Karlsruher Institut für Technologie

Technologiemonitoring: Das Potenzial von Metaverse und KI für extremistische Verwendungszwecke

Octavia Madeira, Georg Plattner, Alexandros Gazos, Tim Röller, Christian Büscher

Technologiemonitoring

Zusammenfassung

*Im Rahmen des Technologiemonitorings wurden zwei Technologiekomplexe als besonders relevant identifiziert und aufgrund ihrer dynamischen Entwicklung untersucht. Das ist zum einen die Entwicklung eines „erweiterten Internets" in Form des virtuellen Raums „Metaverse" und zum anderen Anwendungen im Themenkreis „künstlicher Intelligenz" (KI). Beiden Technologiekomplexen sind bestimmte Eigenschaften eingeschrieben, deren Entwicklung Affordanzen hervorbringen können, die sich aktuell nur antizipieren lassen und in absehbarer Zukunft auch für Extremist*innen von Interesse sind. Für die Vision des Metaverses wurden mithilfe von Expert*innen Szenarien generiert, die (extremistische) Freiheit und Sicherheit in virtuellen sozialen Interaktionen miteinander in Relation setzen. Für den Fall der KI wurden verschiedene Anwendungen sowohl im extremistischen Kontext als auch aus der Perspektive der Prävention betrachtet. Die Kontingenz zukünftiger Entwicklungspfade und die Möglichkeit unerwünschter Ausprägungen technischer Affordanzen erfordern die frühzeitige transdisziplinäre Einbindung verschiedener Stakeholder*innen. Mehr noch als technische Lösungsansätze ist jedoch der soziotechnische Kontext entscheidend, da beide Technologiekomplexe die Anforderungen an Präventionsarbeit und Bildungsangebote verändern. Die höchst dynamische Entwicklung neuer Technologiekomplexe erfordert deshalb ein Echtzeitmonitoring, bei dem die Technikfolgenabschätzung geeignete Methoden zur Verfügung stellt.*

Schlüsselworte

Technikfolgenabschätzung | Technologiemonitoring | Metaverse | KI | Radikalisierung

Die Ambivalenz neuer Technologien als Problem für das Technologiemonitoring

In den vergangenen Jahren hat das Innovationspotenzial extremistischer Akteur*innen (Personen, Gruppen, Organisationen) im Hinblick auf den Missbrauch oder die schädliche Kombination von Technologien großes Interesse geweckt. Ein Technologiemonitoring soll helfen, diese Entwicklungen zu beobachten und einzuschätzen. Zu den konzeptionellen Herausforderungen eines solchen Monitorings gehören Fragen nach geeigneten Begrifflichkeiten, Theorien und Methoden.

Die Technikfolgenforschung (Technology Assessment, TA) hat sich auf den Begriff des „Dual Use" festgelegt, um das ambivalente Potenzial von Technologien auszuloten (Forge, 2010; Mahfoud et al., 2018). Für das Technologiemonitoring ist dieser Begriff unzureichend. Gerade in den vergangenen Jahren sind zahlreiche „Multi Use"-Technologien entstanden, die sich nicht auf einen Zweck festlegen lassen, wie soziale Medien, eine Fülle an Software-Applikationen, kryptografische Techniken und Anwendungen der künstlichen Intelligenz (KI). Alle diese Technologien haben das Potenzial, *malevolente* Modi Operandi zu unterstützen. Sie sind Ergebnis von Forschung und Innovation mit teils militärischen oder teils zivilen Verwendungszwecken. Sie veranschaulichen, wie Technologien einerseits äußerst nützlich sein können, andererseits aber auch ein hohes Missbrauchspotenzial aufweisen – gerade weil sie eine weite Verfügbarkeit und Nützlichkeit für jetzt noch gar nicht bekannte Zwecke erreichen (Cronin, 2020). Im Weiteren lohnt es sich deshalb, mit Begriffen der malevolenten Kreativität (Cropley et al., 2008) und technologischen Affordanzen[1] zu arbeiten. Sie erlauben es, sowohl den Blick auf das Innovationspotenzial extremistischer Akteur*innen zu werfen als auch den Möglichkeitsraum zukünftiger Nutzungsweisen über Dual Use hinaus zu erfassen. Dazu haben wir an anderer Stelle tiefer gehende Analysen ausgearbeitet (Büscher & Kusche, 2023).

Um dazu beizutragen, den schädlichen Einsatz von Technologie im Bereich der Radikalisierung und des Extremismus zu verstehen, abzuschwächen und zu verhindern, muss die TA ihre Herangehensweise auf der einen Seite

[1] *Das Internetmonitoring der LMU arbeitet mit dem Konzept der Affordanzen, um die kontingenten Nutzungsmöglichkeiten von digitalen Räumen im Hinblick auf Radikalisierung zu untersuchen (siehe Greipl et al. 2022).*

ausweiten. In den vergangenen Jahrzehnten hat diese ihre Reichweite in sozialer und zeitlicher Hinsicht mit deliberativ, inklusiven Ansätzen wie Constructive TA oder frühzeitig im Innovationsprozess ansetzenden Ansätzen wie Real-Time TA (Guston & Sarewitz, 2002) oder Responsible Research and Innovation (Stilgoe et al., 2013) erweitert. Auf der anderen Seite limitiert der Zugang zu Wissensquellen bestimmte Herangehensweisen für unser Thema; einige Daten- und Informationsquellen bleiben unzugänglich. Der kreative Prozess der Planung, um mithilfe von Technologien Schaden anzurichten, verläuft klandestin und ist somit kaum beobachtbar. Für die TA ist es keine vertretbare Option, die Arbeit der Nachrichtendienste nachzuahmen. Folglich wird sie auf Expert*inneneinschätzungen zurückgeworfen, mit all den Einschränkungen, die dies mit sich bringt. Zum einen müssen wir damit rechnen, dass auch Expert*innen durch ihr Wissen und Vorstellungsvermögen limitiert sind. Malevolente Kreativität zeichnet sich gerade dadurch aus, dass ihre innovativen Ergebnisse überraschen. Zum anderen haben wir die Erfahrung gemacht, dass diejenigen Expert*innen, die etwas zu spezifischen Technologien und deren Potenzialen aussagen können, in seltenen Fällen auch Einblick in Prozesse der Radikalisierung und des Extremismus haben – und umgekehrt.

Gleichwohl bietet der breite Kanon an Theorien und Methoden in der TA anpassungsfähige Instrumente zur Beobachtung von Trends in der malevolenten Nutzung von Technologien. In dem vorliegenden Text berichten wir von Methoden des „Vision Assessments“ (VA) und des Expert*innenworkshops.

Angeleitet werden wir dabei durch ein kontinuierliches Monitoring relevanter Publikationen zu den Themenbereichen Technological Foresight und Extremismus/Terrorismus sowie durch Erkenntnisse, die in einem im Jahr 2021 durchgeführten Online-Delphi hervorgehoben wurden (Büscher et al., 2022). Dort wurde unter anderem die These vertreten, dass die Relevanz neuer Technologien für extremistische und terroristische Akteur*innen davon abhängig sei, inwiefern diese ein Bestandteil des Alltagslebens würden und damit als Gelegenheitsstrukturen auch für böswilliges Handeln zur Verfügung stünden.[2] In diesem Sinne ist die von Mark Zuckerberg 2021 vorgestellte

[2] *Auch die Erkenntnisse aus dem MOTRA-Modul „Expertenpanels“ weisen auf die zunehmende Verlagerung von extremistischen Aktivitäten in den virtuellen Raum hin.*

Vision eines Metaverses von besonderem Interesse. Hierzu wollen wir am ITAS mithilfe der Methode des Vision Assessments den Möglichkeitsraum einer Zukunft im Metaverse ausloten, um so Hinweise auf die Ausgestaltung dieser Technologie geben zu können. Visionen sind handlungsleitend für die Entwicklung einer Technologie und VA ermöglicht es, verknüpfte Prozesse kritisch zu beleuchten und Risiken frühzeitig zu erkennen.

