, und ihr Einsatz sowohl Chancen als auch negative potenzielle Auswirkungen für Individuen und die Gesellschaft. Im folgenden Abschnitt werden ausgewählte Schadensrisiken, die diesen Modellen innewohnen, und das Potenzial ihrer Ausnutzung durch

böswillige Akteure mit Schwerpunkt auf Desinformation, Hass und Extremismus näher beleuchtet. Anschließend werden einige damit zusammenhängende politische Implikationen erörtert.

Infobox: Generative KI

Während künstliche Intelligenz (KI) ein Oberbegriff für Systeme und Technologien ist, die menschliche Intelligenz nachahmen, bezieht sich Generative KI auf KI-Anwendungen, die neuen Code, Text, Bilder, Audio, Video und multimodale Simulationen als Reaktion auf Aufforderungen generieren können. Der Begriff bezieht sich ebensoauf die zugrunden liegenden Sprachmodelle (»Language Models«, LMs). Auf diesen können die KI-Anwendungen aufgebaut werden. So generiert ChatGPT von OpenAI beispielsweise Text. DALL-E generiert Bilder und Kunst. ChatGPT und DALL-E basieren auf GPT-4.0, 3.5 bzw. 3.0 und DALL-E (eine Version von GPT-3.0), alle sind Beispiele für große Sprachmodelle (LLMs). Sprachmodelle sind Architekturen neuronaler Netzwerke – einer Reihe von Algorithmen, die, grob gesagt, die Operationen eines Gehirns nachahmen und komplexe Muster erkennen können. Diese neuronalen Netzwerke bestehen aus Knotenschichten, »die eine Eingabeebene, eine oder mehrere versteckte Ebenen und eine Ausgabeebene enthalten«.[27] Neuronale Netzwerke sind besonders nützlich für die Bündelung und Klassifizierung von Informationen. Je mehr Knotenebenen vorhanden sind, desto besser ist das neuronale Netzwerk in der Lage, sehr große und komplizierte Datensätze zu verarbeiten und Muster in unbeschrifteten und unmarkierten Daten zu erkennen. Viele dieser LLMs können in unzähligen nachgelagerten KI-Anwendungen wiederverwendet werden.

Die Bedrohungslage

Die vollständigen Auswirkungen von Generativer KI auf Einzelpersonen und die Gesellschaft sind noch nicht absehbar. Dennoch machen sich Forscher:innen[28], politische Entscheidungsträger:innen[29], Strafverfolgungsbehörden[30] und die Industrie[31] zunehmend Gedanken über die möglichen Schäden, die von diesen Systemen ausgehen können. Weidinger et al. beispielsweise leiteten 21 Risiken aus LMs in sechs potenziellen Risikobereichen ab, darunter »Diskriminierung, Ausgrenzung und Toxizität«, »Schäden durch Fehlinformationen« und »böswillige Nutzung«.[32] Andere Expert:innen konzentrierten sich speziell auf Audio-, Text-, Bild- oder multimodale Inhalte, die durch KI erzeugt oder modifiziert werden können. So erstellte die Partnership on AI eine Liste der Risiken im Zusammenhang mit synthetischen Medien und den verantwortungsvollen Praktiken und erklärte, dass »mit der zunehmenden Zugänglichkeit und Ausgereiftheit der synthetischen Medientechnologie auch ihre potenziellen Auswirkungen zunehmen«[33]. Daher ist es wichtig, dass Abhilfemaßnahmen sowohl auf der Angebots- als auch auf der Nachfrageseite ansetzen, wobei noch weitere Forschungsarbeiten erforderlich sind, um die spezifischen Schadensrisiken der KI-gestützten Informationsmanipulation vollständig zu verstehen.[34] Die Anbieter sind stets bemüht, die Genauigkeit ihrer Anwendungen zu verbessern. Die sich entwickelnden Kapazitäten von LLMs bleiben jedoch ungewiss, da es derzeit keine zuverlässigen Techniken zur »Steuerung des Verhaltens von LLMs« gibt.[35]

Inhärente Vorurteile: Toxizität, einschließlich Hass, wurde sowohl in großen LLMs als auch in Webtext-Korpora als ein weit verbreitetes Problem identifiziert.[36] Diese Modelle können manchmal Äußerungen, die schädliche Inhalte wie Fehlinformationen oder Hass darstellen, hohe Wahrscheinlichkeiten zuweisen. Wenn beispielsweise die Trainingsdaten die Ansichten von Minderheiten nicht berücksichtigen, besteht die Gefahr, dass LLM-Verteilungen die Ansichten und Werte der Mehrheit gegenüber denen der Minderheit verstärken. Studien zeigten, dass LLMs unerwünschte Stereotypen aufweisen, wie beispielsweise die anhaltende Assoziation von Muslim:innen mit Gewalt.[37] In einer neueren groß angelegten Analyse von einer halben Million generierter Ergebnisse aus GPT-3.5 legten die Forschungsergebnisse ebenfalls nahe, dass LLMs signifikant toxisch sein können, wenn etwa Rollen wie »eine schlechte Person« in den Einstellungen fürs Konversationsverhalten zugewiesen werden.[38] Im Zusammenhang mit Fehlinformationen können selbst fortgeschrittene LLMs nicht zuverlässig nachweislich wahre Informationen generieren – diese Modelle liefern unter bestimmten Umständen detaillierte und korrekte Informationen, unter anderen jedoch falsche.[39] Dies zeigt nicht nur, dass LLM-Ausgaben (je nach Kontext) inhärent toxisch oder irreführend sein können, sondern auch, dass scheinbar harmlose Aufforderungen durch alltägliche Nutzer:innen Inhalte dieser Art generieren können. In diesem Zusammenhang sollten insbesondere Modelle, die auf den Korpora von Plattformen basieren, deren

Inhalte bereits als toxisch oder irreführend identifiziert wurden, kritisch bewertet werden.

Vorsätzliche böswillige Verwendung: LLMs sind anfällig für die absichtliche böswillige Verwendung zur Erzeugung schädlicher Inhalte oder zur Durchführung schädlicher Verhaltensweisen. In einem Experiment wurde GPT-3.5 angewiesen, auf eine Reihe von Aufforderungen zu reagieren, die sich auf 100 nachweislich falsche Narrative bezogen.[40] Der Chatbot generierte 80 der 100 falschen Narrative. Der Anbieter versprach daraufhin Verbesserungen. Es wurde jedoch festgestellt, dass GPT-4.0 noch häufiger und überzeugender als das GPT-3.5 nachweislich falsche Narrative aufstellte.[41] Abgesehen von der potenziellen Ausnutzung von KI-generierten Texten durch böswillige Akteure können auch überzeugend realistische KI-generierte Deepfakes die Wirkung der Informationsmanipulation verstärken. Dadurch können die bereits bekannten Gefahren für die Qualität des öffentlichen Diskurses und digitalen Informationsraums verschärft werden.[42] Bild- und Videoinhalte werden nach wie vor häufig bei der ausländischen Informationsmanipulation und Einflussnahme eingesetzt.[43] Das liegt daran, dass solche Inhalte nicht nur ansprechend, sondern auch billig und einfach produziert werden können. Folglich werden böswillige Akteure auch ein besonderes Augenmerk auf den Kostenunterschied zwischen von Menschen erstellten und KI-generierten Inhalten legen.[44]

Generierung von Code: LLMs können zum Analysieren, Fehlererkennung aber auch Generieren von Computer-Code eingesetzt werden. Folglich gibt es Spekulationen über die Gefahr, dass böswillige Akteure LLMs ausnutzen, um Computer-Code für Cyberkriminalität, den Betrieb von Social Bots oder politisch motivierte Cyberangriffe effizienter zu generieren. Politisch motivierte Cyberangriffe sind natürlich keine neue Bedrohung: Im November 2022 wurde die Website des Europäischen Parlaments »nur wenige Momente« nach einer Abstimmung, in der Russland zum Sponsor von Terrorismus erklärt wurde, Ziel eines ausgeklügelten Cyberangriffs.[45] Für böswillige Akteure mit begrenzten Ressourcen und beschränktem technischen Know-how existieren nur begrenzte Möglichkeiten, die sich entwickelnden Fähigkeiten von LLMs zu nutzen. Fortschrittlichere Akteure mit ausreichenden Ressourcen, wie beispielsweise ausländische Nachrichtendienste, könnten die Modelle jedoch in Kombination mit anderen Technologien nutzen, um etwa ihre Aktivitäten der Informationsmanipulation zu verbessern, zu automatisieren und auszuweiten.[46] Gleichzeitig könnten LLMs aber auch zur Erhöhung der Cybersicherheit eingesetzt werden, indem sie etwa zur Prüfung und Fehlerbeseitigung von Code verwendet werden.[47]

Automatisierte Distribution: LLMs könnten in Kombination mit bekannten Techniken zur automatisierten Verbreitung von Inhalten ausgenutzt werden. So wurden Social Bots bereits in Informationsmanipulationskampagnen integriert und beispielsweise zur Verbreitung von Verschwörungsideologien eingesetzt.[48] Im März 2023 veröffentlichten Investigativjournalist:innen durchgesickerte E-Mails und andere Dokumente (die sogenannten »Vulkan Files«), die die Entwicklung von Sabotagesoftware durch das russische Unternehmen NTC Vulkan belegen. Einige der im Auftrag der russischen Regierung entwickelten Programme waren bereits in der Lage, Inhalte automatisch zu generieren und zu verbreiten.[49] Die Nutzung von LLMs, um »Sprachmuster nachzubilden« und sich überzeugend als Zielpersonen und -gruppen auszugeben[50], könnte daher solche Programme noch ausgefeilter machen. Folglich wird die Kombination von Techniken für die automatische Verbreitung mit LLMs die Tür für maßgeschneiderte und personalisierte Informationsmanipulation, einschließlich Desinformation, weiter öffnen.[51] Darüber hinaus könnten mit LLMs die derzeit bestehenden Techniken zur Aufdeckung von Informationsmanipulationen untergraben werden, indem sie die Abhängigkeit von kopiertem und eingefügtem Text verringern oder ganz aufheben.[52]

Verbesserte Überzeugungskraft: Fehlinformationen und Informationsmanipulation gab es schon immer, vor allem in der Politik. Ihre Wirkung auf Einzelpersonen und die Gesellschaft hängt jedoch letztlich von ihrer Fähigkeit ab, die Zuhörer:innen oder Leser:innen von einer bestimmten Botschaft zu überzeugen. In einem Experiment im Spieldesign wurde aufgezeigt, dass LLMs anspruchsvolle Gespräche und Dialoge mit Menschen auf höchst überzeugende Weise führen können.[53] In einer anderen experimentellen Studie wurde festgestellt, dass die von GPT-3.0 generierten Botschaften sogar bei verschiedenen politischen Themen, wie etwa ein Verbot von Angriffswaffen oder die Einführung einer Kohlenstoffsteuer, überzeugend argumentierten.[54] Eine Studie vom Juni 2023 mit 697 Teilnehmer:innen bestätigte diese Ergebnisse.[55] GPT-3.0 konnte nicht nur korrekte Tweets erstellen, die leichter zu verstehen waren, sondern auch

überzeugendere synthetische Fehlinformationen generieren. Zudem zeigten die Autor:innen auf, dass die Befragten nicht in der Lage waren, zwischen Tweets, die von GPT-3.0 generiert wurden, und solchen, die von echten Nutzer:innen geschrieben wurden, zu unterscheiden.

Unmittelbarer Zugang: Die Überzeugungskraft von KI-generierten Inhalten kann auch durch die Tatsache verstärkt werden, dass der Zugang zu LLM-Ausgaben oft unvermittelt ist. Herkömmliche Suchmaschinen wie Google oder Bing liefern gemeinhin vergleichbare Ergebnisse. Dies ist besonders problematisch, wenn LLMs für die meisten Suchanfragen korrekte Informationen liefern, was dazu führen kann, dass die Nutzer:innen den Ergebnissen übermäßig vertrauen, obwohl es sich in Einzelfällen auch um Fehlinformationen handeln kann.[56] Angesichts der potenziellen Überzeugungskraft von LLM-Ausgaben ist es nicht verwunderlich, dass die Menschen mit Sorgen auf KI-generierte Fehlinformationen blicken.[57] Die Auswirkungen von LLMs auf das Vertrauen in die Medien müssen jedoch noch nachgewiesen und weiter erforscht werden.[58]

Politische Implikationen

Politische Entscheidungsträger:innen und Regulierungsbehörden, die LLMs regulieren wollen, sehen sich mit neuartigen und drängenden Fragen konfrontiert – und dabei handelt es sich nicht um die gleichen Fragen, die im Zusammenhang mit Architekturen neuronaler Netzwerke mit engen und beabsichtigten Anwendungsfällen oder anderen Arten von Software aufgekommen sind. Dies hängt zum Teil mit der Unvorhersehbarkeit von LLMs zusammen, die gefährliche und für böswillige Akteure zugängliche Fähigkeiten enthalten oder entwickeln können.[59] Während neue Bemühungen zur Regulierung Mitte 2023 unter Parlamentarier:innen in den USA an Fahrt gewinnen[60], haben sieben der einflussreichsten KI-Unternehmen, darunter Amazon, Google, Meta, Microsoft und OpenAI, im Juli 2023 mit dem Weißen Haus ab sofort geltende Selbstverpflichtungen zum Umgang mit den von KI ausgehenden Schadensrisiken vereinbart[61]. Diese Verpflichtungen unterstreichen zwar die amerikanischen Grundsätze der Sicherheit und des Vertrauens, gelten aber nur für KI-Systeme, die leistungsfähiger sind als GPT-4.0, Claude 2, PaLM 2, Titan und – im Falle von KI-generierten Bildern – DALL-E 2. Wie in diesem Bericht erörtert, birgt jedoch bereits die aktuelle Generation von KI-Anwendungen ernste Schadensrisiken, für die dringend Risikominderungsmaßnahmen erforderlich sind.

Im Juni 2023 hat das Europäische Parlament (EP) bereits seine Verhandlungsposition zum AI Act angenommen, einschließlich neuer Regeln für Basismodelle – ein Begriff, der parallel zu LLMs verwendet wird.[62] Im Allgemeinen wird dem derzeit vorgeschlagenen Gesetz mit großer Wahrscheinlichkeit ein »risikobasierter« Ansatz zugrunde liegen, bei dem Systeme entsprechend ihrem Potenzial, die öffentliche Sicherheit und die Grundrechte zu gefährden, eingestuft werden. Die vom EP vorgeschlagenen Regeln beinhalten zusätzliche Transparenzanforderungen für Basismodelle, wie beispielsweise die Offenlegung, dass Inhalte von ihnen generiert wurden. Auch eine Pflicht, dass das Modelldesign illegale Inhalte generiert, gehört dazu. Die endgültige Fassung des AI Acts kann sich jedoch nach August 2023 noch ändern und das Gesetz wird erst im Jahr 2025 in Kraft treten. Aus diesem Grund versucht die Europäische Kommission, KI-Entwickler:innen davon zu überzeugen, dem KI-Gesetz zuvorzukommen, indem sie einen freiwilligen Pakt schließen.[63] Gleichzeitig überlegen politische Entscheidungsträger:innen und Regulierungsbehörden auf der ganzen Welt, welche bereits bestehenden Regelwerke Anwendung finden könnten.

Feststellung der Verantwortlichkeit und Haftung: Die Feststellung der individuellen Verantwortung der Anbieter für schädliche Ergebnisse und Unzulänglichkeiten der KI ist kompliziert, da die Systeme undurchsichtig und unvorhersehbar sind sowie eine Vielzahl von Akteuren und Ressourcen involvieren.[64] Die Rechenschaftspflicht ist daher der »Eckpfeiler« der Steuerung von KI-Technologien.[65] Es besteht Unsicherheit darüber, ob Anbieter von LLMs für ihre Ergebnisse rechtlich haften sollten. Während Anbieter von Online-Plattformen in den meisten liberal-demokratischen Ländern nur begrenzt für nutzergenerierte Inhalte haften (siehe dazu zum Beispiel Section 230 US Communications Decency Act oder Art. 4 DSA), gilt dies möglicherweise nicht für Anbieter von LLMs, da LLMs mit großer Wahrscheinlichkeit (zumindest teilweise) als inhaltserstellende Dienste und nicht als Vermittlungsdienste eingestuft werden.[66] Für die Kläger:innen ist es schwierig nachzuweisen, ob ein bestimmter Schaden aufgrund eines Fehlverhaltens des Anbieters eingetreten ist, da die Verfahren von fortgeschrittenen LLMs im Detail nicht nachvollzogen werden können (beispielsweise die Frage, wie bestimmte Ergebnisse erzeugt wurden). Daher sollten die politischen Entscheidungsträ-

ger:innen die Einführung von Regeln in Erwägung ziehen, die von den Anbieter:innen mehr Modelltransparenz verlangen (beispielsweise durch regelmäßige Transparenzberichte, Modell- oder Systemkarten[67]) und alternative Wege finden, um das Fehlverhalten von Anbieter:innen in bestimmten Fällen zu definieren (zum Beispiel bei der Verletzung etwaiger Sorgfaltspflichten).

Beschränkung des Zugriffs: Der Kostenunterschied zwischen von Menschen erstellten und von KI generierten Inhalten ist besonders für böswillige Akteure von großer Bedeutung. Ein zentraler Faktor ist dabei das Ausmaß, in dem böswillige Akteure LLMs über Entwicklungs-Programmierschnittstellen (APIs) in ihre eigenen Anwendungen integrieren können. Die politischen Entscheidungsträger:innen und Regulierungsbehörden könnten die Anbieter etwa dazu verpflichten, strenge Prüfkriterien für APIs einzuführen und Hintergrundprüfungen hinsichtlich der möglichen Nutzung ihrer Dienste durch böswillige Akteure durchzuführen. Jedoch ist nach wie vor unklar, ob Prüfverfahren den Zugriff böswilliger Akteure auf Entwickler-APIs verhindern können, und wenn ja, ob diese dann einfach zu Open-Source-Modellen wechseln würden.[68] Zudem haben auch Regierungen oder andere Akteure mit reichlich Ressourcen die Möglichkeit, ihre eigenen Systeme zu trainieren und zu betreiben.[69] Dies würde es diesen Akteuren ermöglichen, mögliche Leitplanken, die etablierte LLMs für die Filterung schädlicher Inhalte besitzen, gänzlich zu umgehen.