Das Kriterium der alltagsweltlichen Präsenz lässt sich auch auf die zweite hier untersuchte Zukunftstechnologie anwenden, die künstliche Intelligenz. Im Verlauf des Technologiemonitorings hat sich gezeigt, dass auch Anwendungen aus diesem Gebiet im Phänomenbereich Extremismus und Radikalisierung eine hohe Relevanz besitzen. Dies bestätigen auch die Ergebnisse unserer Delphi-Studie, in der verschiedene neue Technologien anhand der Kriterien Nützlichkeit und Verfügbarkeit von einschlägigen Expert*innen eingeordnet wurden. Bezüglich verschiedener KI-Anwendungen (zum Beispiel Adversarial Attacks, Gesichtserkennung) kam es jedoch teilweise zu widersprüchlichen Bewertungen. Folglich eruierten wir das malevolente Potenzial von KI tiefergehend mithilfe eines Expert*innenworkshops, um neues Wissen über unerwünschte Folgen dieser Innovation im Themenfeld Radikalisierung und Extremismus zu generieren.

Das Metaverse: Extremismus und Radikalisierung im Internet von morgen

Willkommen im Metaverse:
Die Vision von der Zukunft des Internets als erweiterte Wirklichkeit

Laut der von Meta-CEO Mark Zuckerberg im Oktober 2021 vorgestellten Idee soll das sogenannte „Metaverse“ das mobile Internet revolutionieren und zu einer plattformübergreifenden Entität ausgebaut werden, die das Leben aller Nutzer*innen im digitalen Raum erweitern soll (Meta, 2021b, 2021a). Im Gegensatz zum heutigen Stand der Technik sollen die Nutzer*innen mittels Virtual Reality (VR) immersiv und mit eigenem Avatar am Metaverse teilnehmen können. Die Nutzer*innen sollen sich aktiv an der Entstehung und Ausbreitung des Metaverses beteiligen. Hierfür sollen zahlreiche technologische Schnittstellen zur Verfügung gestellt werden,

die diese Nutzer*innenbeteiligung in relevanten Lebensbereichen wie zum Beispiel Arbeit, Handel oder Freizeit ermöglichen sollen. Damit ist eine umfangreiche *Vision* umrissen, wie in Zukunft soziale Interaktion in einer virtuellen Realität in einem für alle zugänglichen Erlebnis- und Handlungsraum stattfinden soll.

Wie zu erwarten, trifft die Vision des Metaverses auch auf erhebliche Kritik. Neben dem Vorbehalt, lediglich auf ökonomische Interessen abzuzielen, spielen strafrechtlich relevante Aspekte eine wichtige Rolle. Nutzer*innen des Metaverse-Vorläufers „Horizon World" berichteten von sexuellen Übergriffen, gegen die kaum oder gar nicht vorgegangen wurde (Bazu, 2021; Chohan, 2022; Diaz, 2022; Singh, 2022; Wiederhold, 2022). Auch für die aktuelle Iteration virtueller interaktiver Welten werden durch Forscher*innen unterschiedliche Missbrauchspotenziale durchgespielt, die vor allem auf terroristische Aktivitäten Bezug nehmen (Weimann & Dimant, 2023). Im Gegensatz zu heutigen Social-Media-Plattformen steht im zukünftigen Metaverse nicht allein Text- und Bildmaterial zur Verfügung, sondern es zeichnet sich vor allem durch Echtzeitkommunikation aus. Dies wird die Dynamik der sozialen Interaktion beeinflussen. Es bleibt abzuwarten, inwiefern Erfahrungswerte im Umgang mit vorangegangenen Versuchen der virtuellen Interaktion beziehungsweise mit aktuellen Social-Media-Plattformen auf das neue Projekt übertragbar sind.

Regelungsmöglichkeiten über (digitale) Staatsgrenzen hinweg werden eine große Bedeutung erhalten. Die Möglichkeiten der Aufrechterhaltung und Durchsetzung demokratischer Werte und Standards im Metaverse als „platform-based public sphere" werden aktuell kontrovers diskutiert (Huq, 2022). Während die Strafverfolgungsbehörden in der heutigen digitalen Welt auf einen gewissen Erfahrungsschatz aufbauen können, stehen Organisationen mit Sicherheitsaufgaben angesichts der rasanten Entwicklung neuer Technologien vor neuen Herausforderungen bezüglich Beobachtung und Intervention (Europol, 2022). Die Entwicklung des Metaverses könnte beispielsweise neue Arten von virtuellen Straftaten hervorbringen, aber auch „klassischen" Verbrechen wie (Identitäts-)Betrug eine neue Dimension verleihen (Chohan, 2022). Gleichzeitig ist sicherzustellen, dass Grundrechte wie freie Meinungsäußerung, aber auch Datenschutz beziehungsweise -souveränität nicht eingeschränkt werden. Folglich kommt dem Balanceakt zwischen Sicherheit und Freiheit eine besondere Bedeutung zu.

Vision Assessment: Vier Szenarien einer möglichen Zukunft im Metaverse

Im Rahmen des Technologiemonitorings setzt sich das Team am ITAS mit der Frage auseinander, wie sich neue Technologien auf Radikalisierung und Extremismus auswirken können, und nutzt hierfür unterschiedliche Methoden der Technikfolgenabschätzung. Für die Analyse der Vision des Metaverses stellt das Vision Assessment eine geeignete Methode dar. Visionen sind Erzählungen, die von bestimmten Akteur*innen, Netzwerken oder Organisationen strategisch ins Spiel gebracht werden. Das Ziel der Ausformulierung einer Vision ist es, bestimmten Zukunftstechnologien und ihrem Innovationspotenzial für Wissenschaft und Gesellschaft Legitimität und Bedeutung zu verleihen, um so die Schwerpunkte von Politik und Forschung mitzugestalten (Hausstein & Lösch, 2020; Schneider et al., 2020). Als analytischer Rahmen ermöglicht das Vision Assessment den Blick über die hegemoniale Vision des Metaverses, wie sie von ihren Befürwortern präsentiert wird, hinaus, um andere Entwicklungsmöglichkeiten dieser Technologie zu betrachten (Lösch et al., 2016).

Auch wenn die Metaverse Vision von Meta-CEO Zuckerberg am Anfang ihrer Entwicklung steht, werden darin bereits Ansprüche an Funktionalität, soziale Wirkung und technologischen Prozess formuliert. Die aktuelle Konzeptualisierung entspricht einer soziotechnischen Vision, die nicht einem wissenschaftlichen Diskurs entstammt und (noch) sehr eindimensional erscheint. Im Hinblick auf die Entwicklung der sozialen Medien in den letzten Jahrzehnten erscheint es umso dringlicher, die mögliche zukünftige (Fehl-)Nutzung des Metaverses zu skizzieren, um daraus Handlungsbedarfe abzuleiten.

Zu diesem Zweck haben wir im Mai 2022 Expert*innen aus den Bereichen Extremismusforschung, Technologie und Kriminologie zu einem zweitägigen Workshop nach Karlsruhe eingeladen. Die dort entwickelten Szenarien zeigen, wie sich extremistisches Handeln und Radikalisierung in einem zukünftigen Metaverse möglicherweise verändern und – je nachdem, wie das Metaverse gestaltet sein wird – kontrolliert und abgesichert werden könnten.

In Vorbereitung auf den Workshop entschieden wir uns für die Vorgabe von zwei Schlüsselfaktoren: Freiheit und Sicherheit. In einem abstrakten Sinne bezeichnet Freiheit die Abwesenheit externer Festlegungen und Sicherheit die Erwartung des Ausbleibens negativer Folgen eigenen Handelns (siehe Tabelle 1). Da das Metaverse eine umfassende virtuelle Erlebniswelt der sozialen Interaktion sein soll, ist es naheliegend, dass es entlang zweier Faktoren gestaltet wird, die für funktionierende demokratische Gesellschaften von essenzieller Bedeutung sind. Darüber hinaus zeigt die Geschichte der Einführung und Verbreitung neuerer Social Media, die von einer permanenten Aushandlung von Freiheit und Sicherheit begleitet wurde, welche Bedeutung beiden Konzepte zukommt (Neuberger, 2023). Diese Faktoren sollten in den Ausprägungen hoch und niedrig für die Erstellung einer Szenarien-Matrix mit vier Feldern verwendet werden.