Inklusives Design: Wie bereits erläutert, können voreingenommene Trainingskorpora unter Umständen zur Erstellung von toxischen Inhalten, einschließlich Hass und Fehlinformationen, führen. Die Anbieter von LLMs sollten daher verpflichtet werden, angemessene und wirksame Maßnahmen zu ergreifen, um der inhärenten Toxizität und Fehlinformationen in LLMs entgegenzuwirken. Dies könnte beispielsweise eine Änderung der Trainingsdaten, das Trainieren eines Klassifikators und dessen Verwendung zur Verringerung der Wahrscheinlichkeit schädlicher Inhalte oder die Anpassung von Belohnungsmodellen im Zusammenhang mit schädlichen Inhalten umfassen. Anbieter von LLMs und darauf aufbauende KI-Anwendungen könnten ferner verpflichtet werden, öffentlich zugängliche Berichte zu veröffentlichen, die Stresstests auf Toxizität und Fehlinformationen enthalten, um Nutzer:innen darüber in Kenntnis zu setzen.[70]

Safety-by-Design: »Safety by Design« beruht auf drei Grundprinzipien: Verantwortlichkeit der Diensteanbieter, Befähigung und Autonomie der Nutzer:innen, und Transparenz und Rechenschaftspflicht. Politische Entscheidungsträger:innen und Regulierungsbehörden sollten die Anbieter von LLMs und nachgelagerten KI-Anwendungen verpflichten, diese Grundsätze einzuhalten, indem sie sicherstellen, dass sie in jeder Phase des Produktlebenszyklus Sicherheitsmaßnahmen einbauen. Dazu müssen sie Interessengruppen aus verschiedenen Sektoren konsultieren und mit der Nutzergemeinschaft zusammenarbeiten, einschließlich derjenigen, die normalerweise unterrepräsentiert sind oder für die ein größeres Risiko besteht, Opfer im digitalen Informationsraum zu werden.

Schutz der Grundrechte: KI-generierte Deepfakes können die Wirksamkeit der Informationsmanipulation erhöhen. KI-generierte Inhalte können jedoch unter bestimmten Umständen direkt gegen Grundrechte verstoßen. Im Jahr 2008 verklagte zum Beispiel ein Fußballspieler Electronic Arts erfolgreich vor einem Gericht in Hamburg, um zu verhindern, dass das Videospiel FIFA sein Konterfei ohne seine Zustimmung verwendet.[71] Entscheidend ist hier, dass der Schaden aus dem »Recht des Klägers, über die Verwendung seines Namens zu entscheiden« und nicht aus kommerziellen Erwägungen resultiert.[72] Die Durchsetzung der Grundrechte kann daher eine weitere Option sein, um zu verhindern, dass böswillige Akteure LLMs ausnutzen. Dazu müssen jedoch zunächst Haftungsfragen geklärt werden.

Verbesserung der Detektion: Die Kennzeichnung von KI-generierten Inhalten, beispielsweise durch digitale Wasserzeichen[73], ist ein häufig diskutiertes Thema unter politischen Entscheidungsträger:innen und Regulierungsbehörden.[74] In der Praxis ist es jedoch eine schwierige Aufgabe, KI-generierte Inhalte zu erkennen.[75] Zum einen gestaltet sich der Aufbau von LLMs mit besser nachweisbaren Ausgaben aus technischer Sicht als schwierig (beispielsweise direkte Manipulation von LLM-Parametern zur Erstellung statistischer Fingerabdrücke oder Trainingsmodelle mit sogenannten »radioaktiven« Daten) und erfordert weitere Forschung und Koordination zwischen den Entwickler:innen. Zum anderen wirft die Verbreitung »radioaktiver« Daten direkt im Internet, wo sie mit großer Wahrscheinlichkeit zur Anlernung von LLMs abgegriffen würden, ethische Bedenken auf, da große Datenmengen verbreitet werden

müssten. Folglich ist eine weitere Abstimmung der Vorgehensweisen von LLM-Anbietern, Online-Plattformen, politischen Entscheidungsträger:innen, Regulierungsbehörden und Forscher:innen erforderlich, um die Erkennung von KI-generierten Inhalten zu verbessern. Die Coalition for Content Provenance and Authenticity (C2PA)[76] und die Partnership on AI[77] sind in dieser Hinsicht vielversprechende Initiativen.

Abschnitt 3: Extended Reality

Das Pew Research Center befragte zwischen Februar und März 2022 mehr als 600 Expert:innen nach deren Prognose über die Entwicklung und die Auswirkungen des Metaversums bis zum Jahr 2040. Ein beachtlicher Anteil vertrat die Ansicht, dass sich der Einzug von Extended Reality (XR) in den Alltag der Menschen mehr auf Augmented Reality (AR) und Mixed Reality (MR) als auf eine noch immersivere virtuelle Realität (VR) konzentrieren wird. Sie warnten, dass diese Technologien »jede menschliche Eigenschaft und Tendenz – sowohl eine schlechte als auch eine gute – enorm verstärken können«.[78] Toby Shulruff, leitender Spezialist für technologische Sicherheit beim US-amerikanischen National Network to End Domestic Violence, stellte beispielsweise fest, dass »XR wie andere Technologien keine menschlichen Probleme wie Vorurteile, Angst oder Gewalt löst«, sondern vielmehr »das, was in der Gesellschaft bereits vorhanden ist, beschleunigt und verstärkt«. Er warnte zudem vor der ernstzunehmenden Möglichkeit, dass sich Menschen, die entsprechende Dienste in Anspruch nehmen, immer mehr von der Welt um sie herum abkoppeln könnten.[79] Inwieweit der Alltag der Menschen von XR geprägt sein wird, ist zwar noch unklar, aber wir kennen bereits die zahlreichen Herausforderungen, die sich aus den derzeitigen digitalen Informationsräumen ergeben. Ganz wie bei Online-Plattformen werden auch XR-basierte Informationsräume wie das Metaversum mit Fragen der Haftung, der Strafverfolgung, der Inhaltemoderation, des Datenschutzes, der Transparenz und des Schutzes der Grundrechte ihrer Nutzer:innen konfrontiert werden. In diesem Abschnitt werden die Schadensrisiken des Metaversums und dessen Nutzung untersucht, bevor die politischen Implikationen erörtert werden.

Infobox: Das Metaversum

Zeitgleich mit der Umbenennung von Facebook in Meta im Oktober 2021 kündigte[80] CEO Mark Zuckerberg seine Absicht an, »den Nachfolger des mobilen Internets«[81] zu bauen und nannte es »Metaverse«, oder das »Metaversum«. Das Metaversum wird als »ein Zusammenfluss von physischer, erweiterter und virtueller Realität in einem gemeinsamen Online-Raum« beschrieben.[82] Wie andere XR-Welten soll die Erfindung damit über spielähnliche Ziele und Gamifizierung hinausgehen. Es gibt viele Beschreibungen und Analogien zum Metaversum, daher lohnt es sich, aufzuschlüsseln und zu definieren, was es eigentlich sein sollte. In einem Essay[83], der im Januar 2020 veröffentlicht wurde, skizzierte Matthew Ball sieben Kernmerkmale des Metaversums: (1) persistent – auf unbestimmte Zeit fortbestehend; (2) synchron und live – eine lebendige Erfahrung, die konsistent für jeden und in Echtzeit existiert; (3) ein individuelles Gefühl der »Präsenz« – das Gefühl, mit anderen Nutzer:innen (beispielsweise bei einer Veranstaltung) und Objekten so zu interagieren, als wäre man physisch mit ihnen verbunden; (4) eine umfassende Wirtschaft – Einzelpersonen und Unternehmen werden in der Lage sein, etwas zu schaffen, zu besitzen, zu investieren, zu verkaufen und für ihre Arbeit belohnt zu werden; (5) eine Verschmelzung virtueller und physischer Welten, die private und öffentliche Netzwerke/Erfahrungen sowie offene und geschlossene Plattformen umfasst; (6) vollständig interoperabel – Nutzer:innen können ihre Avatare und digitalen Gegenstände/Assets von einer Plattform im Metaversum auf eine andere mitnehmen; und (7) gefüllt mit Inhalten und Erfahrungen, die von einer Reihe von Mitwirkenden erstellt und betrieben werden. Im Dezember 2021 eröffnete Meta den Zugang zu dessen Multiplayer-VR-Plattform Horizon Worlds.[84] Das Metaversum, wie es oben beschrieben wurde, existiert jedoch (noch) nicht und hängt von mehreren Bedingungen ab, darunter die Verfügbarkeit von Hardware und damit von Zugangsmöglichkeiten (zum Beispiel Helme, Linsen, sensorische Anzüge oder sogar neuronale Verbindungen), eine starke Internetverbindung und technische Standards für Software, die Interoperabilität ermöglichen[85].

Die Bedrohungslage

Im Jahr 2022 veröffentlichten das Europol Innovation Lab[86] und der EU-Koordinator für die Terrorismusbekämpfung[87] jeweils Berichte über das Potenzial für die schädliche Nutzung, Kriminalität und Radikalisierung im Metaversum. Die Dokumente befassen sich mit den Modalitäten des Metaversums, einschließlich seines immersiven Charakters, der Erfassung von Emotionen und der Verwendung von Avataren. Zudem werden deren Auswirkungen auf die Verbreitung schädlicher und illegaler Inhalte, Belästigung und Missbrauch sowie Terrorismus, insbesondere bei der Rekrutierung, Finanzierung und beim Training analysiert. Im Februar 2023 veröffentlichte auch die WeProtect Global Alliance[88] eine Analyse, die einen Überblick über die neuesten XR-Trends und deren potenzielle Auswirkungen auf die sexuelle Ausbeutung und den Missbrauch von Kindern im Internet gibt. Im Mai 2023 veröffentlichte die gemeinnützige Organisation Standards Australia außerdem ein Whitepaper, in dem die wichtigsten Definitionen für das Metaversum, verschiedene Risiken, einschließlich »menschlicher Risiken« und »gesellschaftlicher Risiken«, sowie die bestehenden Standards dargelegt werden. Kurz darauf, im Juli 2023, befasste sich das Weltwirtschaftsforum (WEF) [89] mit Datenschutz- und Sicherheitsfragen im Zusammenhang mit dem Metaversum. Während einige Anbieter die Vorteile des Metaversums hervorheben (wir verwenden den Begriff im Folgenden als Synonym für XR-basierte Informationsräume, die sich auf soziale Verbindungen fokussieren, und nicht für das gleichnamige Produkt von Meta), birgt dieses gleichzeitig Schadensrisiken, die näher in Betracht gezogen werden sollten.