Im weiteren Verlauf des Workshops sollten die Teilnehmenden in einer Gruppendiskussion fünf weitere Faktoren identifizieren, deren Ausgestaltungen den größten Einfluss auf Freiheit und Sicherheit im Metaverse haben würden. Dabei arbeiteten sie 1) die Moderation von Inhalten, 2) den Schutz von Daten, 3) den Jugendschutz, 4) die staatlichen Regulierungsmöglichkeiten und 5) die technologischen Standards im Metaverse heraus. Die Aufgabe der Expert*innen bestand anschließend darin, die jeweiligen Ausprägungen von Freiheit und Sicherheit einzuschätzen, die sich im Zuge der Ausgestaltung dieser Faktoren ergeben würden.

Der Aufbau des Workshops sah vor, dass zunächst allgemeine gesellschaftliche Szenarien des Metaverses entwickelt wurden, um die Teilnehmenden mit der Methode vertraut zu machen und die ursprüngliche Vision von Meta einzuordnen. Am zweiten Tag wurde ein Perspektivenwechsel vorgenommen und die Teilnehmenden wurden gebeten, die Szenarien durch eine extremistische Brille zu entwickeln. Die Frage war also, wie sich die identifizierten fünf Faktoren in Bezug auf Freiheit und Sicherheit für Extremist*innen in den unterschiedlichen Szenarien gestalten.

Im Kontext einer extremistischen Version des Metaverses definierten wir Freiheit und Sicherheit wie folgt:

Tabelle 1

Definitionen von Freiheit und Sicherheit im extremistischen Metaverse

Freiheit im (extremistischen) Metaverse	Sicherheit im (extremistischen) Metaverse
• Ausmaß der Verwirklichungsmöglichkeiten, die sich extremistischen Akteur*innen bieten: - offen rassistische, sexistische, religions- oder menschenfeindliche Äußerungen - Nutzung verbotener Symbole und Schriften	• Risikominimierung gegenüber der - Ent- oder -Aufdeckung bzw. Zugriff durch staatliche (Sicherheitsbehörden) - Konfrontation mit zivilgesellschaftlichen Gegenspieler*innen (Antifa, Medien, NGOs)
• Bandbreite an Möglichkeiten, die für extremistisches Handeln entscheidend sind, für: - die Identifikation von Gleichgesinnten - Grooming-Kommunikation - öffentlichkeitswirksame Aktionen	• Interne/externe Kommunikation (Strategie, Planung, Propaganda) • Durchsetzung eines Regelwerks - Verschwiegenheit - Einigkeit bei Zielen - einen gemeinsamen Willen mit Gewalt durchsetzen • Klandestinität der eigenen Aktivitäten

In der Szenarienentwicklung zeigte sich, dass vor allem die Moderation beziehungsweise die Inklusion der Nutzer*innen sowie die Frage der Sanktionsgewalt die extremistische Freiheit beeinflussen. Die Rolle des Staates bedingt ganz entscheidend die Verwirklichungsmöglichkeiten von malevolenten Akteur*innen im Metaverse, doch ebenso können zivilgesellschaftliches Engagement und Content-Moderation einen Beitrag dazu leisten, malevolenten Inhalten etwas entgegenzusetzen. Aus der Kreuztabellierung anhand der Ausprägungen von Sicherheit (hoch/niedrig) und Freiheit (hoch/niedrig) ergaben sich vier Entwicklungsszenarien (siehe Abbildung 1).

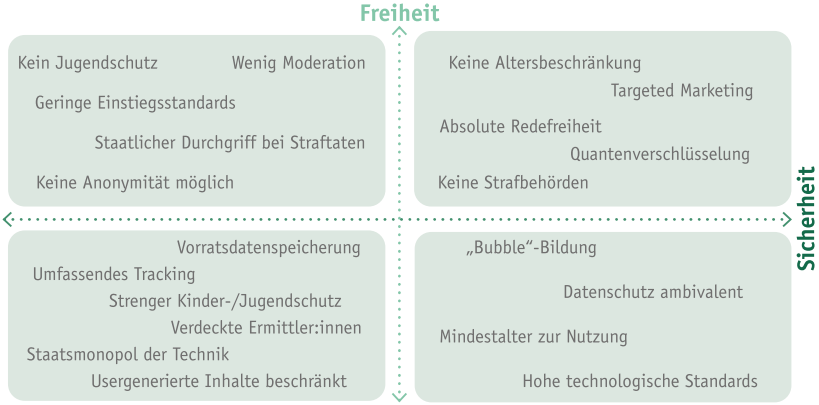

Abbildung 1: Szenarien einer extremistischen Perspektive auf die Nutzung des Metaverses

Lehren aus Vergangenheit und Gegenwart: Wie soll ein zukünftiges Metaverse gestaltet werden?

Das Potenzial des Metaverses für extremistische Aktivitäten hängt davon ab, wie viel Spielraum die jeweils entstehenden Plattformen malevolenten Akteur*innen geben. Freiheit und Sicherheit sind nicht nur die Eckpfeiler liberaler Gesellschaften, sondern auch – in ihrem Gegenteil – die wichtigsten Faktoren für die Verbreitung von extremistischem Gedankengut und Gewalt. Die Szenarien haben deutlich gemacht, wie heikel es sein kann, ein Gleichgewicht zwischen der Einschränkung von Freiheit und Sicherheit für extremistische Akteur*innen zu finden, da dieses Gleichgewicht auch die Gesellschaft im Metaverse als Ganzes betrifft. Die Vision von Mark Zuckerberg und Meta liegt wahrscheinlich zwischen dem ersten und dem dritten Szenario. Die Rolle des Staates und das Ausmaß der Integration der Nutzer*innen in Moderation und Entwicklung werden wichtige Indikatoren dafür sein, welches Szenario in stärkerer Übereinstimmung mit der Vision stehen wird.

Ein weiterer Diskussionspunkt unter den Expert*innen war auch die Frage, ob und inwieweit sich das Metaverse wesentlich von heutigen Social-Media-Plattformen unterscheiden wird. Sollten sich ähnliche Problemlagen abzeichnen, besteht eine realistische Chance, auf der Basis bisheriger Erkenntnisse präventiv arbeiten zu können. In diesem Kontext wurde durch die Workshopteilnehmer*innen wiederholt auf Parallelen zur Gaming-Industrie hingewiesen. Diese ist in der Regel sehr früh in technologische Entwicklungen involviert. Aus der Extremismusforschung ist bekannt, dass extremistische Gruppierungen Spiele und Gamification-Elemente zu Rekrutierungs- und Propagandazwecken nutzen (Englund & Bunmathong, 2022; Regeni, 2023). Sie zeigen dabei eine hohe Innovationskraft und versuchen auf diesem Weg, insbesondere junge Menschen anzusprechen (Koehler et al., 2022). Gleichzeitig existieren gamifizierte Ansätze zur Extremismusprävention beziehungsweise Bemühungen zur Sensibilisierung der Gaming-Community.[3]

Folglich ist die wichtigste Botschaft aus dem Prozess der Szenarienentwicklung die, dass die Metaverse-Entwickler*innen die zukünftigen Nutzer*innen möglichst umfangreich und frühzeitig einbinden. Somit soll eine partizipative Entwicklung sozialer Aspekte im Metaverse gefördert werden. Dies setzt aber auch zahlreiche Bildungsangebote zu partizipativen Gestaltungs-, Handhabungs- und Beteiligungsmöglichkeiten voraus und überträgt auch eine größere Verantwortung auf die zukünftigen Nutzer*innen. Angesichts des Einflusses eines Metaversums auf die politische Meinungsbildung ist jedoch eine demokratische Ausgestaltung unumgänglich (Engelmann et al., 2020). Als Instrumente der Bürgerbeteiligung kommen dafür in der politischen Praxis erprobte Verfahren wie „Mini-Publics“ oder virtuelle Bürgerräte in Frage (Buergerrat.de, 2022; Escobar & Elstub, 2017; Smith & Setälä, 2018).