Sexuelle Belästigung, Missbrauch und geschlechtsspezifische Gewalt: Das Metaversum ist ein geschlechtsspezifischer Raum, der von der frauenfeindlichen und hypermaskulinen Kultur in Spielräumen beeinflusst wird.[90] Das Europol Innovation Lab entdeckte etwa einen Vorfall, in dem eine Frau beschrieb, wie sie innerhalb von 60 Sekunden nach ihrem Beitritt zu Metas Horizon Venues »virtuell gruppenvergewaltigt« wurde.[91] Spielumgebungen können, wie andere soziale Räume auch, Systeme struktureller Diskriminierung und Ungleichheiten wie Rassismus, Sexismus und Behindertenfeindlichkeit reproduzieren.[92] Dabei wird die geschlechtsspezifische digitale Gewalt auf »traditionellen« Online-Plattformen als eine Kontinuität der Offline-Gewalt verstanden, welche sich auf viele Arten gegen Frauen, Mädchen und marginalisierte Geschlechtsidentitäten richten kann, insbesondere gegen Betroffene mit überschneidenden Identitätsmerkmalen wie Abstammung, Sprache, Heimat und Herkunft, Glauben, religiöse oder politische Anschauung. Die Verkörperung und das Gefühl der Präsenz, die das Metaversum bietet, können das Gefühl der Belästigung verstärken und Verletzungen des persönlichen Raums und der körperlichen Präsenz ermöglichen. Dies kann besonders schädlich für Kinder sein, die von betrügerischen Avataren, die sich zunächst wie andere Kinder verhalten, sexuell belästigt und missbraucht werden könnten.[93] Ebenso erschwert die »Flüchtigkeit« des Metaversums die Meldung unerwünschter Verhaltensweisen.[94] Es ist daher wichtig, die Möglichkeit von sehr realen, einschneidenden Erfahrungen mit sexueller Belästigung und Missbrauch anzuerkennen.

Psychologische Auswirkungen: Das Erlebnis von Gewalt im Metaversum kann sich im realen Leben auf die psychische Gesundheit der Nutzer:innen auswirken, insbesondere bei jungen Nutzer:innen. Nach bisherigen Untersuchungen wird eine hohe oder mehrfache Gewalterfahrung in der frühen Adoleszenz mit einer emotionalen Desensibilisierung in Verbindung gebracht, die zu schwerer Gewalt im späten Jugendalter beitragen kann.[95] Die Desensibilisierung gegenüber Gewalt ist eine Form der Gewöhnung, eine Art des nicht-assoziativen Lernens, das nach wiederholter Exposition häufig zu einer verminderten Reaktion auf einen Reiz führen kann, und zwar kontext- und umgebungsübergreifend.[96] So kann beispielsweise das Miterleben einer Schlägerei zu einer Desensibilisierung gegenüber anderen Arten von Gewalt im selben Kontext sowie gegenüber Gewalt in anderen Umgebungen (beispielsweise zu Hause oder in der Schule) führen.[97] Wenn man beispielsweise Zeuge von Massenmorden an Avataren durch Terroristen im Metaversum wird, könnte dies zu einer ähnlichen Desensibilisierung und anderen psychologischen Schäden führen.

Ideologischer Wegbereiter: Das Metaversum könnte böswilligen Akteuren einen Raum bieten, um Propaganda zu verbreiten, Nutzer:innen zu rekrutieren und durch Veranstaltungen und regelmäßige Treffen Kontrolle über eine radikalisierte Gemeinschaft auszuüben. Erstens könnten die emotionale Investition und die unklare Unterscheidung zwischen der realen und der virtuellen

Welt die Nutzer:innen anfälliger für emotionale Manipulation machen.[98] Avatare könnten etwa dazu missbraucht werden, Emotionen zu wecken und extremistische Ideologien zu verbreiten, indem beispielsweise verstorbene Terroristenanführer »virtuell wiederauferstehen«.[99] Zweitens könnte der Einsatz von Hardware, die körperbasierte Daten wie Augenbewegungen oder andere Körperbewegungen erfasst, die Möglichkeiten zur Aufzeichnung biometrischer Daten beschleunigen, was wiederum neue Möglichkeiten zum Identitätsdiebstahl oder zur Auswahl und Beeinflussung von gefährdeten Nutzer:innen sowie zur Anpassung von Botschaften an deren Vorlieben schaffen würde. Dadurch werden böswillige Akteure in der Lage sein, ihre Propaganda wirksamer zu gestalten und Menschen zu rekrutieren. Drittens könnte das Metaversum dazu genutzt werden, um emotionale historische Ereignisse zu reproduzieren oder eine bestimmte Weltanschauung zu schaffen (beispielsweise ein virtuelles Kalifat oder einen Staat der weißen Vorherrschaft), um Anhänger zu gewinnen. Das Europol Innovation Lab stellte fest, dass sich derartige Räume zu einer Parallelwelt entwickeln könnten, die die Rechtsstaatlichkeit untergräbt.[100] So wurden beispielsweise bereits Nazi-Gaskammern in Roblox gemeldet, einer Plattform, auf der Menschen ihre eigenen Erlebnisse oder Minispiele erstellen und diese mit anderen Nutzer:innen teilen können.[101]

Finanzielle und operative Voraussetzungen: Auf der Grundlage von Blockchain-Technologie, Kryptowährungen und NFTs werden die in der Blockchain gespeicherten digitalen Eigentumsnachweise voraussichtlich eine wichtige Rolle in der Wirtschaft des Metaversums spielen.[102] Kryptowährungen könnten zur Geldwäsche missbraucht werden, was die Überwachung von Überweisungen – insbesondere länderübergreifend – erschwert. Finanzielle Mittel könnten durch den Verkauf von Artefakten wie zum Beispiel Hakenkreuzen als NFTs beschafft werden, die dann zur individuellen Gestaltung von Avataren oder zur Anzeige ihrer Zugehörigkeit zu terroristischen oder extremistischen Organisationen verwendet werden. Das Metaversum könnte auch als eine Umgebung zur Durchführung von Trainingseinheiten und zur Realisierung von Szenarien fungieren, beispielsweise zum Üben von Präzisionsschüssen, Geiselnahmen oder gar Aufklärungsarbeiten. Die emotionale und immersive Charakter des Metaversums würde diese Art von Training realistischer und fesselnder machen. Die Modellierung könnte terroristischen Organisationen ein Instrument an die Hand geben, mit dem sie Ziele aus der realen Welt nachbilden könnten, um einen Anschlag zu üben und dessen Wirkung zu maximieren.

Politische Implikationen

Das Metaversum wirft viele der gleichen politischen Debatten auf wie das Internet. Eigenschaften – insbesondere die Unmittelbarkeit und Flüchtigkeit – werden im Zusammenhang mit der Privatsphäre und der Sicherheit der Nutzer:innen berücksichtigt werden müssen. Darüber hinaus wird das Metaversum Risiken für Einzelpersonen und die Gesellschaft mit sich bringen, denen auf besondere Weise begegnet werden muss. In einem 2022 veröffentlichten Bericht des Analyse- und Forschungsteams des Rates der Europäischen Union wird ein potenzieller Kampf zwischen den jeweiligen Rollen von Regierung, Industrie und Nutzer:innen vorausgesagt, der zu verschiedenen Modellen führen könnte: ein regulatorischer Rahmen mit Schwerpunkt auf dem Schutz der Grundrechte; ein Ansatz, der sich auf ein freies, dezentralisiertes und offenes Internet mit Währungs- und Eigentumsrechten konzentriert; oder ein geschäfts- und gewinnorientiertes Modell, das von der Industrie unterstützt wird und das Eigentum am Metaversum beansprucht.[103] Die Herausforderung der unterschiedlichen Interessen verdeutlicht die Wichtigkeit einer frühzeitigen Entwicklung von gemeinsamen Ansätzen und Modellen.

Angemessenheit und Anwendbarkeit der Strafgesetze: Berichte über sexuelle Übergriffe im Metaversum werfen neue Fragen zu den rechtlichen Voraussetzungen für »herkömmliche« Straftaten auf. Eine Vergewaltigung beispielsweise erfordert einen physischen Akt, während ein Avatar virtuell ist. Es sollte jedoch argumentiert werden, dass die Immersivität des Metaversums zu einem sehr realen (emotionalen) Erlebnis von Vergewaltigung führen kann, was die Frage nach den Anforderungen für Straftatbestände im Zusammenhang mit Vergewaltigung erneut aufwirft. Es wird wichtig sein, die bestehenden Strafgesetze zu überprüfen und zu definieren, was ein kriminelles Verhalten im Metaversum darstellt. Genauso sollten neue Gesetze in Erwägung gezogen werden, um die Verfolgung dieser Straftaten zu ermöglichen. Eine rechtliche Überprüfung sollte sich auf mehrere Perspektiven aus der Zivilgesellschaft, der Industrie, der Strafverfolgung und der Wissenschaft stützen und sowohl die Angemessenheit

als auch die Anwendbarkeit der bestehenden Strafgesetze und -vorschriften bewerten, insbesondere im Hinblick auf psychologische und körperliche Schäden und die Abgrenzung zwischen physischen und virtuellen Schäden. Im Rahmen einer solchen Überprüfung könnten die Bedingungen entwickelt werden, unter denen ein Avatar mit einer Person gleichgesetzt werden kann, und es könnte erörtert werden, ob etwaige beobachtete Schäden durch neue Rechtsvorschriften kriminalisiert werden sollten. Politische Entscheidungsträger:innen, Regulierungsbehörden und Strafverfolgungsbehörden könnten eine Metaversum-spezifische Taxonomie im Zusammenhang mit Schäden und Straftaten entwickeln, um proaktiv neue Arten von Schäden und Verbrechen zu identifizieren.[104]

Ermittlung von Straftaten und Beweisaufnahme: Das Europol Innovation Lab betont, dass der flüchtige Charakter des Metaversums zu einem Mangel an virtuellen Spuren und zu Schwierigkeiten bei der Beweisaufnahme für kriminelle Vorfälle führen könnte.[105] Angesichts der flüchtigen Nutzererfahrung und der potenziell kurzen Reaktionszeiten nach der Meldung von Vorfällen wird sich die Datenerfassung im Zusammenhang mit Vorfällen im Metaversum schwierig gestalten. Ein weiteres Hindernis für kriminalpolizeiliche Ermittlungen ist die Schwierigkeit, eingeschränkte »Räumlichkeiten« im Metaversum zu betreten, die einen Schlüssel oder ein NFT erfordern. Da es zudem schwierig ist, den Standort der Nutzer:innen (oder des verwendeten Zugangsgeräts) zu ermitteln, werden die Strafverfolgungsbehörden vor die Herausforderung gestellt werden, die tatsächliche Identität der Nutzer:innen festzustellen und die Gerichtsbarkeit zu etablieren. Eine Koordinierung der Strafverfolgungsbehörden wird erforderlich sein, um die Möglichkeiten der Entschädigung und die Rechtsmittel in den verschiedenen Gerichtsbarkeiten zu prüfen. Diese Bemühungen sollten auch auf die Entwicklung einer datenschutzgerechten Vorgehensweise abzielen.[106]

Entwicklung von Moderationsverfahren: Das Europol Innovation Lab schätzt ein, dass »einfach nur in einem virtuellen Auto durch ein Metaversum zu patrouillieren, bei potenziell endlosen Welten wahrscheinlich nicht sehr gut funktionieren [wird], sowohl zur Abschreckung als auch um ansprechbar zu sein«.[107] Die bereits bekannten Herausforderungen bei der Inhaltemoderation, insbesondere im Hinblick auf die Abwägung zwischen Meinungsfreiheit und Persönlichkeitsrechten, werden sich im Metaversum noch verstärken, da die Anbieter ihren Schwerpunkt eher auf das Verhalten der Nutzer:innen als auf Inhalte legen werden müssen. Die Herausforderungen beziehen sich auf den Einsatz von menschlicher und KI-gestützter Moderation, einschließlich Fragen zu Ressourcen, Datenverzerrungen und algorithmischer Diskriminierung. Die Anbieter müssen Wege finden, um Interaktionen zu moderieren und gleichzeitig die Durchsetzung von Gemeinschaftsstandards in Echtzeit zu ermöglichen, damit die Privatsphäre der Nutzer:innen gewahrt bleibt. In privaten Räumen werden andere Formen der Moderation erforderlich sein. Sollte zum Beispiel die Rhetorik einer Person in ihrem virtuellen Zuhause der gleichen Moderation unterliegen wie auf dem öffentlichen Marktplatz eines Metaversums? Die Unterscheidung zwischen öffentlichen und privaten Online-Räumen (die für politische Entscheidungsträger:innen bereits eine schwierige Herausforderung darstellt)[108] erfordert ein klareres Verständnis darüber, welche Inhalte und Verhaltensweisen an welchen Stellen im Metaversum erlaubt sein sollten. Ebenso gibt es geografische Unterschiede in der Gesetzgebung, was eine Variation bei der Einstufung des Verhaltens und bei der Moderation von Inhalten erfordert – so könnten beispielsweise in einigen Ländern bestimmte Gesten oder der Konsum von Alkohol verboten sein.[109] Die Nutzer:innen sollten auch die Möglichkeit haben, Verhaltensweisen, die gegen die Gemeinschaftsstandards verstoßen, in Echtzeit zu melden.

Safety- and privacy-by-design: Die politischen Entscheidungsträger:innen und Regulierungsbehörden sollten die (co-)regulatorischen und freiwilligen Rahmenbedingungen für Online-Plattformen überprüfen, insbesondere im Hinblick auf die Bestimmungen, die die Anbieter verpflichten, die Risiken ihrer Dienste zu bewerten und zu mindern, um ihre Anwendbarkeit auf das Metaversum zu beurteilen. In der EU könnte dies den DSA und AI Act betreffen. Darüber hinaus sollten die Anbieter von XR-Diensten die Standardisierung von Sicherheitstools wie Stummschaltung, Sperrung und andere Sicherheitsressourcen im Zusammenhang mit dem Metaversum in Betracht ziehen. Meta weist zum Beispiel darauf hin, dass die Nutzer:innen in Horizon Worlds eine »sichere Zone« nutzen könnten, die als »ein persönlicher Raum, in dem man sich für einen Moment von anderen Menschen und der Umgebung zurückziehen kann«, beschrieben wird. Mithilfe des »Schild«-Symbols können die Nutzer:innen die Welt, in der sie sich

gerade befinden, melden, Nutzer:innen in ihrer Nähe ansehen, melden oder blockieren, jemanden stummschalten oder die Stummschaltung aufheben oder in ihren »persönlichen Bereich« gehen.[110] Meta hat außerdem eine »persönliche Grenze« eingeführt, einen Abstand von etwa einem Meter zwischen dem Avatar und Nicht-Freunden.[111] Diese Funktion beinhaltet bestimmte Vorgaben, wenn sich zwei Nutzer:innen zum ersten Mal treffen. Wenn zum Beispiel die »persönliche Grenze« der einen Nutzer:in ausgeschaltet ist, während die des:der anderen für alle aktiv ist, dann richtet das Metaversum einen Abstand von einem Meter zwischen beiden Nutzer:innen ein. Dies kann zwar das Sicherheitsgefühl der Nutzer:innen erhöhen, doch die Neuerungen wurden erst nach Berichten über sexuelle Übergriffe eingeführt. Daher besteht auch der Bedarf, Risiken proaktiv zu bewerten und Maßnahmen zu ergreifen, um zu verhindern, dass die Schäden des heutigen Internets in großem Umfang verschlimmert werden. Letztlich sollte die Verantwortung aber nicht bei den Nutzer:innen, sondern bei den Anbieter:innen liegen, die ihre Nutzer:innen schützen müssen.

Förderung der Festlegung von Standards: Das Metaversum umfasst eine vielfältige Landschaft verschiedener Anwendungen, die von einzelnen Entwickler:innen erstellt, geleitet und betrieben und den Nutzer:innen über hardwarespezifische Stores zur Verfügung gestellt werden. Folglich sollten politische Entscheidungsträger:innen und Regulierungsbehörden bei der Regulierung des Metaversums die Entwicklung und Anwendung horizontaler Regeln für verschiedene Dienste in Betracht ziehen.[112] Eine auf Reddit durchgeführte Forschung legt nahe, dass Normen auf Makroebene »Moderator:innen neuer und entstehender Gemeinschaften dabei helfen können, ihre Regulierungsvorschriften bereits während der Gründungsphase der Gemeinschaft zu gestalten« – allerdings nur, wenn solche systemweiten Normen den Moderator:innen bekannt sind.[113] Allgemein gültige Normen und Standards für akzeptables Verhalten im Metaversum befinden sich noch in der Entstehungsphase. Einige Nutzer:innen verstoßen vielleicht unwissentlich gegen die Erwartungen an akzeptierte Verhaltensweisen, während andere absichtlich Schaden anrichten wollen.[114] Gleichgesinnte politische Entscheidungsträger:innen und Regulierungsbehörden sollten die Anbieter, die die Technologien des Metaversums entwickeln, aktiv einbeziehen, um sie frühzeitig zur Verantwortung zu ziehen. An diesem Austausch sollten auch die Entwickler:innen von Metaversum-Anwendungen beteiligt sein, um ihre Richtlinien zur Verwendung von VR-Hardware zu diskutieren. Auf EU-Ebene hat die Europäische Kommission zu öffentlichem Feedback aufgerufen, um eine »Vision für neu entstehende virtuelle Welten« zu entwickeln, die auf der »Achtung der digitalen Rechte und der Gesetze und Werte der EU« beruht.[115] Die Aufforderung unterstreicht die Notwendigkeit für »bereichsübergreifende Voraussetzungen, wie beispielsweise geeignete Verwaltungsmodelle, um die Führungsposition der EU bei der Entwicklung und Standardisierung virtueller Welten zu gewährleisten«.[116]