Es ist jedoch nicht allein Aufgabe der Bürger*innen, demokratische Grundprinzipien im Metaverse zu entwickeln und aufrechtzuerhalten. In ihrem Arbeitspapier berichten Rau et al. (2022) über Herausforderungen und Interventionsmöglichkeiten bezüglich der Plattform-Governance im Kontext von Rechtsextremismus. Sie weisen in diesem Zusammenhang vor allem auf die Notwendigkeit einer vollständigen Transparenz hin, um

[3] *Siehe beispielsweise das Projekt „Good Gaming – Well Played Democracy“ der Amadeu-Antonio-Stiftung (https://www.amadeu-antonio-stiftung.de/projekte/good-gaming-well-played-democracy/).*

(Nutzer*innen-)Partizipation überhaupt erst zu ermöglichen, aber auch um die quasi-staatlichen Monopolstrukturen der Plattformbetreiber*innen aufzubrechen (Bundtzen & Schwieter, 2023; Jiang, 2020; Popiel, 2022; Rau et al., 2022). Damit einhergehend wird durch Rau et al. (2022) angemerkt, dass es neben repressiven Maßnahmen wie Löschung oder Sperrung ebenso wichtig ist, die demokratische Resilienz gemäß dem Prinzip der „wehrhaften Demokratie" durch die Stärkung demokratischer Akteur*innen zu fördern. Mit Blick auf aktive Handlungsstrategien hat sich dabei beispielsweise das Konzept der Gegenrede (Counter Speech) etabliert (Morten et al., 2020; Rau et al., 2022).

Inwiefern sich diese Handlungsvorschläge auch im Metaverse umsetzen lassen, ist noch offen (Schmitt, 2022). Auch zeichnet sich bereits jetzt die Problematik des „legal but harmful content" ab, also Inhalte, welche nicht per se offen extremistisch oder sogar terroristisch sind (und somit auch im Rahmen gesetzlich festgelegter Bedingungen gelöscht werden müssen, siehe TCO-VO),[4] sondern subtil radikalisierend wirken können. Gleichwohl ist hier das Grundrecht auf freie Meinungsäußerung zu berücksichtigen. Nicht zuletzt ist bei der Umsetzung all dieser Governance-Vorschläge die Multistakeholder-Perspektive im Sinne der Beteiligung von politischen Entscheidungsträger*innen, Plattformbetreiber*innen, Sicherheitsbehörden, Zivilgesellschaft und Forschung zu bedenken (Jiang, 2020).

[4] *Die Terrorist Content Online-Verordnung (TCO-VO) beinhaltet die am 28.04.2021 durch das Europäische Parlament beschlossene Regelung, wonach Plattformbetreiber*innen terroristische Inhalte innerhalb von einer Stunde nach Meldung zu entfernen oder zu sperren haben (Verordnung (EU) 2021/784 des Europäischen Parlaments und des Rates vom 29. April 2021 zur Bekämpfung der Verbreitung terroristischer Online-Inhalte (Text von Bedeutung für den EWR), 2021).*

Künstliche Intelligenz und mögliche Auswirkungen auf Extremismus und Radikalisierung

Künstliche Intelligenz als ambivalente Erweiterung menschlicher Möglichkeiten

Künstliche Intelligenz ist ein sehr umfassender Begriff, dessen Definition nicht immer eindeutig ist. Die Europäische Union definiert in ihrem Strategiepapier „Künstliche Intelligenz für Europa" KI als „Systeme mit einem ‚intelligenten' Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen" (Europäische Kommission, 2018, S. 1). Künstliche Intelligenz ist der Überbegriff für Anwendungen, bei denen Maschinen menschenähnliche Intelligenzleistungen wie Lernen, Urteilen und Problemlösen automatisiert erbringen (Brundage et al., 2018; UNICRI & UNCCT, 2021). Viele Publikationen verwenden die Begriffe KI und maschinelles Lernen (ML) gleichermaßen und gleichbedeutend (Schroeter, 2020), dies wird im vorliegenden Beitrag ebenso gehandhabt.

KI wird bereits von sozialen Medien eingesetzt, um extremistische Inhalte schneller löschen zu können (DPA, 2017), in der Terrorabwehr wird an kritischen Orten in manchen Staaten bereits auf Gesichtserkennungssoftware zurückgegriffen (Huszti-Orbán & Ní Aoláin, 2020). Weiteren Anwendungsmöglichkeiten zur Gefahrenabwehr stehen unter anderem Bedenken hinsichtlich des Daten- und Privatsphärenschutzes (Aden & Fährmann, 2020) sowie Fragen der Nützlichkeit für polizeiliche Aufgaben entgegen (Pelzer, 2018).

Im extremistischen Kontext könnten Deepfakes und Social Bots genutzt werden, um Menschen zu verunsichern, ihr Vertrauen in die demokratische Gesellschaft zu erschüttern und in weiterer Folge zu radikalisieren. Doch könnten mithilfe von KI bald ganz neue Möglichkeiten der Radikalisierung geschaffen werden, ebenso könnten sich neue Angriffsvektoren für gewaltbereite Extremist*innen eröffnen (Brundage et al., 2018; Ciancaglini et al., 2020; UNICRI & UNCCT, 2021).

Bereits diese knappen Ausführungen zeigen Ambivalenzen auf, die mit der Erweiterung menschlicher Möglichkeiten durch gegenwärtig verfügbare KI-Anwendungen im Hinblick auf Radikalisierung und Extremismus einhergehen. Zugleich ist damit aber wohl erst der Anfang einer Entwicklung

beschrieben. KI ist schon heute ein fester Bestandteil des Alltags. Für die Zukunft ist es denkbar, dass sowohl die niederschwellige Verfügbarkeit als auch die Verbreitung sowie die Leistungsfähigkeit von KI-Anwendungen weiter zunehmen werden. Diesbezüglich steht ein Technologiemonitoring vor der Herausforderung, mögliche problematische Folgen zukünftiger Entwicklungen frühzeitig zu erkennen.

*KI im Spiegel der Expertise von Informatik, Extremismusforschung und Kriminologie: Erkenntnisse aus einem Expert*innenworkshop*

Um die mit neuen Technologien auf der Basis von KI einhergehenden gesellschaftlichen Chancen und Risiken in Bezug auf Radikalisierung und Extremismus genauer abschätzen zu können, haben wir am ITAS einen Online-Expert*innenworkshop konzipiert und im November 2022 durchgeführt. Ein Erkenntnisgewinn für Wissenschaft und Praxis sollte dabei durch einen interdisziplinären Dialog von Expert*innen aus verschiedenen für die Thematik relevanten Wissensgebieten entstehen.