Multi-Stakeholder-Dialoge: Der multilaterale Austausch zum Metaversum muss ebenfalls ausgeweitet werden, um eine Angleichung an weltweit akzeptierte Bedingungen, wie beispielsweise die grundlegenden Menschenrechte, zu gewährleisten. Ein guter Ausgangspunkt könnte ein erweiterter Austausch im Rahmen des EU-US-Handels- und Technologierats (TTC) sein, um eine transatlantische Abstimmung zu gewährleisten. Das Weltwirtschaftsforum hat in Zusammenarbeit mit INTERPOL, Meta, Microsoft und anderen Akteuren aus der Wissenschaft, der Zivilgesellschaft, der Regierung und der Industrie die Initiative »Defining and Building the Metaverse«[117] (Definition und Erstellung des Metaversums) ins Leben gerufen, die die Entwicklung von Governance-[118] und Wertschöpfungsmechanismen beinhaltet.[119] Ein im Juli 2023 veröffentlichter Bericht erkennt die derzeitige Dominanz des globalen Nordens in diesem Bereich an und betont die Notwendigkeit einer »weltweiten Zusammenarbeit zwischen verschiedenen Interessengruppen, einschließlich Wissenschaftler:innen, Regulierungsbehörden, politischen Entscheidungsträger:innen und Designteams, um das Verständnis fürs Metaversum zu fördern und Schutzmaßnahmen einzuführen«. Gleichzeitig besteht die Erkenntnis, dass »in verschiedenen Ländern oder Regionen oder für verschiedene Gemeinschaften besondere Risiken entstehen könnten«.[120] Die Berücksichtigung von Inklusivität und Diversität wird entscheidend sein, um eine Diskussion unter Einbezug von intersektionellen Perspektiven zu ermöglichen.

Fazit

In demselben Ausmaß, wie der digitale Informationsraum dezentraler, generativer und immersiver wird, werden sich auch die Schwere und die Wahrscheinlichkeit von Schadensrisiken verändern. Es ist noch zu früh, um die genauen Veränderungen vorherzusagen, doch einige relevante Trends lassen sich bereits jetzt beobachten. Erstens bietet das dezentralisierte Social Web (z. B. Odysee, PeerTube) einen neuen, unregulierten und vernetzten Zufluchtsort und Finanzierungsmechanismus für böswillige Akteure. Zweitens können LLMs (z. B. GPT-4.0, DALL-E), denen KI-Anwendungen zugrunde liegen, nicht nur inhärent schädliche Inhalte erzeugen, sondern auch von böswilligen Akteuren für billige, automatisierte und überzeugende Informationsmanipulation ausgenutzt werden. Und drittens bieten immersive Welten (z. B. Horizon Worlds) böswilligen Akteuren nicht nur neue Möglichkeiten der Mobilisierung, Finanzierung und Planung von Aktivitäten, sondern sie verstärken auch möglicherweise die Auswirkungen schädlicher Inhalte und Verhaltensweisen. Dies ist insbesondere dann der Fall, wenn immersive Welten für alle unsere Sinne zugänglich werden (zum Beispiel durch Berührung). In Anbetracht der erheblichen Schadensrisiken sollten sich politische Entscheidungsträger:innen und Regulierungsbehörden bereits frühzeitig mit einer durchdachte Zukunftsgestaltung der analysierten neuen Plattformen und Technologien beschäftigen. Die Fragen der Haftung, der Sicherheit, der Transparenz und der Durchsetzung müssen neu diskutiert werden. Vor dem Hintergrund der Konvergenz neuer Plattformen und Technologien muss ein enger Austausch mit der Wissenschaft, der Zivilgesellschaft und mit verschiedenen Bereichen der Industrie stattfinden, um ein ganzheitliches und tiefes Verständnis der sich entwickelnden Bedrohungslage zu gewinnen. Darüber müssen die Regelwerke kontinuierlich anpasst und weiterentwickelt werden, um die jeweiligen Schadensrisiken zu mindern, und nach Wegen zu suchen, um bereits bestehende Regelvorschriften durchzusetzen.

Endnoten

1 **Yu, M., 2022**. *SPC Releases Typical Cases on Protection of Personality Rights*. [Online] Available at: https://www.chinajusticeobserver.com/a/spc-releases-typical-cases-on-protection-of-personality-rights [Accessed 31 July 2023]

2 **McLuhan, M., Hutchon, K., McLuhan, E., 1980**. Media, message, and language. National *Textbook Company. Skokie, IL*.

3 **World Economic Forum (WEF), 2023**. *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. Available at: https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf [Accessed 24 August 2023]

4 **ISD, 2023**. *Digital Policy Lab*. [Online] Available at: https://www.isdglobal.org/digital-policy-lab/ [Accessed 29 August 2023]

5 **Heath, A., 2023**. *This is what Instagram's upcoming Twitter competitor looks like*. [Online] Available at: https://www.theverge.com/2023/6/8/23754304/instagram-meta-twitter-competitor-threads-activitypub [Accessed 8 August 2023]

6 **Edelman, G**. *The Father of Web3 Wants You to Trust Less*. [Online] Available at: https://www.wired.com/story/web3-gavin-wood-interview/ [Accessed 31 July 2023]

7 **Wilson, J., 2023**. *Extremist-friendly tech company closes after legal fine, The Guardian*, 16 July. Available at: https://www.theguardian.com/technology/2023/jul/16/lbry-closes-odysee-cryptocurrency-tech-sec-fraud-extremist [Accessed 8 August 2023]

8 **European Data Protection Supervisor (EDPS), 2022**. *EDPS launches pilot phase of two social media platforms*. [Online] Available at: https://edps.europa.eu/press-publications/press-news/press-releases/2022/edps-launches-pilot-phase-two-social-media_en [Accessed 31 July 2023]

9 https://social.bund.de/about

10 **TechAgainstTerrorism, 2023**. *State of play. Trends in terrorist and violent extremist use of the internet. 2022*. [Online] Available at: https://www.techagainstterrorism.org/wp-content/uploads/2023/01/FINAL-State-of-Play-2022-TAT.pdf [Accessed 8 August 2023]

11 **Christchurch Call, 2022**. *Christchurch Call Initiative on Algorithmic Outcomes*. [Online] Available at: https://www.christchurchcall.com/media-and-resources/news-and-updates/christchurch-call-initiative-on-algorithmic-outcomes/ [Accessed 8 August 2023]

12 **Hammer, D., Gerster, L. & Schwieter, C., 2023**. *Im digitalen Labyrinth. Rechtsextreme Strategien der Dezentralisierung im Netz und mögliche Gegenmaßnahmen*. [Online] Available at: https://isdgermany.org/im-digitalen-labyrinth/ [Accessed 31 July 2023]

13 **Ibid.**

14 **Gerster, A., Arcostanzo, F., Prieto-Chavana, N., et al., 2023**. *The Hydra on the Web: Challenges Associated with Extremist Use of the Fediverse – A Case Study of PeerTube*. [Online] Available at: https://www.isdglobal.org/isd-publications/the-hydra-on-the-web-challenges-associated-with-extremist-use-of-the-fediverse-a-case-study-of-peertube/ [Accessed 29 August 2023]

15 **Rozenshtein, A. Z., 2023**. *Moderating the Fediverse. Content moderation on distributed social media*. [Online] Available at: https://www.journaloffreespeechlaw.org/rozenshtein2.pdf [Accessed 31 July 2023]

16 **Kessels, B., 2022**. *The Fediverse never Forgets*. [Online] Available at: https://berk.es/2022/12/23/fediverse-never-forgets/ [Accessed 31 July 2023]

17 **Hamari, J., Koivisto, J., & Sarsa, H., 2014**. Does gamification work? A literature review of empirical studies on gamification. *2014 47th Hawaii international conference on system sciences*. 3025-3034.

18 **Matlach, P., Hammer, D. & Schwieter, C., 2022**. *On Odysee: The Role of Blockchain Technology for Monetisation in the Far-Right Online Milieu*. [Online] Available at: https://www.isdglobal.org/isd-publications/on-odysee-the-role-of-blockchain-technology-for-monetisation-in-the-far-right-online-milieu/ [Accessed 31 July 2023]

19 **Ibid.**

20 **Wilson, J., 2023**. *Extremist-friendly tech company closes after legal fine, The Guardian*, 16 July. Available at: https://www.theguardian.com/technology/2023/jul/16/lbry-closes-odysee-cryptocurrency-tech-sec-fraud-extremist [Accessed 8 August 2023]

21 **Odysee, 2022**. *Declaration of Indifference: Community Guidelines*. [Online] Available at: https://help.odysee.tv/communityguidelines/ [Accessed 29 August 2023]

22 **Hammer, D., Gerster, L. & Schwieter, C., 2023**. *Im digitalen Labyrinth. Rechtsextreme Strategien der Dezentralisierung im Netz und mögliche Gegenmaßnahmen*. [Online] Available at: https://isdgermany.org/im-digitalen-labyrinth/ [Accessed 31 July 2023]

23 **Bundesamt für Justiz (BfJ), 2023**. Änderung der Pressemitteilung vom 2. März 2023: Bundesamt für Justiz hält Bußgeldbescheide in Höhe von 5,125 Millionen Euro gegen das soziale Netzwerk Telegram aufrecht. [Online] Available at: https://www.bundesjustizamt.de/DE/ServiceGSB/Presse/Pressemitteilungen/2023/20230302.html [Accessed 31 July 2023]

24 **Framasoft, 2023**. *PeerTube instances*. [Online] Available at: https://instances.joinpeertube.org/instances [Accessed 29 August 2023]

25 **Future of Life Institute, 2023**. *Pause Giant AI Experiments: An Open Letter.* [Online] Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [Accessed 10 July 2023]

26 **Coulter, M., 2023**. *AI experts disown Musk-backed campaign citing their research, Reuters, 5 April.* Available at: https://www.reuters.com/technology/ai-experts-disown-musk-backed-campaign-citing-their-research-2023-03-31/ [Accessed 24 August 2023]

27 **IBM, 2023.** *What is a neural network?* [Online] Available at: https://www.ibm.com/topics/neural-networks#:~:text=Artificial%20neural%20networks%20(ANNs)%20are,an%20associated%20weight%20and%20threshold [Accessed 8 August 2023]

28 **Weidinger, L., Mellor, J., Rauh, M, et al., 2021**. *Ethical and social risks of harm from.* [Online] Available at: https://www.deepmind.com/publications/ethical-and-social-risks-of-harm-from-language-models [Accessed 10 July 2023]; Goldstein, J. A. et al., 2023. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

29 **European Parliament, 2023**. *EU AI Act: first regulation on artificial intelligence.* [Online] Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence?&at_campaign=20226-Digital&at_medium=Google_Ads&at_platform=Search&at_creation=RSA&at_goal=TR_G&at_advertiser=Webcomm&at_audien [Accessed 17 July 2023]; **eSafety Commissioner, 2023**. *Tech Trends Position Statement. Generative AI.* [Online] Available at: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai [Accessed 24 August 2023]

30 **Europol, 2023**. *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Luxemburg: Publications Office of the European Union.