Im Vorfeld des Workshops wurden zunächst in Vorarbeiten des Technologiemonitorings, unter anderem mithilfe einer zweistufigen Delphi-Befragung, KI-Anwendungen identifiziert, die von Expert*innen als besonders wirkungsvoll angesehen wurden oder deren Einschätzungen mit höherer Unsicherheit verbunden waren und deshalb einer vertiefenden Betrachtung bedürfen. Die so ermittelten zu untersuchenden einzelnen Anwendungen wurden anschließend vom Projektteam geclustert und zu drei Technologiekomplexen zusammengefasst:

1. Targeted Communication
2. Monitoring durch Bilderkennung
3. Kommunikationsoptimierung durch Spracherkennung

Am Workshop nahmen letztlich Expert*innen aus den Bereichen Extremismusforschung, Extremismusprävention, Informatik, maschinelles Lernen, Kriminologie und Sicherheitsbehörden teil. Konkret sollten diese in drei Kleingruppen das Potenzial von KI-Anwendungen in Bezug auf die Technologiekomplexe sowohl für Extremismusbekämpfung und -prävention als auch für Radikalisierung und extremistische Akteur*innen in naher Zukunft (bis 2040) diskutieren. Gewünschtes Ergebnis war hierbei

einerseits das gesammelte Expert*innenwissen, andererseits aber auch Handlungsempfehlungen an Stakeholder*innen zur frühzeitigen Identifikation ungewollter Folgen dieser Innovationen und mögliche Wege, um gegenzusteuern.

Die erste Gruppe befasste sich mit dem Themenkomplex Targeted Communication. Dieser beinhaltete sämtliche KI-gestützten Anwendungen, die zur gezielten Erreichung von Personen oder Gruppen mittels personalisierter Inhalte benutzt werden können, beispielsweise Social Bots oder Fake Accounts/Content. Dabei wurde herausgearbeitet, dass automatisierte Manipulation und individuell angepasste radikalisierende Kommunikation bereits möglich sind. Algorithmen auf TikTok oder YouTube führen innerhalb kürzester Zeit zu extremistischen Inhalten (Matamoros-Fernández et al., 2021; Thomas & Balint, 2022). In Zukunft könnten in diesem Bereich vor allem die vollständige, authentisch wirkende Automatisierung von Kurznachrichten (zum Beispiel auf Messenger-Diensten) sowie die Manipulation von Diskussionen und Debatten durch Bots eine große Rolle für Radikalisierung und Extremismus spielen. Für Akteur*innen in den Bereichen Deradikalisierung, Prävention und Strafverfolgung bietet KI im Bereich Targeted Communication laut den Expert*innen vielfältige Möglichkeiten. So ist beispielsweise vorstellbar, dass eine KI automatisiert allgemeine Disclaimer über extremistische Inhalte legt, um somit noch nicht radikalisierte Personen zu warnen. Ebenso könnten die Algorithmen der verschiedenen sozialen Medien für Aufklärung genutzt werden, um auf ein verschwörungstheoretisches Video ein einordnendes Video zu empfehlen. Als weitere Möglichkeit wurde die Einrichtung von automatisierten Umleitungen zu Gegennarrativen genannt, die durch bestimmte Suchanfragen getriggert werden könnten.

Die zweite Gruppe setzte sich mit Technologien zum automatisierten Monitoring mittels Bilderkennung auseinander. Hierunter fielen Technologien zur Identifizierung über Mustererkennung (beispielsweise Objekte und Orte), aber auch biometrische Mustererkennungen, zum Beispiel Gesichts- oder Gangerkennung. Außerdem fielen in diesen Bereich KI-gestützte Bildgeneratoren. Hier waren sich die Expert*innen einig, dass das Problem der Deepfakes weiter an Relevanz und Qualität gewinnen werde. Dadurch könnte gesamtgesellschaftliches Vertrauen unterminiert werden, um den Boden für radikale und extremistische Lösungen zu bereiten. Ferner könnte Bilderkennung zur Identifikation von Gleichgesinnten eingesetzt

werden, zum Beispiel über automatisierte Erfassung von Sharepics oder Memes.[5] Eine dritte Nutzungsmöglichkeit von KI zur Bilderkennung für extremistische Zwecke wäre die Verfolgung einer Person mittels Drohnen, bei der Gesichtserkennungssoftware eingesetzt wird.[6]

Für die Gegenseite stellen sich hier sehr ähnliche Anwendungsmöglichkeiten dar, und zwar die Aufdeckung von Deepfakes mittels KI-Tools und die dementsprechende automatisierte Platzierung von Gegennarrativen. Des Weiteren kann KI genutzt werden, um gefährdete Personen zu identifizieren, die beispielsweise vermehrt beginnen, extremistische Memes zu teilen, um hier frühzeitig mit präventiven Maßnahmen eingreifen zu können. Das Potenzial dieser Technologien schätzten die Expert*innen durchgehend als hoch ein. Zwar sei große Rechenleistung zum Betrieb der Applikationen notwendig, dies könne jedoch durch die zu erwartende Verbreitung als Dienstleistung (as-a-Service) und damit verknüpfter Cloud-Computing-Kapazität für illegale Vorhaben leicht ausgeglichen werden.

Die dritte Gruppe befasste sich mit dem Technologiekomplex der Spracherkennung. Darin wurden alle KI-Anwendungen zusammengefasst, die dazu dienen, Kommunikation zu optimieren, wie beispielsweise „Natural Language Processing (NLP)“ und „Text-to-Speech-Applikationen“. Mithilfe NLP-produzierter Texte können extremistische Akteur*innen unter anderem Ressourcen einsparen und Propagandamaterial schneller generieren. Die Expert*innen schätzten vor allem Übersetzungsanwendungen als besonders relevant ein, wodurch verschiedene Sprachräume miteinander verbunden werden können. So könnten bedeutsame Reden simultan übersetzt oder extremistisches Material von erfahrenen Gruppierungen bearbeitet werden, um die Übersetzungen für andere, ressourcenschwächere Sprachgruppen aufzuarbeiten. Auf diese Weise wäre eine einfachere transnationale Vernetzung denkbar. Die Kontaktaufnahme und Radikalisierung über Messenger und soziale Medien könnte sich in Zukunft noch effektiver mithilfe von „Recruitern“ gestalten. Weiterhin könnten technische Mittel zur Emotionserkennung verfügbar werden,

[5] *Dies sei für heutige KI-Anwendungen oftmals noch schwierig, da die Maschinen bei Bildern ohne Text keinen Kontext erkennen und mithin Subtext, Sarkasmus, Humor oder szeneinterne Codes nicht erschließen können.*

[6] *Die zuverlässige Erkennung einer Person von hinten oder oben ist für moderne biometrische Applikationen durch das Fehlen klarer Merkmale (zum Beispiel eines farbigen Rucksacks) noch unmöglich.*

mit denen Menschen leichter von extremistischen Inhalten erreicht werden können. Die Expert*innen stellten die These auf, dass Recruiter und die Emotionserkennung gegebenenfalls den Trend zu Einzeltäter*innen verstärken, da klassische Organisationsstrukturen nicht mehr zwingend benötigt werden.

Mensch oder Maschine: Wie lassen sich die Probleme unerwünschter Folgen von zukünftigen Anwendungen künstlicher Intelligenz in Bezug auf Extremismus und Radikalisierung lösen?

Die Handlungsempfehlungen der Expert*innen beziehen sich auf verschiedene Adressat*innen und Gebiete. So richteten sich technologische Empfehlungen überwiegend an die Adresse der Entwickler*innen und Technologie-/Plattformbetreiber*innen sowie an die Sicherheitsbehörden und thematisieren dabei das technische Design oder technisch gestützte Mittel zur gezielten Information, Deradikalisierung und Prävention. Die sozialen Empfehlungen nahmen zivilgesellschaftliche Prozesse, Regulierungen und Bildungsangebote in den Blick, um mit ihrer Hilfe malevolente Verwendungen zu verhindern, zu erschweren und/oder zu bewältigen.

Um malevolenter Verwendung von KI-Technologien technisch zu begegnen, könnten die Entwickler*innen und Betreiber*innen die Affordanzen der Technologie gezielt verändern oder aber ausnutzen. Im Fall von „Targeted Communication" könnten Betreiber*innen von sozialen Netzwerken/Plattformen Disclaimer schalten, problematische Inhalte löschen, problematische Akteur*innen mit „shadow bans" versehen, gezielte Inhalte moderieren („content moderation") und technisch gestützt organisierte Gruppen erkennen beziehungsweise leichter aufbrechen. Hier könnten staatliche oder supranationale Vorgaben helfen, wie sie zum Beispiel bereits im Digital Services Act der EU teilweise angedacht sind. Zusätzlich gewinnen KI-Zertifizierungskonzepte an Bedeutung für Missbrauchsprävention, so die Expert*innen. Hier steckt die Zertifizierung jedoch noch in den Kinderschuhen, und wichtige Fragen bleiben ungeklärt, vor allem jene nach den Kriterien und der Nachvollziehbarkeit der Entwicklung.

KI-Anwendungen könnten perspektivisch ebenfalls Ressourcen schonen. Beispielsweise könnten NLP-Anwendungen und automatisierte Übersetzungsanwendungen bei flächen- und sprachenübergreifender Content-

Moderation sehr hilfreich sein. Auch in diesem Fall könnte auf Sprachräume mit viel Erfahrung zurückgegriffen werden, um weniger ressourcenreiche Räume mithilfe von Übersetzungen zu moderieren. Hier muss laut Expert*innen jedoch klar die Grenze zwischen verboten und „harmful but legal" gezogen werden. Wichtig sei außerdem, dem Staat nicht leichtfertig weitere technologische Zugriffsmöglichkeiten zu gewähren. Sicherheitsbehörden hätten bereits jetzt ein sehr großes technisches Überwachungspotenzial, und die Grenzen, die gegeben sind, beziehen sich lediglich darauf, dass ein Verdacht gegeben sein muss, der die Überwachung rechtfertigt. Aus Expert*innensicht sollten Sicherheitsbehörden eher beobachten, beraten und die Zivilgesellschaft fördern.

Als einen bedeutsamen Faktor für die zukünftige Auseinandersetzung und Verwendung von KI-Applikationen erachten die Expert*innen die soziale Vermittlung: Falls es dazu kommt, dass künstlich erstellte und übersetzte Medien (Texte, Bilder und Videos) verbreitet werden, um zu radikalisieren, zu verunsichern oder zu spalten, dann ist es von zentraler Bedeutung, dass die Gesellschaft diese als solche identifizieren kann. Im Bereich der digitalen Medien seien ältere Menschen ohne Medienkompetenz vergleichsweise anfälliger für solche Fälschungen als jüngere Bevölkerungsgruppen. Gleichzeitig müsse auch die „Data-Literacy" in allen Sicherheitsbehörden erhöht werden. Beiden Herausforderungen könnte durch gezielte und umfassende Bildungsmaßnahmen begegnet werden, doch könne dies keinesfalls der einzige Lösungsansatz bleiben. Zuletzt betonten die Expert*innen noch die Notwendigkeit einer phänomenunabhängigen Aufklärung über technische Möglichkeiten (beispielsweise Deepfakes) für Stakeholder*innen. Hier könnten etwa Mediator*innen eingesetzt werden, die an der Schnittstelle von technischem Know-how zu KI und staatlichen Entscheidungsträger*innen vermitteln, da ein tiefes Verständnis von KI auch für letztere oft nur schwer erreichbar sei.

Ein finaler wichtiger Punkt, der durch die Expert*innen immer wieder betont wurde, ist, dass natürlich auch das Potenzial von KI als Technologie einem gesellschaftlichen Aushandlungsprozess unterliegen sollte. Die Grenzen dessen, was KI leisten kann, und die Grenze dessen, was sie leisten soll, sind nicht grundsätzlich im Einklang.

Diskussion

Im Rahmen des MOTRA-Technologiemonitorings wurde mithilfe von Expert*innenworkshops sowohl für das Metaverse als auch für drei KI-Technologiekomplexe versucht, deren Potenzial im Kontext von Extremismus und Radikalisierung abzuschätzen. Dabei zeigte sich, dass technologische Lösungsansätze für Folgeprobleme nur ein Teil des Puzzles sind und der soziotechnische Kontext als wesentlich relevanter erachtet wird. Die gesammelten Einschätzungen legen nahe, dass die betrachteten Technologien extremistische Aktivitäten bedeutend erleichtern beziehungsweise verändern können. Die Expert*innen hoben deren offenen, in ihren Verwendungsmöglichkeiten nicht vollständig festgelegten, Charakter hervor, in dem inhärente Affordanzen angelegt sind. Auf der einen Seite wirkt sich dieser Umstand auf etwaige Nutzungsszenarien aus, sprich malevolente oder benevolente Use Cases. Auf der anderen Seite können präventive Maßnahmen genau hier ansetzen. Beispielsweise indem jene eingeschriebenen Eigenschaften vor einer Veröffentlichung identifiziert, zur Diskussion gestellt, sozial verhandelt und vor der endgültigen Implementierung modifiziert werden.

Aus TA-Perspektive und aus der Sicht der Expert*innen ist es darüber hinaus sinnvoll, Technologien wie das Metaverse und KI im Zusammenspiel zu denken, da sie jeweils als Ermöglichungstechnologien aufeinander wirken können. Beispielsweise basieren derzeitige Visionen des Metaverses auf KI-beschleunigter Infrastruktur sowie KI-Anwendungen (siehe die NVIDIA Keynote auf der Computex 2023; NVIDIA, 2023). Wird diesen Visionen entsprochen, ist das Metaverse geeignet, die Entwicklung von KI zu beschleunigen und zu fördern. Parallel dazu ist die Forderung nach einer demokratischen Kontrolle der Technologien und der frühzeitigen Einbindung zivilgesellschaftlicher Akteur*innen folgerichtig und wichtig, da davon auszugehen ist, dass die Vermeidung malevolenter Use Cases auch im Interesse der Betreiber*innen ist. Dies steht auch im Einklang mit verschiedentlich publizierten Forderungen, die Weiterentwicklung von KI-Werkzeugen vorerst zu stoppen, um nachhaltigen Schaden von der Menschheit abzuwenden (CAIS, 2023). Offensichtlich wird befürchtet, dass sich diese Technologien zu Entitäten entwickeln, die sich jeglichem Kontrollzugriff widersetzen.

Extremist*innen nutzen digitale Alltagstechnologien für ihre Zwecke und Aktivitäten. Gleichzeitig verändern digitale Technologien den Kreis an Personen, die gegenüber jenen Aktivitäten vulnerabel sind. Das verlangt Anstrengungen zur Aufklärung und zu politischer Bildung. Ein Ausschnitt aktueller Fragen könnte in Anlehnung an die Ergebnisse des KI-Workshops wie folgt formuliert werden:

- Wie kann Aufklärung außerhalb des Bildungssystems erfolgen beispielsweise für Berufstätige und ältere Menschen?
- Wie können ressourcenärmere Räume effektiver unterstützt werden?

Es wurde im Kreis des Expert*innenworkshops formuliert, dass für die Bearbeitung dieser Fragen die entsprechende Infrastruktur bereitgestellt werden muss, um das Zusammentreffen von Wissenschaft und Praxis gewährleisten zu können. Der Aufruf zur Vernetzung betrifft ebenso und im Besonderen sicherheitsbehördliche Einrichtungen, um deren Perspektiven und Bedürfnisse – im demokratischen Sinne ausreichend – berücksichtigen zu können. Hinsichtlich der generierten Ergebnisse aus beiden Workshops kann der Mehrwert einer engeren Verzahnung von Akteur*innen aus Wissenschaft, Praxis sowie Sicherheitsbehörden exemplarisch angenommen werden.