31 **Solaiman, I., Brundage, M., Clark, J., 2019**. *OpenAI Report: Release Strategies and the Release Strategies and the.* [Online] Available at: https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf [Accessed 12 July 2023]; Partnership on AI, 2023. *PAI's Responsible Practices for Practices for A Framework for Collective Action.* [Online] Available at: https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf [Accessed 12 July 2023]

32 **Weidinger, L., Mellor, J., Rauh, M, et al., 2021**. *Ethical and social risks of harm from.* [Online] Available at: https://www.deepmind.com/publications/ethical-and-social-risks-of-harm-from-language-models [Accessed 10 July 2023]

33 **Partnership on AI (PAI), 2023**. *PAI's Responsible Practices for Practices for A Framework for Collective Action.* [Online] Available at: https://partnershiponai.org/wp-content/uploads/2023/02/PAI_synthetic_media_framework.pdf [Accessed 12 July 2023]

34 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

35 **Bowman, S. R., 2023**. *Eight Things to Know about Large Language Models.* [Online] Available at: https://arxiv.org/abs/2304.00612 [Accessed 31 July 2023]

36 **Gehman, S., Gururangan, S., Sap, M., et al., 2020**. *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models.* [Online] Available at: https://arxiv.org/abs/2009.11462 [Accessed 11 7 2023]

37 **Abid, A., Farooqi , M. & Zou, J., 2021**. Large language models associate Muslims with violence. *Nat Mach Intell,* Volume 3, 461–463.

38 **Deshpande, A., Murahari, V., Rajpurohit, T., et al., 2023**. *Toxicity in CHATGPT: Analyzing Persona-assigned Language Models.* [Online] Available at: https://arxiv.org/pdf/2304.05335.pdf [Accessed 11 July 2023]

39 **Rae, J. W., Borgeaud, S., Cai, T., et al., 2021**. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher.* [Online] Available at: https://arxiv.org/abs/2112.11446 [Accessed 17 July 2023]

40 **Brewster, J., Arvanitis, L. & Sadeghi, M., 2023**. *The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation At Unprecedented Scale.* [Online] Available at: https://www.newsguardtech.com/misinformation-monitor/jan-2023/ [Accessed 11 July 2023]

41 **Arvantis, L., Sadeghi, M. & Brewster, J., 2023**. *Despite OpenAI's Promises, the Company's New AI Tool Produces Misinformation More Frequently, and More Persuasively, than its Predecessor.* [Online] Available at: https://www.newsguardtech.com/misinformation-monitor/march-2023/ [Accessed 11 July 2023]

42 **Horvitz, E., 2023**. *On the Horizon: Interacive and Compositional Deepfakes.* [Online] Available at: https://arxiv.org/abs/2209.01714 [Accessed 31 July 2023]

43 **EEAS, 2023**. *1st EEAS Report on Foreign Information Manipulation and Interference Threats.* [Online] Available at: https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-DataTeam-ThreatReport-2023..pdf [Accessed 11 July 2023]

44 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

45 **Van Sant, S. & Goujard, C., 2022**. *European Parliament website hit by cyberattack after Russian terrorism vote.* [Online] Available at: https://www.politico.eu/article/cyber-attack-european-parliament-website-after-russian-terrorism/ [Accessed 31 July 2023]

46 **Europol, 2023.** *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Luxemburg: Publications Office of the European Union.

47 **Hill, M., 2023**. *6 ways generative AI chatbots and LLMs can enhance cybersecurity (CSO).* [Online] Available at: https://www.csoonline.com/article/575377/6-ways-generative-ai-chatbots-and-llms-can-enhance-cybersecurity.html [Accessed 31 July 2023]

48 **Bessi, A. & Emilio, F., 2016**. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday,* 21(11); Wang, E. L., Luceri, L., Pierri, . F. & Ferrara, E., 2023. *Identifying and characterizing behavioral classes of radicalization within the QAnon conspiracy on Twitter.* s.l., Proceedings of the International AAAI Conference.

49 **Heubl, B., Sabolwski, N. & Weinmann, L., 2023**. *Vulkan Files #michgibtsgarnicht*, SZ, 31 March. Available at: https://www.sueddeutsche.de/projekte/artikel/politik/russland-cyberkrieg-desinformation-propaganda-fakenews-twitter-vulkan-files-ukraine-e287057/?reduced=true [Accessed 31 July 2023]

50 **Europol, 2023.** *ChatGPT - The impact of Large Language Models on Law Enforcement, a Tech Watch Flash Report from the Europol Innovation Lab.* Luxemburg: Publications Office of the European Union.

51 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

52 **Ibid.**

53 **Meta Fundamental AI Research Diplomacy Team (FAIR), Brown, N., Dinan, E., et al., 2022**. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science,* 378(6624).

54 **Bai, H., Voelkel, J. G., Eichstaedt, J. C., et al., 2023**. *Artificial Intelligence Can Persuade Humans on Political Issues.* [Online] Available at: https://osf.io/stakv/ [Accessed 17 July 2023]

55 **Spitale, G., Biller-Andorno, N. & Germani, F., 2023**. AI model GPT-3 (dis)informs us better than humans. *ScienceAdvances,* 9(26).

56 **Weidinger, L., Mellor, J., Rauh, M, et al., 2021**. *Ethical and social risks of harm from.* [Online] Available at: https://www.deepmind.com/publications/ethical-and-social-risks-of-harm-from-language-models [Accessed 10 July 2023]

57 **YouGov, 2023**. *KI – Chance oder Bedrohung?.* [Online] Available at: https://yougov.de/topics/technology/articles-reports/2023/05/17/ki-chance-oder-bedrohung [Accessed 17 July 2023]

58 **Etienne, H., 2021**. The future of online trust (and why Deepfake is advancing it). *AI Ethics*, 1, 553-562.

59 **Anderljung, M., Barnhart, J., Leung, J., et al., 2023**. *Frontier AI Regulation: Managing Emerging Risks to Public Safety.* [Online] Available at: arxiv.org/pdf/2307.03718.pdf [Accessed 26 July 2023]

60 **CSIS, 2023**. *Sen. Chuck Schumer Launches SAFE Innovation in the AI Age at CSIS.* [Online] Available at: https://www.csis.org/analysis/sen-chuck-schumer-launches-safe-innovation-ai-age-csis [Accessed 31 July 2023]

61 **White House, 2023**. *Ensuring Safe, Secure, and Trustworthy AI.* [Online] Available at: Ensuring-Safe-Secure-and-Trustworthy-AI.pdf (whitehouse.gov) [Accessed 26 July 2023]

62 **European Union, 2023**. *MEPs ready to negotiate first-ever rules for safe and transparent AI.* [Online] Available at: https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai [Accessed 26 July 2023]

63 **European Commission, 2023**. *Artificial intelligence: in Europe, innovation and safety go hand in hand. Statement by Commissioner Thierry Breton.* [Online] Available at: Artificial intelligence | Statement by Commissioner Breton (europa.eu) [Accessed 26 July 2023]

64 **Novelli, C., Taddeo, M. & Floridi, L., 2023**. *Accountability in artificial intelligence: what it is and how it works.* [Online] Available at: https://link.springer.com/article/10.1007/s00146-023-01635-y [Accessed 31 July 2023]

65 **Ibid.**

66 **Ariyaratne, H., 2023**. *ChatGPT and intermediary liability: Why Section 230 does not and should not protect generative algorithms.* [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4422583 [Accessed 31 July 2023]

67 **eSafety Commissioner, 2023**. *Tech Trends Position Statement. Generative AI.* [Online] Available at: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai [Accessed 24 August 2023]

68 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

69 **Buchanan, B., Lohn, A., Musser, M., et al., 2021**. *Truth, Lies, and Automation. How language models could change disinformation.* Center for Security and Emerging Technology, Washington, DC.

70 **Deshpande, A., Murahari, V., Rajpurohit, T., et al., 2023**. *Toxicity in CHATGPT: Analyzing Persona-assigned Language Models*. [Online] Available at: https://arxiv.org/pdf/2304.05335.pdf [Accessed 11 July 2023]

71 **Greer, C., 2017**. *International Personality Rights and Holographic Portrayals*. [Online] Available at: https://doi.org/10.18060/7909.0052 [Accessed 26 July 2023]

72 **Celli, F., 2020**. *Deepfakes Are Coming: Does Australia Come Prepared?* [Online] Available at: http://classic.austlii.edu.au/au/journals/CanLawRw/2020/18.pdf [Accessed 26 July 2023]

73 **eSafety Commissioner, 2023**. *Tech Trends Position Statement. Generative AI*. [Online] Available at: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai [Accessed 24 August 2023]

74 **Goujard, C., 2023**. EU wants Google, Facebook to start labeling AI-generated content. *POLITICO*. Available at: https://www.politico.eu/article/chatgpt-dalle-google-facebook-microsoft-eu-wants-to-start-labeling-ai-generated-content/ [Accessed 29 August 2023]

75 **Goldstein, J. A., Sastry, G., Musser, M., et al., 2023**. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. [Online] Available at: https://cdn.openai.com/papers/forecasting-misuse.pdf [Accessed 12 July 2023]

76 **Coalition for Content Provenance and Authenticity (C2PA)**. *Overview*. [Online] Available at: https://c2pa.org/ [Accessed 8 August 2023]

77 **Partnership on AI (PAI)**, 2023. *About Us*. [Online] Available at: https://partnershiponai.org/about/ [Accessed 8 August 2023]

78 **Anderson, J. & Rainie, L., 2023**. *The Metaverse in 2040*. [Online] https://www.pewresearch.org/internet/2022/06/30/the-metaverse-in-2040/ [Accessed 31 July 2023]

79 **Ibid.**

80 **The New York Times, 2021.** The Metaverse Is Mark Zuckerberg's Escape Hatch, *The New York* Times, 29 October. Available at: https://www.nytimes.com/2021/10/29/technology/meta-facebookzuckerberg.html [Accessed 31 July 2023]

81 **Meta, 2021**. *Connect 2021: Our vision for the metaverse*. [Online] Available at: https://tech.fb.com/ar-vr/2021/10/connect-2021-our-vision-for-the-metaverse/ [Accessed 31 July 2023]

82 **Newton, C., 2021**. *Mark in the metaverse*. [Online] Available at: https://www.theverge.com/22588022/mark-zuckerberg-facebook-ceo-metaverse-interview [Accessed 31 July 2023]

83 **Ball, M., 2020**. *The Metaverse: What It Is, Where to Find it, and Who Will Build It*. [Online] Available at: https://www.matthewball.vc/all/themetaverse [Accessed 31 July 2023]

84 **Heath, A., 2021**. Meta opens up access to its VR social platform Horizon Worlds, *The Verge*, 9 December. Available at: https://www.theverge.com/2021/12/9/22825139/meta-horizon-worlds-access-open-metaverse [Accessed 31 July 2023]

85 **World Economic Forum (WEF), 2023**. Interoperability in the Metaverse. [Online] Available at: https://www.weforum.org/reports/interoperability-in-the-metaverse/ [Accessed 24 August 2023]

86 **Europol, 2022**. *Policing in the metaverse: what law enforcement needs to know*. Luxembourg: Publications Office of the European Union.

87 **Council of the European Union, 2022**. *The Metaverse in the context of the fight against terrorism*. EU Counter-Terrorism Coordinator. [Online] Available at: https://data.consilium.europa.eu/doc/document/ST-9292-2022-INIT/en/pdf [Accessed 8 August 2023]

88 **WeProtect Global Alliance, 2023**. Intelligence briefing. Extended Reality technologies and child sexual exploitation and abuse. [Online] Available at: https://www.weprotect.org/library/extended-reality-technologies-and-child-sexual-exploitation-and-abuse/ [Accessed 24 August 2023]

89 **World Economic Forum (WEF), 2023**. Metaverse Privacy and Safety. [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 24 August 2023]

90 **Ashraf, M. 2023**. *Gender-based Abuse on the Metaverse: The New Internet is Being Coded on a Toxic Palimpsest*. [Online] Available at: https://botpopuli.net/gender-based-abuse-on-the-metaverse-the-new-internet-is-being-coded-on-a-toxic-palimpsest/ [Accessed 31 July 2023]

91 **Patel, N. J., 2021**. *Reality or Fiction?* [Online] Available at: https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0 [Accessed 31 July 2023]

92 **Gray, K.L, 2016**. Solidarity is for white women in gaming. Using Critical Discourse Analysis To Examine Gendered Alliances and Racialized Discords in an Online Gaming Forum. In: Kafai, Y.B., Richard, G.T., Tynes, B.M., et al., 2016. *Diversifying Barbie and Mortal Kombat: Intersectional perspectives and inclusive designs in gaming*. Carnegie Mellon University, ETC Press, Pittsburgh, PA.

93 **WeProtect Global Alliance, 2023**. Intelligence briefing. Extended Reality technologies and child sexual exploitation and abuse. [Online] Available at: https://www.weprotect.org/library/extended-reality-technologies-and-child-sexual-exploitation-and-abuse/ [Accessed 24 August 2023]

94 **Blackwell, L., Ellison, N., Elliott-Deflo, N., & et al., 2019**. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction Vol. 3* Issue CSCW, Article No.: 100,1–25.

95 **Mrug, S., Madan, A. & Windle, M., 2016**. Emotional Desensitization to Violence Contributes to Adolescents' Violent Behavior. *Journal of Abnormal Child Psychology,* 44 (1), 75-86.

96 **Rankin C.H. et al., 2009**. Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory.* 92, 135-138.

97 **Mrug, S., Madan, A. & Windle, M., 2016**. Emotional Desensitization to Violence Contributes to Adolescents' Violent Behavior. *Journal of Abnormal Child Psychology,* 44 (1), 75-86.

98 **Hayward, K.J. & Cottee, S., 2011**. Terrorist (e)motives: the existential attractions of terrorism. *Studies in Conflict and Terrorism,* 34(12), 963-986.

99 **Council of the European Union, 2022**. *The Metaverse in the context of the fight against terrorism.* EU Counter-Terrorism Coordinator. [Online] Available at: https://data.consilium.europa.eu/doc/document/ST-9292-2022-INIT/en/pdf [Accessed 8 August 2023]

100 **Europol, 2022**. *Policing in the metaverse: what law enforcement needs to know.* Publications Office of the European Union, Luxemburg.

101 **The Algemeiner, 2022**. *Children's Gaming Platform Removes 'Disturbing' Nazi Concentration Camp 'Experience' With Gas Chambers.* [Online] Available at: https://www.algemeiner.com/2022/02/21/childrens-gaming-platform-removes-disturbing-nazi-concentration-camp-experience-with-gas-chambers/ [Accessed 31 July 2023]

102 **Financial Times**. *NFTs: The metaverse economy.* [Online] Available at: https://www.ft.com/partnercontent/crypto-com/nfts-the-metaverse-economy.html [Accessed 31 July 2023]

103 **Council of the European Union, 2022**. *The Metaverse in the context of the fight against terrorism.* EU Counter-Terrorism Coordinator. [Online] Available at: https://data.consilium.europa.eu/doc/document/ST-9292-2022-INIT/en/pdf [Accessed 8 August 2023]

104 **World Economic Forum, 2023**. *Metaverse Privacy and Safety.* [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

105 **Europol, 2022**. *Policing in the metaverse: what law enforcement needs to know.* Publications Office of the European Union, Luxemburg.

106 **World Economic Forum, 2023**. *Metaverse Privacy and Safety.* [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

107 **Ibid.**

108 **Tuck, H., Guhl, J., et al., 2023**. *Researching the Evolving Online Ecosystem: Telegram, Discord & Odysee.* [Online] Available at: https://www.isdglobal.org/isd-publications/researching-evolving-online-ecosystem-telegram-discord-odysee/ [Accessed 8 August 2023]

109 **World Economic Forum, 2023**. *Metaverse Privacy and Safety.* [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

110 **Meta, 2023**. *What is the Safe Zone in Horizon Worlds?* [Online] Available at: https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/safe-zone-in-horizon/ [Accessed 31 July 2023]

111 **Meta, 2022**. *Introducing a Personal Boundary for Horizon Worlds and Venues.* [Online] Available at: https://about.fb.com/news/2022/02/personal-boundary-horizon/ [Accessed 31 July 2023]

112 **Blackwell, L., Ellison, N. Elliott-Deflo, N., & Schwartz, R., 2019**. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Art. No. 100, 1-25.

113 **Chandrasekharan, E. et al., 2018**. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer* Interaction, 2(CSCW), Art. No. 32, 1–25.

114 **Blackwell, L., Ellison, N. Elliott-Deflo, N., & Schwartz, R., 2019**. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Art. No. 100, 1-25.

115 **European Commission, 2023**. *Virtual worlds (metaverses) – a vision for openness, safety and respect.* [Online] Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13757-Virtual-worlds-metaverses-a-vision-for-openness-safety-and-respect_en [Accessed 31 July 2023]

116 **Ibid.**

117 **World Economic Forum, 2023**. *Partners.* [Online] Available at: https://initiatives.weforum.org/defining-and-building-the-metaverse/partners [Accessed 31 July 2023]

118 **Ibid.**

119 **Ibid.**

120 **World Economic Forum, 2023**. *Metaverse Privacy and Safety.* [Online] Available at: https://www3.weforum.org/docs/WEF_Metaverse_Privacy_and_Safety_2023.pdf [Accessed 31 July 2023]

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdgermany.org

Gefördert durch:

Auswärtiges Amt

aufgrund eines Beschlusses des Deutschen Bundestages