Kritisch angefügt werden muss, dass die angewendeten Methoden auch Limitationen ausgesetzt sind. Bezüglich beider Workshops hängt das Ergebnis von der erreichbaren fachlichen Expertise, aber vor allem auch von dem, durch die Methoden aktivierbaren, Vorstellungsvermögen der Teilnehmer*innen ab. Deshalb erzeugt der Blick auf unterschiedliche Szenarien keine Gewissheit, wie die Entwicklungen der Technologiekomplexe konkret ablaufen werden, regt aber Diskussion an, wie mit diesen umgegangen werden könnte. Im Sinne der Einflussnahme auf die Entwicklung der besprochenen Ideen bewegt sich TA auch hier in einem Spannungsfeld zwischen Koordinations-, Planungs- und Aktivierungsfunktion, welche sich im sogenannten Collingridge-Dilemma widerspiegelt. Demnach ist TA entweder frühzeitig involviert, muss dann aber auf unzureichende Datenbasis zurückgreifen oder beteiligt sich erst so spät, dass eine durch TA angeregte Veränderung schon nicht mehr möglich ist, da technologie-charakteristische Attribute bereits unwiderruflich implementiert worden sind (Grunwald, 2022; Wehling, 2021)

Nichtsdestotrotz zeigte die Diskussion der Expert*innen im TA-Kontext deutlich, dass bezüglich des Metaverses auf die Erfahrungswerte in der Präventionsarbeit im Umgang mit sozialen Medien zurückgegriffen werden kann und auch soll. Besonders Inhalte im Sinne von „legal but harmful" werden wohl auch in Zukunft Plattformmoderation vor eine große Herausforderung stellen. Ebenso werden extremistische Anwendungsfälle von KI-Technologien bereits jetzt breit diskutiert (zum Beispiel Siegel & Doty, 2023). Folglich ergeben sich für die TA basierend auf den Resultaten des KI-Workshops neue Impulse, vor allem die Funktionsweise der „niedrigschwelligen Automatisierung" bietet hier Ansatzpunkte zu weiteren Analysen.[7] Der rasanten Entwicklungsgeschwindigkeit neuer Technologien folgend, ist es naheliegend, dass die TA-Perspektive größere Bedeutung einnehmen wird, welche sich auch in einer Form des Echtzeitmonitorings manifestieren sollte, um „am Puls der Zeit" zu bleiben. Nicht zuletzt ist in diesem Kontext die transdisziplinäre Arbeit verschiedener Stakeholder*innen ein vielversprechendes Mittel zur Implementierung demokratischer Werte und Schutzmechanismen, um der malevolenten Nutzung technologischer Innovationen etwas entgegenzusetzen.

[7] *Siehe auch die Möglichkeiten zur automatisierten Gegenrede via Social Bots (Clever et al., 2022)*

Literatur

Aden, H., Fährmann, J. (2020). Datenschutz-Folgenabschätzung und Transparenzdefizite der Techniknutzung: Eine Untersuchung am Beispiel der polizeilichen Datenverarbeitungstechnologie. *TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, *29* (3), Article 3. https://doi.org/10.14512/tatup.29.3.24

Bazu, T. (2021). *The metaverse has a groping problem already*. MIT Technology Review. Abgerufen von https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S. J., Belfield, H., Farquhar, S. & Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Abgerufen von https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

Buergerrat.de. (2022). *Bürgerrat diskutierte über künstliche Intelligenz*. Buergerrat.de. Abgerufen von https://www.buergerrat.de/aktuelles/buergerrat-diskutierte-ueber-kuenstliche-intelligenz/

Bundtzen, S., Schwieter, C. (2023). *Datenzugang zu Social-Media-Plattformen für die Forschung: Lehren aus bisherigen Maßnahmen und Empfehlungen zur Stärkung von Initiativen inner- und außerhalb der EU*. Institute for Strategic Dialogue (ISD). Abgerufen von https://isdgermany.org/datenzugang-zu-social-media-plattformen-fuer-die-forschung/

Büscher, C., Kusche, I. (2023). *Monitoring new and emerging technologies in order to prevent extremism and terrorist violence* [Manuscript].

Büscher, C., Kusche, I., Röller, T., Andres, F., Gazos, A., Hahn, J., Ladikas, M., Madeira, O., Plattner, G. & Scherz, C. (2022). Trends der zukünftigen Technologienutzung im Kontext von Extremismus und Terrorismus: Erste Erkenntnisse aus dem MOTRA-Technologiemonitoring. In U. Kemmesies et al. (Hrsg.), *MOTRA-Monitor 2021* (S. 248–281). Wiesbaden: MOTRA.

CAIS (2023). *Statement on AI Risk*. Center for AI Safety. Abgerufen von https://www.safe.ai/statement-on-ai-risk

Chohan, U. W. (2022). *Metaverse or Metacurse?* SSRN Scholarly Paper. https://doi.org/10.2139/ssrn.4038770

Ciancaglini, V., Gibson, C., Sancho, D., McCarthy, O., Eira, M., Amann, P. & Klayn, A. (2020). *Malicious Uses and Abuses of Artificial Intelligence. Trend Micro Research, United Nations Interregional Crime and Justice Research Institute (UNICRI), Europol's European Cybercrime Centre (EC3)*. Abgerufen von https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence

Clever, L., Klapproth, J. & Frischlich, L. (2022). Automatisierte (Gegen-)Rede? Social Bots als digitales Sprachrohr ihrer Nutzer*innen. In J. Ernst, M. Trompeta & H.-J. Roth (Hrsg.), *Gegenrede digital: Neue und alte Herausforderungen interkultureller Bildungsarbeit in Zeiten der Digitalisierung* (S. 11-26). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-36540-0_2

Cronin, A. K. (2020). *Power to the People: How Open Technological Innovation is Arming Tomorrow's Terrorists*. New York: Oxford University Press.

Cropley, D. H., Kaufman, J. C. & Cropley, A. J. (2008). Malevolent Creativity: A Functional Model of Creativity in Terrorism and Crime. *Creativity Research Journal*, *20* (2), 105–115. https://doi.org/10.1080/10400410802059424

Diaz, A. (2022). *Disturbing reports of sexual assaults in the metaverse: 'It's a free show'*. New York Post. Abgerufen von https://nypost.com/2022/05/27/women-are-being-sexually-assaulted-in-the-metaverse/

DPA (2017). Internet: Facebook sucht Terrorinhalte mit künstlicher Intelligenz. *ZEIT ONLINE*. Abgerufen von https://www.zeit.de/news/2017-06/15/internet-facebook-sucht-terrorinhalte-mit-kuenstlicher-intelligenz-15213203

Engelmann, S., Grossklags, J. & Herzog, L. (2020). Should users participate in governing social media? Philosophical and technical considerations of democratic social media. *First Monday*. https://doi.org/10.5210/fm.v25i12.10525

Englund, G., Bunmathong, L. (2022). Understanding the Game: Bridging Research Gaps at the Nexus of Gaming and Extremism (*GNET Insights*). GNET. Abgerufen von https://gnet-research.org/2022/03/09/understanding-the-game-bridging-research-gaps-at-the-nexus-of-gaming-and-extremism/

Escobar, O., Elstub, S. (2017). Forms of Mini-Publics: An introduction to deliberative innovations in democratic practice [Research and Development Note]. *New Democracy.* Abgerufen von https://www.newdemocracy.com.au/2017/05/08/forms-of-mini-publics/

Europäische Kommission (2018). Mitteilung der Kommission an das Europäische Parlament, den Europäischen Rat, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen: Künstliche Intelligenz für Europa. Abgerufen von https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

Verordnung (EU) 2021/784 des Europäischen Parlaments und des Rates vom 29. April 2021 zur Bekämpfung der Verbreitung terroristischer Online-Inhalte (Text von Bedeutung für den EWR), 172 OJ L 79 (2021). Abgerufen von http://data.europa.eu/eli/reg/2021/784/oj/deu

Europol (2022). Policing in the metaverse: What law enforcement needs to know: An observatory report from the Europol Innovation Lab. Publications Office of the European Union. https://doi.org/10.2813/81062

Forge, J. (2010). A Note on the Definition of "Dual Use". *Science and Engineering Ethics*, *16 (1),* 111-118. https://doi.org/10.1007/s11948-009-9159-9

Greipl, S., Hohner, J., Schulze, H. & Rieger, D. (2022). Radikalisierung im Internet: Ansätze zur Differenzierung, empirische Befunde und Perspektiven zu Online-Gruppendynamiken. In U. Kemmesies et al. (Hrsg.), *MOTRA-Monitor 2021* (S. 42–70). Wiesbaden: MOTRA.

Grunwald, A. (2022). Aus der Werkstatt der Technikfolgenabschätzung. In A. Grunwald (Hrsg.), *Technikfolgenabschätzung: Einführung* (S. 167-214). Baden-Baden: Nomos. https://doi.org/10.5771/9783748928775

Guston, D. H., Sarewitz, D. (2002). Real-time technology assessment. *Technology in Society*, *24 (1),* 93-109. https://doi.org/10.1016/S0160-791X(01)00047-1

Hausstein, A., Lösch, A. (2020). Clash of Visions: Analysing Practices of Politicizing the Future. *BEHEMOTH – A Journal on Civilisation*, *13 (1),* 83-97. https://doi.org/10.6094/behemoth.2020.13.1.1038

Huq, A. Z. (2022). Militant Democracy Comes to the Metaverse. *Emory Law Journal*, *72*.

Huszti-Orbán, K.,, Ní Aoláin, F. (2020). Use of Biometric Data to Identify Terrorists: Best Practice or Risky Business? Human Rights Center University of Minnesota. Abgerufen von https://www.ohchr.org/Documents/Issues/Terrorism/biometricsreport.pdf

Jiang, J. A. (2020). *Toward a Multi-Stakeholder Perspective for Improving Online Content Moderation.* University of Colorado.

Koehler, D., Fiebig, V. & Jugl, I. (2022). From Gaming to Hating: Extreme-Right Ideological Indoctrination and Mobilization for Violence of Children on Online Gaming Platforms. *Political Psychology*, *44 (2),* 419–434. https://doi.org/10.1111/pops.12855

Lösch, A., Böhle, K., Coenen, C., Dobroć, P., Ferrari, A., Heil, R., Hommrich, D., Sand, M., Schneider, C., Aykut, S., Dickel, S., Fuchs, D., Gransche, B., Grunwald, A., Hausstein, A., Kastenhofer, K., Konrad, K., Nordmann, A., Schaper-Rinkel, P. & Wentland, A. (2016). Technikfolgenabschätzung von soziotechnischen Zukünften. *Diskussionspapiere: Institut für Technikzukünfte.* https://doi.org/DOI:10.5445/IR/1000062676

Mahfoud, T., Aicardi, C., Datta, S. & Rose, N. (2018). The Limits of Dual Use. *Issues in Science and Technology*, *34 (4),* 73–78.

Matamoros-Fernández, A., Gray, J. E., Bartolo, L., Burgess, J. & Suzor, N. (2021). What's "Up Next"? Investigating Algorithmic Recommendations on YouTube Across Issues and Over Time. *Media and Communication*, *9 (4),* 234–249. https://doi.org/10.17645/mac.v9i4.4184

Meta (2021a). *C*onnect 2021: Our vision for the metaverse. Tech at Meta. Abgerufen von: https://tech.facebook.com/reality-labs/2021/10/connect-2021-our-vision-for-the-metaverse/

Meta (2021b). The Metaverse and How We'll Build It Together—Connect 2021. Youtube. Abgerufen von: https://www.youtube.com/watch?v=Uvufun6xer8

Morten, A., Frischlich, L. & Rieger, D. (2020). Gegenbotschaften als Baustein der Extremismusprävention. In J. B. Schmitt, J. Ernst, D. Rieger & H.-J. Roth (Hrsg.), *Propaganda und Prävention* (S. 581-589). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-28538-8_32

Neuberger, C. (2023). Sicherheit und Freiheit in der digitalen Öffentlichkeit. In N. J. Saam, H. Bielefeldt (Hrsg.), *Die Idee der Freiheit und ihre Semantiken* (S. 297–308). Bielefeld: transcript Verlag. https://www.transcript-open.de/doi/10.14361/9783839461884-028

NVIDIA (Regisseur) (2023, Mai 30). NVIDIA Keynote at COMPUTEX 2023. Abgerufen von: https://www.youtube.com/watch?v=i-wpzS9ZsCs

Pelzer, R. (2018). Policing of Terrorism Using Data from Social Media. *European Journal for Security Research*, *3 (2),* 163-179. https://doi.org/10.1007/s41125-018-0029-9

Popiel, P. (2022). Digital Platforms as Policy Actors. In T. Flew, F. R. Martin (Hrsg.), Digital Platform Regulation: Global Perspectives on Internet Governance (S. 131–150). Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-95220-4

Rau, J., Kero, S., Hofmann, V., Dinar, C. & Heldt, A. P. (2022). Rechtsextreme Online-Kommunikation in Krisenzeiten: Herausforderungen und Interventionsmöglichkeiten aus Sicht der Rechtsextremismus- und Platform-Governance-Forschung. *Arbeitspapiere des Hans-Bredow-Instituts.* https://doi.org/10.21241/SSOAR.78072

Regeni, P. (2023). *Accelerationism Meets Gamification: A Look at the Convergence in the Framing of Online Narratives (Insights).* Abgerufen von https://gnet-research.org/2023/05/31/accelerationism-meets-gamification-a-look-at-the-convergence-in-the-framing-of-narratives-online/

Schmitt, J. B. (2022). Virtuelle Welten und künstliche Intelligenzen als Herausforderungen und Chancen digitaler Gegenrede. In J. Ernst, M. Trompeta & H.-J. Roth (Hrsg.), *Gegenrede digital. Interkulturelle Studien.* (S. 27–40). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-36540-0_3

Schneider, C., Roßmann, M. & Lösch, A. (2020). Sociotechnical Visions of 3D Printing – After the First Hype? Report of the Vision Assessment Study in the Cluster of Excellence 3D Matter Made to Order *(Nr. 138; KIT Scientific Working Papers)*. Karlsruhe. https://doi.org/10.5445/IR/1000117984/V2

Schroeter, M. (2020). Artificial Intelligence and Countering Violent Extremism: A Primer. GNET. Abgerufen von https://gnet-research.org/2020/09/28/artificial-intelligence-and-countering-violent-extremism-a-primer/

Siegel, D., Doty, M. B. (2023). Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda (GNET Insights). Global Network on Extremism & Technology. Abgerufen von https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/

Singh, K. (2022). *In The Metaverse, Sexual Assault Is Very Real.* Abgerufen von https://www.refinery29.com/en-us/2022/06/11004248/is-metaverse-sexual-assault-illegal

Smith, G., Setälä, M. (2018). Mini-Publics and Deliberative Democracy. In A. Bächtiger, J. S. Dryzek, J. Mansbridge & M. Warren (Hrsg.), *The Oxford Handbook of Deliberative Democracy* (S. 300–314). Online Ausgabe: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198747369.013.27

Stilgoe, J., Owen, R. & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42 (9),* 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Thomas, E., Balint, K. (2022). Algorithms as a Weapon Against Women: How YouTube Lures Boys and Young Men into the 'Manosphere' (ISD Discussion Paper). Abgerufen von https://www.isdglobal.org/isd-publications/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/

UNICRI & UNCCT (2021). Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes. United Nations Interregional Crime and Justice Research Institute; United Nations Office of Counter-Terrorism. Abgerufen von https://unicri.it/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes

Wehling, P. (2021). Technikfolgenabschätzung und Wissenschaft. In S. Böschen, A. Grunwald, B.-J. Krings & C. Rösch (Hrsg.), *Technikfolgenabschätzung: Handbuch für Wissenschaft und Praxis* (S. 178–190). Baden-Baden: Nomos.

Weimann, G., Dimant, R. (2023). The Metaverse and Terrorism: Threats and Challenges. *Perspectives on Terrorism*, *17 (2),* 92–107.

Wiederhold, B. K. (2022). Sexual Harassment in the Metaverse. *Cyberpsychology, Behavior, and Social Networking*, *25 (8),* 479–480. https://doi.org/10.1089/cyber.2022.29253.editorial